Research Report

Data sharing in credit markets:

Does comprehensiveness matter?

Giovanni Barci
Galina Andreeva
Sylvain Bouyon

European Credit Research Institute

No. 23
August 2019

Final Study

Study sponsored by ACCIS

Data sharing in credit markets:

Does comprehensiveness matter?

Giovanni Barci

Galina Andreeva

Sylvain Bouyon

CEPS-ECRI and the University of Edinburgh Business School, gratefully acknowledge the sponsorship received for this report from ACCIS, BIK, BISZ, BKR, CRIF, Equifax, Experian, KSV1870 and SCHUFA. The views expressed in this report are those of the authors alone and do not reflect the opinion of ACCIS or of any of its members.

The European Credit Research Institute (ECRI) is a research institution that operates within the Centre for European Policy Studies (CEPS) in Brussels. Established in 1999 for the study of banking and credit in Europe, ECRI focuses on institutional, economic and legal aspects related to retail finance and credit reporting. The institute provides expert analysis and academic research for a better understanding of the economic and social impact of credit. ECRI supports and funds independent academic research projects. The institute monitors markets and regulatory changes and looks at their impact nationally and internationally.

ECRI gratefully acknowledges the sponsorship received for this report from ACCIS. The views expressed in this report are those of the authors alone and do not reflect the opinion of ACCIS or any other member of CEPS.

Giovanni Barci is Research intern at CEPS. Galina Andreeva is Senior Lecturer in Management Science, University of Edinburgh, Business School. Sylvain Bouyon is Research Fellow and Head of Retail Finance and Fintech at CEPS.

This version of the study is available for free downloading from the ECRI website (www.ecri.eu).

ISBN 978-94-6138-745-5

EUROPEAN CREDIT RESEARCH INSTITUTE (ECRI)
Place du Congrès 1
B-1000 Brussels, Belgium
Tel.: +32-2-2293911
Email: info@ecri.eu
Web: www.ecri.eu

Contents

List of Boxes, Tables and Figures

Executive Summary

International institutions such as the World Bank have shown that data sharing on borrowers contributes to improving the risk profile of borrowers and increasing access for more customers to credit markets. Several relevant academic articles underline these results. Thus, the existence of data sharing between lenders is commonly acknowledged as one of the core ingredients of successful credit markets.

The originality of the present study is to analyse whether sharing more comprehensive data improves the functioning of credit markets in European countries. Assuming that mechanisms to share data do exist, does a higher comprehensiveness in the data collected matter for credit markets? The study answers this by firstly analysing whether higher comprehensiveness in the data collected by credit reference agencies (CRAs) in the EU-28 can impact credit markets, and secondly, by analysing to what extent the sharing of non-traditional data, that is, data not directly related to credit activities, can contribute to well-functioning credit markets.

1. The impact of the data collected by European CRAs

Background and methodology used

By adopting an econometric approach, the purpose is to assess whether higher comprehensiveness in the data (both traditional and non-traditional) provided by CRAs can impact credit markets. In particular, whether higher data comprehensiveness can impact on a number of macroeconomic indicators, namely (i) financial inclusion, (ii) financial intermediation (this is the allocation of consumer deposits to credit activities, or the supply of credit), and (iii) the risk of missed repayments. Indexes of comprehensiveness are created by using a database built by European CRAs with the data types they hold. These include indexes to reflect the breadth and the depth of the data collected.

The *level of breadth* of the data collected focuses on the types of borrowers, the types of data collected on borrowers, the types of organisations supplying data to CRAs, and the types of services CRAs collect data on. A general index of breadth is created to aggregate all these features and one sub-index is developed for each main feature. As European CRAs primarily collect structured ‘traditional’ data, a distinct index to cover structured ‘non-traditional’ data is created.

The *level of depth* reflects the granularity of positive and negative data CRAs collect for all the services covered. A general index for depth is developed, as well as distinct sub-indexes for positive data, negative data and structured data on non-loan services, respectively.

The main drivers behind the degree of data comprehensiveness are also analysed to better understand the data sharing mechanisms in credit markets. These include the role of competition in credit markets, as well as the laws that restrict or enable data sharing.

Econometric models which test the impact and drivers of the indexes are built at country-level. As the number of observations is rather limited (between 34 and 60), any interpretation of the results needs to be done carefully. Several control variables have been included and different tests have been conducted in order to achieve an acceptable level of robustness in the findings (see Annex 2 for further details).

Findings on financial inclusion

Financial inclusion is measured by the consumer credit to GDP ratio and the share of the population which has taken credit in the 12 previous months. The latter is divided according to the level of income: the 60% richest and the 40% poorest of the population in each country.

Our findings show that with more detailed information, creditworthiness assessment outcomes could change for some consumers. A first group of consumers could be considered less likely to default than originally anticipated. Conversely, another group could be assessed more likely to default than initially predicted. Therefore, it was assumed that the collection of more data by CRAs should have two opposite effects on inclusion: one that boosts lending to those who can afford it and one that reduces lending to those who cannot.

<u>The findings reveal that more comprehensive data overall boosted lending, thereby increasing financial inclusion</u>. <u>As a result of higher comprehensiveness in the data collected by CRAs, the increase in lending to safer borrowers more than offsets the lower lending to riskier borrowers</u>.

This finding is confirmed by higher general breadth and the collection of more structured data on non-loan services. Both indexes contributed to increasing significantly all variables for financial inclusion, no matter whether the consumers were on lower or higher incomes. Finally, a greater depth of positive data boosted both measures of financial inclusion - the consumer credit to GDP ratio and the share of lower income consumers who had been able to access credit in the last 12 months.

Findings on financial intermediation

Financial intermediation (the supply of credit) is measured by the ratio of consumer credit to deposits. Again, increasing the general breadth and the comprehensiveness index of non-loan structured data have a robust positive impact on this ratio. <u>This shows that when CRAs collect more comprehensive data it encourages banks to channel savings into consumer credit rather than into other financial activities</u>.

Findings on the risk of missed repayments

The risk of consumers missing repayments is measured by the future ratio of non-performing loans (NPL). <u>Higher general breadth in the data collected by CRAs reduces this future (NPL) ratio.</u> <u>This result could link higher comprehensiveness with the improvement in the risk profile of the pool of borrowers</u>.

Findings on drivers behind comprehensiveness

Different drivers are assessed in this study. The capital requirement ratio is used as a proxy for the barriers of entry in credit markets. The degree of concentration in credit markets provides a measure of the degree of competition on those markets. Finally, an index is developed to reflect to what extent domestic laws enable data sharing with CRAs.

Results reveal that established lenders are more willing to share data when capital requirements are more constraining and the risk of competing with new players is limited. In addition, higher concentration in credit markets tends to limit data sharing. Last but not least, laws enabling the sharing of data seem to have a positive significant impact on the breadth of data that is shared.

2. The impact of non-traditional data

Background and methodology used

There is evidence in economic literature that previous credit history is highly predictive of future credit performance. In addition, several relevant publications show that accuracy improves with a greater level of detail, that financial inclusion increases by adding positive data and that CRA-data being used in scoring models increases their accuracy.

Non-traditional data goes beyond the traditional credit history and may include structured data (e.g. rent and utility bill payments and banking transactions) and unstructured data (e.g. mobile phone use, web browsing and psychometrics). The adopted typology of what is non-traditional data varies from country to country. Only recently have some CRAs started to collect certain types of such data.

There is also evidence from international organisations and industry reports that non-traditional data can offer value to both lenders and borrowers. The evidence comes predominantly from developing countries, where a lack of traditional shared credit data is a known obstacle for individuals in getting access to finance. In order to draw parallels with the EU context, qualitative interviews have been conducted with eight relevant stakeholders (three from banking institutions, three from consumer organisations and two from CRAs).

Structured versus unstructured: which is more effective?

Structured non-traditional data is perceived as the most useful type of data (especially for affordability assessments) and the most socially/legally acceptable. Unstructured non-traditional data (texts, images, web browsing) is attracting mixed opinions, with some stakeholders seeing its potential value (some banks and Fintech startups), whilst others (several consumer associations) are completely opposed to its use.

Main benefits of non-traditional data

The main benefits of non-traditional data are seen as increasing the predictive accuracy of credit risk models, increasing financial inclusion, and offering a holistic view of a customer. The use of non-traditional data should be complementary to traditional types, since the latter is still more powerful. However, it depends on the specific country and portfolio.

In cases where traditional data is less powerful, the expected utility of non-traditional data should be higher. In cases where traditional data is limited or not available, the non-traditional data may be used on its own. The customers who are expected to benefit most from the use of non-traditional data include those with ‘thin files’ (people with no or little credit history). This is supported by the evidence from studies mainly conducted in developing countries and the USA.

Main challenges of the use of non-traditional data

Legal compliance is seen by almost all respondents as the main challenge/risk associated with the use of non-traditional data in credit granting. Other challenges include data quality, IT/ technological difficulties and social acceptance/ethics.

3. Policy recommendations

In light of these findings, several policy recommendations can be formulated:

- To boost financial inclusion, policy-makers should facilitate the availability of more comprehensive credit data, notably structured non-loan data.
- More attention should be paid to the extent of credit data sharing in the analysis of borrowers' solvency and the accessibility of credit.
- Policy changes should be introduced to enable wider data sharing in credit markets, particularly those markets with a high degree of concentration and relatively low barriers to entry.
- Legislation should not promote information that is only "up to date and accurate", as mandated by data protection legislation, but also "comprehensive".
- If unstructured, non-traditional information, such as social media data, is to be used as an input to determine creditworthiness, it should be done within a clearer ethical framework and with a better understanding of societal preferences.
- Further research should investigate the impact of non-loan data for specific segments of the population.

Acronyms

ACCIS: Association of Consumer Credit Information Suppliers

AI: Artificial intelligence

AUC: Area under the Curve

CFPB: Consumer Federal Protection Bureau

CRA: Credit reference agency

ECB: European Central Bank

FICO: Fair Isaac Corporation

GAFA: Google, Amazon, Facebook and Apple

GPFI: Global Partnership for Financial Inclusion

GDP: Gross domestic product

GDPR: General Data Protection Regulation

ICCR: International Committee on Credit Reporting

IMF: International Monetary Fund

KYC: Know Your Customer

LTD: Loan-to-deposit ratio

NPL: Non-performing loans

P2P: Peer-to-peer

RWA: Total risk-weighted assets

Glossary

Affordability – a credit applicant's economic ability to repay a loan.

Algorithm – a logical procedure or computational method that is used for problem solving.

Application data – data provided by a credit applicant on the application form.

Artificial Intelligence (AI) – computer programs or systems that perform functions normally requiring the human brain.

Area under the Curve (AUC) – a numeric measure of predictive accuracy of the credit scoring model, with 0.5 indicating no predictive power, and 1 perfect prediction, i.e. that all real defaults are correctly predicted as such. It is a rescaled version of the Gini coefficient, another popular measure of predictive power: AUC = Gini/2+0.5.

Behavioural data – data coming from existing borrowers regarding their financial behaviour, e.g. from monthly bank statements.

Big Tech/GAFA – major technology companies that have exceptional influence, often referred to as GAFA, i.e. Google, Amazon, Facebook and Apple.

Characteristic – a piece of information describing a credit applicant/borrower, a variable in a credit scoring model.

Credit score – an output from a credit scoring model, a numeric summary of the creditworthiness of a credit applicant, most commonly based on the estimated probability of default.

Credit scoring/credit risk assessment – mathematical, statistical, machine-learning predictive models used for assessing and ranking borrowers or credit applicants.

Credit Reference Agency (CRA)/Credit Bureau – a system for collecting, processing and facilitating the exchange of information regarding credit applications and payment behaviour from credit grantors and possibly other private and public sources, normally on the basis of reciprocal data-sharing agreements.

Creditworthiness – a broad concept referring to whether a particular individual should be granted credit or can be regarded as an attractive/acceptable customer by a particular lender. Assessing creditworthiness before granting credit is seen as a key component of responsible lending. The assessment typically evaluates the risk a borrower will not make repayments (credit risk), and the risk a borrower will not have sufficient income to make repayments (affordability risk).

Cut-off score – a predetermined level of predicted probability of default that is used as a threshold to decide which applicants should be accepted for credit. It is normally established during the credit scoring model development and corresponds to a risk appetite of a given lender.

Default – failure to repay the loan on agreed terms, most commonly measured as three missed consecutive monthly repayments.

Fin Tech – "Technology-enabled innovation in financial services that could result in new business models, applications, processes or products with an associated material effect on financial markets and institutions and the provision of financial services" (Financial Stability Board).[1] By extension, any company operating in that field.

Gini coefficient – a numeric measure of predictive accuracy of the credit scoring model, with 0 indicating no predictive power, and 1 – perfect prediction, i.e. that all real defaults are correctly predicted as such. It is a rescaled version of Area under the Curve (AUC), another popular measure of predictive power: Gini = (AUC-0.5) x 2.

Know Your Customer (KYC) – procedures aimed either at the identification of a customer for legal purposes or at understanding a customer's preferences for marketing and service improvement.

Logistic regression – a popular statistical model used in credit scoring, suitable for a binary outcome: e.g. Default/No Default.

Negative data[2] – statements about defaults or arrears and bankruptcies. It may also include statements about lawsuits, liens and judgments that are obtained from courts or other official sources.

Non-traditional/alternative credit data – data coming from non-traditional sources that may not be directly related to credit, such as financial transactions, certain other obligations like rental information, payment of utility bills, geo-location, mobile phone use, or social media posts.

Open banking – an initiative that provides financial services customers with a possibility to securely share their financial data with multiple financial institutions.

Positive data[3] - information that covers facts of contractually compliant behavior. It includes detailed statements about outstanding credit, amount of loans, repayment patterns, assets and liabilities, as well as guarantees and/or collateral. The extent to which positive information is collected typically depends on national legislation, including the data protection regime.

Predictive accuracy – extent to which estimates from a model agree with actual values.

---

[1] www.fsb.org/work-of-the-fsb/policy-development/additional-policy-areas/monitoring-of-fintech/.

[2] The definitions for negative and positive data are the ones of the World Bank, 2011 in "General Principles for Credit Reporting" (see: http://documents.worldbank.org/curated/en/662161468147557554/pdf/70193-2014-CR-General-Principles-Web-Ready.pdf).

[3] See previous footnote.

Structured data – information normally in a numeric form that follows a fixed format and can be easily stored in databases and spreadsheets. The formal structure makes it easy to enter, store and analyse.

Thin file segment/population – people that do not have extensive previous credit history, therefore, their files or credit reports are 'thin'.

Traditional credit data – data normally used for credit risk assessment coming from a credit application form and/or previous credit behaviour, e.g. credit utilisation (loans, mortgages) length of credit history, credit activity, etc.

Unstructured data – information that does not follow a fixed format and is not organised in a predefined manner. Examples include text files, images, social media data and sensor data.

General introduction

It is commonly acknowledged that data sharing between lenders is one of the core components of successful credit markets. Over the last 20 years, the positive impact of this mechanism has been proven by international institutions such as the World Bank, which has developed an active agenda to promote data sharing in credit markets around the world, and several relevant academic articles published in significant peer-review journals. It has notably been shown that the existence of data sharing on borrowers should contribute significantly to improving the risk profile of the pool of borrowers (reducing the risk of default or NPLs), as well as to giving certain consumers and firms a better chance of accessing credit markets.

However, some specific questions have not been addressed in detail by the economic literature. Empirical analyses on the impact of data sharing are generally conducted with simple dummies. For example, is there a credit register? Is positive data collected? Little has been written on the impact of more comprehensive data and the actual role of non-traditional data, which can be defined as all data that is not directly related to credit activities. Assuming that mechanisms to share data do exist, does a higher comprehensiveness in the data shared matter for credit markets?

The purpose of the present study is, therefore, to address these two issues, by assessing whether the collection of more granular data and non-traditional data can have a significant impact on metrics as important as financial inclusion, the risk of missed repayments and the allocation of deposits to credit activities. In order to have a better understanding of the mechanism to share data in credit markets, the main drivers behind the degree of comprehensiveness of the data shared will also be analysed.

The study will address all these issues in two complementary chapters (ECRI covered the first one, while the University of Edinburgh contributed to the second one). The first chapter will analyse whether higher comprehensiveness in the data collected by credit reference agencies (CRAs) can impact credit markets. To do so, it will use a database that provides very granular information on the type of data collected by CRAs across the EU-28. Several indexes of comprehensiveness will be built in order to econometrically assess the impact of more comprehensive data on a set of dependent variables that, among others, proxy financial inclusion and financial risks. In addition, the main drivers of the degree of comprehensiveness of the data collected will be analysed, by determining the role of concentration and competition in credit markets, as well as the laws that restrict or enable data sharing.

While the primary focus of the database used in the first chapter is on the traditional data collected by CRAs, it also covers the case of non-traditional data. However, this only concerns structured data provided by non-bank entities such as utilities. This type of data accounts for a relatively small share of what is usually considered non-traditional data. Therefore, in order to have a full-fledged analysis of the type of data that might affect credit markets, the second chapter will consider a large set of non-traditional data (both structured and unstructured) and, based on several interviews with different stakeholders, well-targeted desk research, and the expertise of the team, will provide qualitative analyses on the role so far of non-traditional data in credit markets and possible future scenarios.

Chapter 1.
Structured data sharing in credit markets: does comprehensiveness matter?

Introduction

This chapter primarily aims to assess how the comprehensiveness of CRA data in Europe can be measured, and the impact the degree of comprehensiveness has. Some emphasis is also placed on the main factors behind this degree of comprehensiveness. In the first section, the focus is placed on four dimensions: the impact of the degree of comprehensiveness on financial inclusion, financial intermediation, the risk of missed repayments, and the drivers behind the degree of comprehensiveness. For each dimension, a literature review of the most relevant publications is conducted, and the specific question addressed by the study is addressed.

In the second section, a description is provided on the methodology used to develop several indexes of data comprehensiveness. Additional analyses are presented on the trends observed in the indexes. The main characteristics of both dependent and control variables are reviewed in the third section. The fourth section covers in more detail technical features of the econometric models and tests conducted to reinforce the robustness of the findings. Finally, for each of the four dimensions analysed, the econometric/economic strategy, the limitations of such a strategy, and the final results are provided in detail.

1. Background, literature review and questions to be addressed

1.1 Financial inclusion

One of the primary focuses of the economic literature on credit reporting systems is to analyse the impact of data sharing on financial inclusion. According to the World Bank, "[f]financial inclusion means that individuals and businesses have access to useful and affordable financial products and services that meet their needs – transactions, payments, savings, credit and insurance – delivered in a responsible and sustainable way".[4]

In the literature analysing the interplay between borrowers' data sharing and financial inclusion, the degree of financial inclusion is generally proxied by the credit to GDP ratio. In 1993, Jappelli et al.[5] found an ambiguous effect of information sharing on the volume of lending. The theoretical model created by the authors revealed that when banks exchange

[4] The full definition of the World Bank can be found here: www.worldbank.org/en/topic/inancialinclusion/overview.

[5] Pagano, M., & Jappelli, T. (1993). Information sharing in credit markets. *The Journal of Finance*, *48*(5), 1693-1718.

information about borrowers, this should result in both additional lending to safe borrowers and lower lending to riskier borrowers.

On one hand, a better understanding of consumers' financial capacity could mean that some consumers who were lent a certain amount could from then on be lent less or nothing, as their ability to make repayments was previously overestimated (effect one).[6] To a certain extent, this could be interpreted as a decrease in financial inclusion for the sake of reducing financial risk. On the other hand, consumers whose ability to make repayments was underestimated could gain access to credit or to higher amounts (effect two). Effect two could be perceived as 'actual' financial inclusion.

As a result, the cumulative impact of data sharing on the volume of lending depends on which of the two effects prevails. A positive significant impact of data sharing on credit to GDP ratios would imply that the additional lending to safer consumers more than offsets the decreased lending to riskier ones. Such a finding would imply that financial inclusion is improved. Conversely, a negative significant impact would mean that the decrease recorded in lending to riskier consumers dominates, thereby reflecting a lower level of financial inclusion.

In 2002,[7] the same authors developed empirical analyses based on their own survey and showed that information sharing (proxied by the presence of a credit bureau and/or credit register, and several other simple metrics) has a positive significant impact on bank lending to GDP. The date existing credit bureaus/registers were founded did not matter and the nature of the ownership of these institutions (privately owned versus publicly owned) did not affect the relationship.

Finally, Djankov et al. (2007) proved that both information sharing institutions and creditor protection through the legal system had a positive significant effect on financial inclusion (proxied by the private credit to GDP ratios). However, the effect of creditor protection on financial inclusion differs by countries, being greater in richer countries. The presence of a credit bureau tends to boost private credit in poorer countries but not in richer ones. The most likely explanation is that almost all rich countries do have such credit bureaus and therefore the presence of such an institution does not really contribute to differentiating such countries from one another.

The use of the credit to GDP ratio as a proxy for financial inclusion has pros and cons. Among the pros, there is a high level of consistency across EU countries in the methodologies used to build statistics on GDP and the outstanding amounts of different types of credit. Also, the measures of outstanding amounts of credit cover the vast majority of loans of a given product issued. Finally, although this ratio is often used as an indicator of financial depth, several relevant publications also used it to mirror financial access, as analysed above.

---

[6] For example, by considering microeconomic data with firms, Doblas-Madrid et al. (2013) showed that information sharing induces creditors to grant smaller loans and leases and to demand more guarantees.

[7] Jappelli, T., & Pagano, M. (2002). Information sharing, lending and defaults: Cross-country evidence. *Journal of Banking & Finance*, *26*(10), 2017-2045.

However, provided that rising comprehensiveness in the data shared on borrowers contributes to improving the accuracy of creditworthiness assessments, this could eventually result in the two opposite effects analysed by Jappelli et al. (1993). The main limitation of using credit to GDP ratios to proxy financial inclusion is that the effects cannot be disentangled. The coefficient of the credit to GDP ratio would simply mirror the aggregate outcome of these effects.

The average amounts lent and the actual share of the population that has credit cannot be deduced directly. In some cases, this could result in a misleading interpretation. For example, if the share of the population that has a credit decreases but the average amount granted to remaining creditors increases to such a degree that it more than overcomes the effect sparked by the decreasing share of creditors, then the credit to GDP ratio would increase (provided that movements in GDP do not cancel out these effects). The ratio's increase could, therefore, be wrongly interpreted as a rise in financial inclusion, whereas underlying mechanisms are more ambiguous.

Furthermore, the use of credit to GDP ratios to proxy financial access cannot provide any detailed insights on the differences across consumer segments. Policies aimed at boosting inclusion typically target consumers who are younger, on lower income or who have lived abroad as they are the groups that are typically excluded from credit markets.

To conclude, the first purpose of the present study is to empirically address different issues related to the impact of data sharing on financial inclusion in Europe. The main question concerns the extent to which additional data sharing on borrowers affects financial inclusion, and whether it contributes to boosting it. Particular focus will be placed on different types of data that can be shared (loan data versus non-loan data; positive data versus negative data; etc.). In addition, the present study will use indexes which can measure the different forms of financial inclusion: amounts lent, the share of the population accessing credit, the impact of the income and saving of consumers on the credit access of those consumers, etc. The use of some of the indicators developed by the World Bank to differentiate financial access across income groups should help in that respect.[8]

1.2 Financial intermediation

Another strand of economic literature examines whether the existence of a mechanism of data sharing encourages banks to intermediate private sector savings into private credit rather than into other financial activities that are unrelated to credit in the real economy. Some authors such as Beck et al. (2009) and Gianetti et al. (2011) have focused on this question by integrating metrics from bank balance sheets. One of these metrics concerns the loan to deposit ratio (LTD), which can be used as a proxy for intermediation efficiency.

[8] This data can be found in the Global Findex Database of the World Bank at https://datacatalog.worldbank.org/dataset/global-financial-inclusion-global-findex-database.

Therefore, the present study will assess to what extent the sharing of more comprehensive data can persuade banks to increase the supply of credit and lend more to consumers.

1.3 Financial risk

One key function of credit reporting systems is to contribute to reducing the financial risk of credit markets. Several relevant research publications have already focused on this issue. In 1993, Jappelli et al. used as a core assumption of their theoretical model that information sharing improves the risk profile of the pool of borrowers, decreases defaults and reduces average interest rates. The article published by the authors in 2002 tested empirically the impact of information sharing on two variables used to proxy default rates: loan loss provision and an index of credit risk. Using only large standard errors, they found a negative impact on both proxies. By adopting a microeconomic approach and focusing on firms, Doblas-Madrid et al. (2013)[9] proved that information sharing (proxied by the affiliation to a credit bureau) reduces contract delinquencies, especially for informationally opaque and risky borrowers.

In the present study, the objective is to analyse at country level the impact of the shared data's granularity on financial risks by focusing exclusively on European economies. Different variables can serve as proxies for financial risks. Given the focus of the study on credit markets at country level, one of the most relevant metrics concerns the ratio of non-performing loans (NPLs). The question is, therefore, to assess whether an increase in the comprehensiveness of the data collected by CRAs has had an impact on NPL ratios across Europe.

1.4 Degree of competition and concentration, and legal environment

The interplay between, on one hand, the structure of credit markets and the local legal environment of CRAs, and, on the other hand, the extent to which information on borrowers is shared can take different forms. Causality can be analysed in both directions. First, does the mechanism for sharing data influence the degree of competition on credit markets? Second, do local laws and the degrees of concentration and competition in credit markets influence the extent to which credit data is shared?

Economic literature addressing the first question is limited. The article published by Pagano et al. (1993) is the only relevant paper identified that focuses on that issue. By using almost exclusively theoretical tools, the authors showed that data sharing in consumer credit markets tends to boost lending and facilitate the emergence of new players. This should, therefore, result in lowering the degree of concentration and boosting competition in the lending market. No empirical publication has been identified that covers this issue.

The second question has also been covered by Pagano et al. (1993), as they proved theoretically that information sharing between lenders increases endogenously. Among the

[9] Doblas-Madrid, A., & Minetti, R. (2013). Sharing information in the credit market: Contract-level evidence from US firms. *Journal of financial Economics*, *109*(1), 198-223.

mechanisms influencing information sharing, fear of competition from potential entrants appeared to be a powerful disincentive. This finding was empirically confirmed by Bruhn et al. (2013), who analysed the relationship between private credit market and credit data sharing. To do so, they used a large sample of countries around the world and focused on the averages of several variables over the 2005-10 period. Among other findings, the authors showed that a greater threat of new entrants had a negative impact on the likelihood of having a private credit bureau, the ability of credit bureaus to report positive information and the right of borrowers to have access to their credit history by law.

Bruhn et al. (2013) also showed that the degree of concentration tends to have a broadly similar impact on the willingness of banks to share proprietary data. When a limited number of banks cover a large market share of credit, they are more willing to maintain their oligopoly rents by limiting the exchange of data with other players. In that context, the emergence of a private credit bureau is less likely.

Given the absence of empirical findings in the economic literature regarding the first question, the present study will focus on the second question by somewhat refining some of the results of Bruhn et al. (2013). In particular, the study will integrate more refined indexes that truly capture the degree of granularity of the data shared. Furthermore, the use of indexes mirroring the degree to which laws can restrict or enable data sharing could help better assess the impact of the regulatory framework. The objective would be therefore to answer the following question: Do European countries with higher concentration in credit markets, lower barriers of entry, or more restrictive laws display lower indexes of comprehensiveness in the data shared?

2. Indexes of comprehensiveness

2.1 Literature review

The literature on the impact of data sharing on credit markets has adopted different approaches to measure the extent to which data is shared. Some authors have developed their own survey or have used a microeconomic database to collect the necessary data. Jappelli et al. (2002) created a survey of 49 countries around the world to assess key information on the presence of credit bureaus, credit registers, etc. Doblas-Madrid et al. (2013) followed a microeconomic approach, by using the Payment Information Network database[10] to conduct their empirical analyses on a large set of firms and assess whether each firm had joined the bureau before the credit contract started.

[10] Some more details on this database can be found in Doblas-Madrid et al. (2013).

However, most of this economic literature has used the indexes developed by the World Bank in the context of the International Finance Corporation (IFC) "Doing business" database.[11] The World Bank database is the result of a survey of more than 180 countries. The geographic area covered differs depending on the objective of each article. The database of the World Bank was used by Karapetyan et al. (2013)[12] for a sample of 26 countries in Central and Eastern Europe and the former Soviet Union, while Bruhn et al. (2013),[13] Djankov et al. (2007)[14] and Giannetti et al. (2011)[15] used it for a large set of countries all around the world.

Based on the data collected by the World Bank, authors have built different indicators of comprehensiveness. Typical information that is used relates to whether a credit bureau and/or a credit registry operate(s) in the country, whether both firms and individuals are covered, whether positive and/or negative data is collected and distributed, whether the registry distributes data which is at least two years old or whether the registry has existed for more than three years. Finally, a transparency index can be built based on the answers to five questions which relate to the ability of borrowers to access their credit data and scoring.

The World Bank survey has several advantages. Firstly, it is used by many authors who focus on the economic and financial issues related to credit data sharing. Therefore, it has become a sort of international standard for this economic literature. Secondly, it is conducted on a yearly basis. Thirdly, it is likely the single largest effort to build consistent data on these issues across a large set of countries on all continents. Nevertheless, for certain variables, the degree of granularity is rather limited. In addition, key information is missing for many countries (see Bruhn et al., 2013). Finally, the very broad coverage of countries also tends to erase some local specificities. In this context, as analysed in the next section, the present study proposes to use a much more granular database specifically designed for Europe: the Survey of the Association of Consumer Credit Information Suppliers (ACCIS).

2.2 Methodology used

Advantages and disadvantages of the ACCIS Membership Survey

One of the main advantages of the ACCIS Membership Survey is to provide a high level of granularity in the data collected. It is the macroeconomic database with the highest degree of comprehensiveness that has been identified. In addition, there is a time dimension, as the

---

[11] One of the aims of this database is to develop an indicator on the ease of getting credit in countries around the world. To do so, the World Bank covers questions on the existence of credit registers. The database is available at www.doingbusiness.org/.

[12] Karapetyan, A., & Stacescu, B. (2013). Information sharing and information acquisition in credit markets. *Review of Finance*, *18*(4), 1583-1615.

[13] Bruhn, M., Farazi, S., & Kanz, M. (2013). *Bank competition, concentration, and credit reporting*. The World Bank.

[14] Djankov, S., McLiesh, C., & Shleifer, A. (2007). Private credit in 129 countries. *Journal of financial Economics*, *84*(2), 299-329.

[15] Giannetti, C., & Jentzsch, N. (2011). Credit Reporting, Access to Finance and Identification Systems: International Evidence.

survey has been conducted four times, in 2009, 2012, 2015 and 2017. However, data is available only for 2012, 2015 and 2017. The survey covers 32 CRAs. Therefore, at best, up to 96 observations can be used at CRA level. The database also includes several questions on the national institutional framework that can influence the activities of CRAs. Finally, it has been designed mainly for the EU-28, so a high degree of consistency is assured.[16]

Among the drawbacks, the ACCIS Membership Survey does not cover all EU-28 countries, as it includes only 19. In the context of this study, the French credit register was asked questions similar to the ones contained in the ACCIS Membership Survey, for 2012, 2015 and 2017. This means that 20 EU-28 countries are covered in the econometric part. The Survey also provides data for three non-EU economies whose level of economic development is broadly similar to countries of Northern-Western Europe and Scandinavia: Iceland, Norway and Switzerland.

This implies that, at country level, up to 69 observations could be used in the regressions without lagged variables and 46 with lagged variables. The use of broadly 40-70 observations should provide a satisfactory level of robustness in the empirical findings, provided that specific issues such as outliers or cross-section/over time heterogeneity have been well addressed. A large share of the empirical literature on credit data sharing that has been published by the World Bank and/or in high-quality peer-review journals have been using a broadly similar number of observations. For instance, La Porta et al. (1998) and Jappelli et al. (2002) focused on 40 observations, and Bruhn et al. (2013) used between 37 and 133 observations, depending on the regressions.

Before using the database, several tests have been conducted with the data provided by the respective CRAs in order to reinforce consistency over time and across CRAs. The database used for the regressions is therefore slightly different from the one that is published on the ACCIS website.[17]

Creation of six indicators for the breadth of data collected

Thanks to the high granularity proposed by the ACCIS Membership Survey, several indicators of comprehensiveness can be built (see Tables 1 and 2). The study created two groups of these indicators. The first one covers six indexes for the breadth of data. These indexes provide data for 2012, 2015 and 2017. As shown in the below tables, Breadth_General includes information on the types of borrowers, the types of data collected on borrowers, the types of organisations which supply data and the types of services for which data is collected.

This index combines 73 questions which are distributed as follows: nine for the types of borrowers, 14 for the types of data collected on borrowers, 31 for the types of organisations and 19 for the types of services. This high number of questions allows the development of

[16] The database is designed for all ACCIS members, including those from non-EU countries: Switzerland, Iceland, Norway, Russia, Serbia and Kosovo.

[17] The findings of the latest survey are at https://accis.eu/wp-content/uploads/2018/11/AUG18_ACCIS-Survey-of-Members-2017_FINAL.pdf.

refined indexes that can control for many local specificities regarding data collecting processes.

For each indicator, the values are ranged from zero to one, and are calculated as the proportion of actually collected items over total possible items. The absence of collected items results in an index whose value is null. On the other hand, when the CRA collects data for all items possible, the value of the index is one.

In order to test for different assumptions and to improve robustness of the findings, five other indicators have been developed for breadth (Breadth_Individual, Breadth_Providers, Breadth_Services, Breadth_Loan-Services and Breadth_Non-loan-Services). Breadth_Individual covers the 14 questions related to the type of data collected on borrowers. Breadth_Providers includes the questions on the 31 types of organisations. Breadth_Services integrates the questions on the 19 types of services.

In the ACCIS Membership Survey, as shown in the Tables 1 and 2, the general question for the 19 types of services is: "Do you hold positive and/or negative data for these products?" Given that the distinction between positive data and negative data has more to do with depth than with breadth, the indicator Breadth_Services simply integrates the answer to the question: "Do you hold data for these products?", whether it is positive or negative. Therefore, when the answer is yes, the value is one.

Finally, based on the 19 types of services, two distinct indexes (Breadth_Loan-Services and Breadth_Non-loan-Services) are created on the breadth of the seven types of loans for which data is or is not collected (mortgage, consumer loans, credit and store cards, education loans, credit line on current account, payday loans and leasing) and on the breadth of the 12 non-loan services for which data is or is not collected (mail order, point of sale credit, energy, water, internet service providers, satellite/cable TV, fixed line telecoms, telecoms-mobile, home rent, health insurance, other insurances and others).

The choice of these six indexes has two specific objectives. The first one has to do with econometric needs. The purpose will be to assess the degree of robustness of specific findings. Should, for example, Breadth_General have a significant impact on a given dependent variable, to what extent do its different components have a significant impact as well?

The second objective is directly related to policy issues. The purpose is to address specific policy questions. Breadth_Individual, for instance, can help assess whether more granular data on borrowers can help credit reporting systems achieve their objectives of inclusion or reduce financial risk. As regards Breadth_Providers, does a higher number of data providers contribute to reinforcing the performance of CRAs? With Breadth_Services, a similar question could be asked regarding the number of services covered.

Last but not least, several policy issues could be addressed with the two indicators Breadth_Loan-Services and Breadth_Non-loan-Services. For example, does the collection of data on more traditional services such as loans contribute to increasing the performance of credit reporting systems? To what extent can the collection of data on non-loan services such as utilities improve financial inclusion and reduce the volume of NPLs?

Creation of five indicators for the depth of data collected

The ACCIS Membership Survey also provides granular data on the type of data that is collected for each of the 19 services covered in Breadth_Services (see Tables 1 and 2). This additional information is defined as the depth of the data collected, and is available for 2015 and 2017.

Five distinct indicators of depth have been developed. The first is the index "Depth_General" and covers a total of 209 questions (11 questions for each of the 19 services covered). As shown in the below table, the 11 questions relate to the types of collected negative data (default – more than three months – in arrears – one to three months – rejected cheque list, number of missed payments and write-offs) and positive data (original contract amount, outstanding amount, payment terms, interest rate, duration of loan and on time payment). The second indicator is the index "Depth_Positive", which provides values on the proportion of total items for which positive data is collected; this indicator is the result of 114 questions. The third indicator is "Depth_Negative", which is based on 95 questions.

The last two indicators are developed on the depth of the data collected for the different types of loans ("Depth_Loan-Services") and for other types of services ("Depth_Non-loan-Services"). Whereas the former covers 77 questions, the latter is the result of 132 questions.

In line with the indicators for breadth, the choice of several depth indexes aims at both strengthening the robustness of the econometric results when possible, and answering some policy questions when relevant. The choice of two indexes for the depth of positive data and negative data can for instance help assess whether the collection of more positive data makes a difference. And thanks to the high granularity of the database, this could be answered with a refined approach.

The distinction between depth for loans and depth for other types of services aims at providing some robust insights to the debate on whether the collection of additional data on non-loan services can help strengthen the performance of CRAs. One of the main limitations of the depth indicators is the lower number of observations, as no equivalent information is available for 2012. As such, at best, 46 observations at country level can be observed for depth.

Creation of two indicators for the legal framework

Two indicators are also developed to mirror the legal environment surrounding CRAs. For 2015 and 2017, the ACCIS Membership Survey provided answers to ten questions on the rules applying in the countries covered (see the list of questions in the below Table 2). The first five questions are compiled to create the Restriction_Index. The last five are combined to develop the Promotion_Index, as those rules are likely to facilitate the sharing of data within credit markets. In line with the indexes for depth and breadth, for each question, a binary rule

(Yes=1; No=0) is applied. Then, the different answers are aggregated, and the aggregate is divided by five in order to range from zero (no type of rules applies in the country) to five (all types of rules apply).

Table 1. Eleven indexes of data comprehensiveness

INDEXES	DESCRIPTION
Breadth Indexes	
Breadth_General*	Proportion of actually collected data over total possible data referring to Question 1, Question 2, Question 3, Question 4
Breadth_Individual*	Proportion of actually collected data over total possible data referring to Question 2
Breadth_Providers*	Proportion of actually collected data over total possible data referring to Question 3
Breadth_Services*	Proportion of actually collected data over total possible data referring to Question 4
Breadth_Loan-Services**	Proportion of actually collected loan data over total possible loan data referring to Question 4
Breadth_Non-loan-Services**	Proportion of actually collected non-loan data over total possible non-loan data referring to Question 4
Depth Indexes	
Depth_General*	Proportion of actually collected data over total possible data (included both loan*/non-loan data* and *positive/negative data*)
Depth_Positive***	Proportion of actually collected positive data over total possible positive data (included loan*/non-loan data*)
Depth_Negative***	Proportion of actually collected negative data over total possible negative data (included loan*/non-loan data*)
Depth_Loan-Services**	Proportion of actually collected loan data over total possible loan data (included *positive/negative data*)
Depth_Non-loan-Services**	Proportion of actually collected non-loan data over total possible non-loan data (included *positive/negative data*)
Regulatory framework Indexes	
Restriction_Index****	Proportion of affirmative answers over total questions
Promotion_Index****	Proportion of affirmative answers over total questions

Source: ACCIS Membership Survey and authors' selection.
Note: The asterisks correspond to the different components described in Table 2 below.

Table 2. Components of the indexes of data comprehensiveness and legal framework

* Survey question regarding Breadth Indexes and Depth Indexes: 1. <u>Question 1</u>: On what type of borrowers does your organisation collect data from providers? 2. <u>Question 2</u>: What type of data do you collect on borrower identity, income and asset, bankruptcy and court judgement? 3. <u>Question 3</u>: What types of organisations supply data to your company? 4. <u>Question 4</u>: Do you collect and hold both positive and negative data for the following items? 5. <u>Depth</u>: For each of the items listed in Question 4, do you collect/hold data on Default (>3 months), In arrears (1-3 months), Rejected cheque list, Number of missed payments, Write-offs, Original contract amount, Outstanding amount, Payment terms, Interest rate, Duration of loan, On time payment?
** 1. <u>Loan data</u>: Mortgage, Consumer loans, credit and store cards, Education loans, Credit line on current account, Payday loans, Leasing 2. <u>Non-loan data</u>: Mail order, Point of sale credit, Energy, Water, Internet service providers, Satellite/cable TV, Fixed line telecoms, Telecoms-mobile, Home rent, Health insurance, Other insurances, Others • Distinction refers to Breadth and Depth Indexes
*** 1. <u>Positive data</u>: Original contract amount, Outstanding amount, Payment terms, Interest rate, Duration of loan, On time payment 2. <u>Negative data</u>: Default (>3 months), In arrears (1-3 months), Rejected cheque list, Number of missed payments, Write-offs • Distinction refers only to Depth Indexes
**** 1. <u>Restriction Index</u>: a. Do you comply with special standards? b. Is there a special legislation/regulation on credit reporting in your country? c. Is your organisation subject to sectorial supervision? d. Are you aware of any national legislation/regulation planned in your country which may affect the credit reporting industry? e. Is there any other law or regulation that restricts data sharing? 2. <u>Promotion Index</u>: a. Is there a regulation requiring to share credit data? b. Is there a regulation requiring to share non-credit data? c. Are there regulations promoting the sharing of data? d. Is access to public/court data permitted by regulation? e. Is access to public/court data required by regulation?
Notes: All indexes range from zero to one. They are calculated as the proportion of actually collected data over total possible data that can be collected. (No data collected → Index = 0; collected data = all possible data that can be collected → Index =1.) (<u>Collected data</u>: data/information actually collected by each CRA; <u>all possible data that can be collected</u>: all the types of data that is assessed in the ACCIS Survey.)

Source: ACCIS Membership Survey and authors' selection.

Box 1. Trends observed in the EU-28 in the indexes for breadth, depth and legal environment

1. Trends observed in the indexes for breadth in the EU-28

Picture in 2017

In 2017, high heterogeneity across CRAs could be observed in Breadth_General in the EU-28. While the maximal value of the index reached 0.775, the minimum stood at only 0.282. The first quartile (Q1) scored below 0.394, the second one (Q2) between 0.394 and 0.504, the third one (Q3) between 0.507 and 0.634, and the fourth one (Q4) above 0.634. This reveals a distribution that is highly dispersed between the minimum and the maximum. This could mirror different business models at CRA level and different domestic realities.

The sample identified for 2017 can be divided into three groups of countries: the first concerns the EU-15 economies from Northern-Western Europe (Austria, Belgium, Denmark, Finland, Germany, Netherlands, Sweden and the UK), the second includes EU-15 economies from Southern Europe (France, Greece, Italy and Spain), and the third relates to the most recent member states (Croatia, Cyprus, Czech Republic, Hungary, Poland, Romania and Slovakia). Considering each quartile, the distribution is as follows: the first group has three countries within Q1, one in Q2, four in Q3 and five in Q4; the second group has one country in Q1, two in Q2, one in Q3 and two in Q4, and the third group includes three in Q1, four in Q2, two in Q3 and none in Q4. Overall, there does not seem to be a clear geographic distribution, as belonging to one of the three groups does not appear to markedly raise the likelihood of being in one particular quartile.

Interestingly, when several CRAs exist within one country, the gap between these CRAs in the value of the index is generally low. This pattern could suggest some 'inertia' at country-level. Two CRAs operating within the same country should comply with similar local rules and therefore should have broadly similar possibilities in terms of what they can and cannot collect.

Evolution between 2012 and 2017

Between 2012 and 2017, a medium upward trend could be observed in Breadth_General. This mirrors the high share of CRAs which have registered an increase over the period (92% of all CRAs analysed, as shown in Table 3 below). Both minimum and maximum increased markedly during those five years (from 0.127 to 0.282 and from 0.592 to 0.775, respectively). In the meantime, a medium divergence could be observed, as reflected by the rising standard deviation (which occurred only between 2015 and 2017), thereby revealing a more fragmented picture in 2017 than in 2012.

As regards the different components of Breadth_General, the second index Breadth_Individual, which covers information on identity, income, asset and bankruptcy/court cases, has followed patterns mostly similar to those of Breadth_General. As 86% of CRAs recorded a rise in the value of this index, the average value of Breadth_Individual has moved around a medium upward path and medium divergence has been observed (reflecting a sharp increase in the standard deviation over the 2012-15 period and a small decrease between 2015 and 2017). Again, both minimum and maximum increased significantly. However, at 0.647, the maximum still scored far from a perfect 1.0. This gap is especially due to the poor coverage by CRAs of consumer/borrower income and asset data.

The index for Breadth_Providers on the number of types of organisations that supply data to CRAs has shown somewhat different dynamics from the index for Breadth_Individual. First, standard deviation has been on average higher and high divergence has been observed since 2012. This could mirror greater diversity in the way business models are shaped with respect to the type of organisations that provide data to CRAs. This could also relate to specific domestic rules that restrict access to certain types of data. Nonetheless, a low upward trend could be registered in the average of the sample, as the index has grown for two-thirds of CRAs.

The standard deviation has been even higher for the index covering Breadth_Services on the types of services for which data is collected. While some CRAs covered data for all services identified, others collected data for a rather limited number of services. The minimal value of the index has nonetheless increased significantly between 2012 and 2017. Interestingly, some specific CRAs might have a low number of organisations from which they collected data, but still had a high number of services for which they can collect data. This might be explained by their technological ability to retrieve data on multiple types of services from a limited number of sources. Overall, the average has increased markedly for Breadth_Services, whereas high divergence has resulted in further fragmentation.

When Breadth Loans-services is considered (mortgage, consumer loans, credit and store cards, point of sale credit, education loans, credit line on a current account, payday loans and SMS loans and leasing), low convergence can be observed over the 2012-17 period. The average has increased significantly to score very high in 2017, at 0.898. While no CRA collected data for all these services in 2012, a small majority of CRAs were doing so in 2017.

Finally, the collection of data on non-loan services (mail order, energy – gas, electricity and oil – water, internet service providers, satellite cable TV, fixed line telecoms, telecoms mobile, home rent, health insurance, other insurances and others) has admittedly shown tremendous growth in its average. However, values of that index have diverged markedly since 2012. On one hand, a group of CRAs collected data for all these non-loan services in 2017, whereas none of these CRAs covered all these services in 2012. On the other hand, another group of CRAs declared they had not collected any data on these non-loan services in 2017. This again can be the result of some combination of local CRA strategies and domestic regulatory restrictions.

Table 3. Trends in indexes for breadth in the EU-28

Index name	Trends in averages	Convergence vs Divergence[18]	% of CRAs for which it has increased
2012-17			
Breadth_General	Upward (medium)	Divergence (medium)	92%
Breadth_Individual	Upward (medium)	Divergence (medium)	86%
Breadth_Providers	Upward (Low)	Divergence (high)	67%
Breadth_Services	Upward (High)	Divergence (high)	92%
Breadth_Loan-Services	Upward (High)	Convergence (low)	77%
Breadth_Non-Loan-Services	Upward (High)	Divergence (high)	79%

Source: ACCIS Membership Survey and author's calculations.

2. Trends observed in the indexes for depth in the EU-28

Data on depth is available only for 2015 and 2017. This makes a priori the analyses on trends less relevant than for breadth. However, some interesting observations can still be made. First, for what concerns Depth_General, low growth was registered in the average of the sample, passing from 0.262 to 0.285. In parallel, rising standard deviation reflected medium divergence across the panel. Among CRAs for which data on Depth_General is available, the minimum remained low for both years (0.072 in 2015 and 0.077 in 2017), whereas the maximum stood at 0.603 for 2015 and 0.746 for 2017.

[18] Convergence implies that the values of the indexes of CRAs gradually change so as to become similar. Conversely, divergence entails that those values develop in a different direction.

As was the case for Breadth_General, no specific geographic distribution could be observed within the sample. The group of EU-15 economies from Northern-Western Europe recorded five countries within the Q1, one within Q2, one within Q3 and three in Q4. The second group covering EU-15 economies from Southern Europe included one country in Q1, three in Q2 and one in Q3. Finally, the group relating to the most recent member states had two countries in Q2, four in Q3 and three in Q4.

The indicator Depth_Negative reveals a slight increase in the average between 2015 and 2017, as well as a noticeable divergence. The 2017 maximum reached 0.758, while the two lowest scores stood at 0.074 and 0.084. Higher divergence was recorded in Depth_Positive. This could be partly explained by the marked growth observed in the index of one CRA (which passed from 0.526 to 0.825), whereas some other bodies continued not to collect any positive data.

Finally, patterns differed significantly between Depth_Loan-Services and Depth_Non-loan-Services. Both indexes recorded small growth in their averages. However, their level differed markedly in 2017, as the index reached on average 0.515 for Depth_Loan-Services and only 0.13 for Depth_Non-loan-Services. This is mainly due to the significant number of CRAs that do not collect (or collect just a little) data on non-loan services. In 2017, there were only four CRAs that scored above 0.25 for non-loan services and 0.20 for loan services. Last but not least, while slow convergence was observed in Depth_Loan-Services, significant divergence was registered between 2015 and 2017 in Depth_Non-loan-Services.

Table 4. Trends in indexes for depth in the EU-28

Index name	Trends in averages	Convergence vs Divergence	% of CRAs for which it has increased
2015-17			
Depth_General	Upward (Low)	Divergence (medium)	52%
Depth_Positive	Stable	Divergence (high)	50%
Depth_Negative	Upward (Low)	Divergence (medium)	55%
Depth_Loan-Services	Upward (Low)	Convergence (low)	52%
Depth_Non-loan-Services	Upward (Low)	Divergence (medium)	25%

Source: ACCIS Membership Survey and author's calculations.

3. Trends observed in the two indexes for the legal environment in the EU-28

In 2017, the average of the EU-28 sample stood at 0.63 for the Restriction index and 0.54 for the Promotion index. The former had a minimum of 0 for one country and a maximum of 1 for three countries, whereas the latter had a minimum of 0.2 for three countries and a maximum of 1 for one country. Between 2015 and 2017, the average remained broadly similar for the Restriction index. This masked disparities across countries for which information is available, since the index remained stable for six countries, decreased for four and increased for seven. The standard deviation was also stable, confirming the absence of convergence or divergence dynamics.

On the other hand, the average recorded for the Promotion index increased significantly, reflecting a desire of local authorities to facilitate the exchange of data. As such, the index rose for 13 countries, did not change for three countries and contracted for only one country. In particular, significant positive changes could be observed in the answers to the question: "Is there a regulation requiring to share credit data?" Finally, the standard deviation contracted somewhat, suggesting a small convergence towards more promotion rules.

3. Analyses of data for other variables

3.1 Dependent variables

The impact of the indexes for comprehensiveness will be tested on five dependent variables. Details and summary statistics are provided in Annex 1, Tables A1 and A2. The values used for all these dependent variables are the ones at the end of the period. The main reason behind this choice is that the indexes of comprehensiveness are also available at the end of the period and no information is available on the evolution of these indexes within each of the periods. Following a similar reasoning, values of control variables are also the ones at the end of the period.

The three first dependent variables capture some elements of financial inclusion. The consumer credit to GDP ratio is the first of these variables. It is based on the ECRI Statistical Package published in 2018.[19] When data on consumer credit is available, it is used by default as dependent variables. This can be explained by the fact that consumer credit markets are in general the most affected by the activities of CRAs. When no information is available specifically for consumer credit, then a proxy mirroring a larger share of total private credit is used and it is assumed that this proxy correlates well with consumer credit.

The two other dependent variables for inclusion are provided by the financial inclusion indicators of the World Bank,[20] which are based on a large survey of countries around the world. These variables provide the percentage of a given population who reported borrowing money from a bank or another type of financial institution in the past 12 months. This data is available for 2015 and 2017. Two indicators have been chosen on the respective percentage observed in two types of economic 'classes' (the 40% poorest and the 60% richest). As expected and as shown in Table A2, the shares are much higher for the 60% richest than for the 60% poorest.

Next, the consumer credit to deposit ratio is calculated by combining the data of the ECRI Statistical Package (for outstanding consumer credit) and the European Central Bank (for outstanding household deposits). Finally, the index of non-performing loans is provided by the International Monetary Fund. It gives the value of non-performing loans divided by the total value of the loan portfolio (including non-performing loans before the deduction of specific loan-loss provisions). The loan amount recorded as non-performing should be the gross value of the loan as recorded on the balance sheet, not just the amount that is overdue. The

[19] This ECRI Statistical Package (www.ceps.eu/publications/ecri-statistical-package-2018-lending-households-europe) is one of the main EU sources for statistics on the outstanding value of household credit in the EU-28.

[20] The indicators can be found in the Global Findex Database 2017 of the World Bank at https://globalfindex.worldbank.org/.

metrics used for loans cover all private loans (both for households and non-financial corporations).[21]

3.2 Control variables

A list of control variables is used in the regressions in order to strengthen the robustness of the findings. As shown in Tables A1 and A2 in Annex 1, there are several macroeconomic variables. First, the GDP per capita ratio is measured according to the purchasing power parity and is provided by the World Bank. The total unemployment rate comes from Eurostat.

Eight control variables that are directly related to the financial sector are also used in some regressions. First, there is the short-term interest rate, which is provided by the ECB and covers consumer loans with an original maturity of between one year and five years. Second, Tier 1 capital, which is also provided by the ECB, estimates the ratio of the bank's core equity capital to its total risk-weighted assets (RWA). Third, the policy rates practised by each domestic central bank and whose value is directly provided by the Bank of International Settlements.

The next four variables are used for the regressions addressing the issues of financial inclusion of different groups of income. They are provided by the financial inclusion indicators of the World Bank for each income group (40% poorest and 60% richest). The first two of them mirror the share of the population which has a financial account. The two others concern the share that has savings in a financial institution. As expected, the value of these variables is significantly higher for the richer group.

For the eighth variable, a proxy is used to mirror the degree of competition in the credit market: the Herfindahl-Hirschman index. The values for this index are provided by the ECB and are calculated by squaring the market share of each firm competing in a market and then summing the resulting numbers. When the value of this index is one, then there is a situation of monopoly in a given market. Conversely, if the value of the index is close to null, then the supply is extremely fragmented. The ECB provides values for this index only for the concentration of total private credit supply (both households and non-financial corporations) and total assets of credit institutions. The present study focuses on the former.

Finally, a set of dummies is added to control for domestic legal framework. This information is provided by Djankov et al. (2007). Four different types of dummies are added in some regressions of the present study: one for legal systems having a UK origin, one for French/Spanish legal origin, one for Scandinavian legal origin and one for legal German origin. By using the classification of Djankov et al. (2007), the legal origin could be identified for all

[21] In its Financial Soundness Indicators (see: http://data.imf.org/?sk=51B096FA-2CD2-40C2-8D09-0699CC1764DA) the IMF promotes the cross-country comparability of the data on NPLs and recommends that loans should be classified as NPL when (1) payments of principal and interest are past due by three months (90 days) or more, or (2) interest payments equal to three months (90 days) interest or more have been capitalised (reinvested into the principal amount), refinanced, or rolled over (that is, payment has been delayed by agreement).

countries of the sample except Cyprus. Desk research has confirmed that the legal origin for Cyprus tends to be French.

The methodologies used for the econometric model and the different tests that have been used are analysed in Annex 1. They notably explain why the regressions have been conducted at country level rather than at CRA level.

4. Financial inclusion: strategy, limitations and results

4.1 Strategy and limitations

The analyses for financial inclusion will focus on four dependent variables. The first one concerns the consumer credit to GDP ratio.[22] The two others are related to different income groups of consumers and provide the share of each group which had borrowed any type of credit in the previous 12 months. As shown in Section 3, the two groups are the 40% poorest and the 60% richest.

The choice of these two segments of consumers has implications in terms of financial inclusion. A shown in Table A2 of Annex 1, the 40% poorest have on average lower access to credit markets than the 60% richest. Several control variables are added to partly explain the gap between both income groups.

First, the share of the population that has access to a financial account can act as a proxy for the share of consumers for which financial data is or is not available. Consumers whose income exclusively comes from the shadow economy are more likely to not have any financial accounts. They might be excluded from credit markets even though their actual wealth would allow them to get a credit. In this context, particular attention will be paid to the impact of non-loan data on the financial inclusion of the two income groups. As expected, Table A2 of Annex 1 reveals that the value of this control variable was on average higher for the richer group. It is anticipated that this control variable has a positive significant impact on financial inclusion.

Another control variable aims to assess the potential financial wealth which is saved in financial accounts. The expectation is that a higher share of the population that has such wealth will positively contribute to the financial inclusion of that population, as credit providers are more likely to provide loans to people with such collateral. As shown in Table A2 of Annex 1, a marked gap can be observed in the average of that variable across income groups.

Regarding other control variables, the GDP per capita is used to capture the effects related to whether the economic development of a country can have an impact on the financial depth and inclusion of the economy. The general assumption is that the more developed the

---

[22] Cyprus was excluded from this panel, as the value of its consumer credit to GDP ratio can be characterised as an outlier.

economy, the higher the values of all dependent variables since the financial system is assumed to be more sophisticated and more accessible to consumers. Therefore, rising GDP per capita should contribute to increasing financial inclusion.

However, although this may be true for the whole household credit market, the result could still be different for specific types of credit. For instance, by considering the figures registered in 2017 across the EU-28, simple correlations reveal a low (but significant) positive correlation between the total household credit to GDP ratio and GDP per capita, a moderate positive correlation between housing credit to GDP ratio and GDP per capita, and a low (but significant) correlation between consumer credit to GDP ratio and GDP per capita.

This could be explained by the fact that richer economies of Northwestern Europe have tended to develop larger mortgages markets – helped by the development of capital markets shaped for the purpose – than the poorer economies of Central and Eastern Europe. In the latter economies, over the last three decades, owner occupancy ratios have also been on average much higher than in Northwestern Europe, hereby limiting the development of mortgage markets. Conversely, in Central and Eastern Europe, the reliance on consumer credit tends to be higher than in Northwestern economies, as it is likely to be more frequently used by segments such as sole entrepreneurs and as regulatory constraints on the price of credit (usury rates, etc.) were often lower (see Bouyon et al., 2018).[23] Therefore, it could be expected that GDP per capita might have a negative significant impact on consumer credit to GDP ratios.

Another control variable concerns the rate of unemployment. The general unemployment rate is used for the consumer credit to GDP ratio. Expectations regarding the sign of the coefficient of these three control variables remain ambiguous. On one hand, higher unemployment should negatively affect the creditworthiness of consumers, resulting in less credit being granted. On the other hand, higher unemployment should result in more arrears and persistent outstanding values of credit. Some unemployed consumers in additional need of funding could even ask for new credits in order to fund existing credit or any other types of expenses they can no longer assume. This mechanism would operate provided that some lenders are willing to provide loans to these borrowers despite higher financial risks of future missed repayments.

Financial control variables are also used to control for the impact of interest rates. For the consumer credit to GDP ratio, this concerns the short-term interest rate (between one year and five years). For the regressions on different groups of consumers, as all household credit is covered, the policy rate is used in order to mirror interest rates practised for any type of household credit. These financial control variables are expected to have a negative impact on

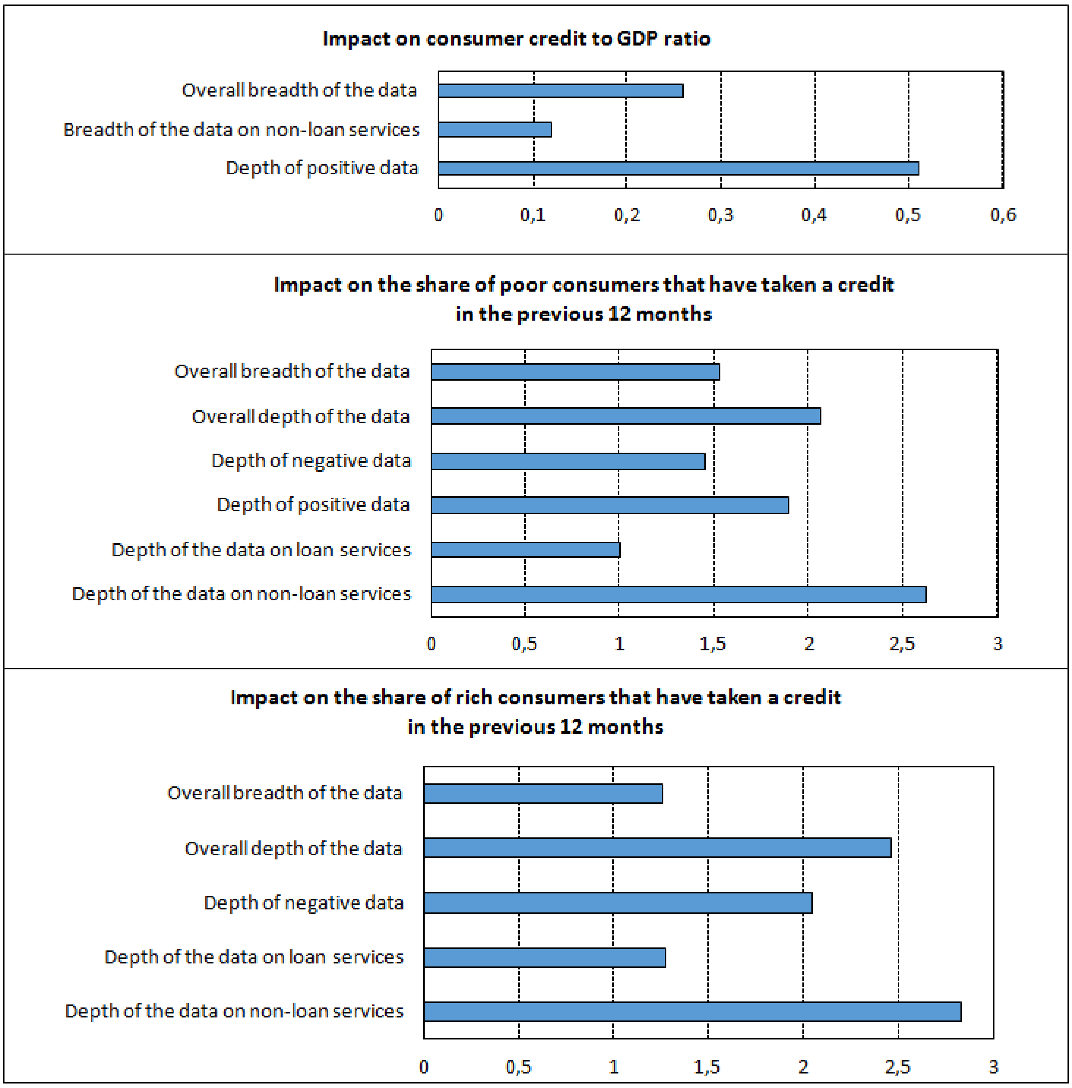
[23] One possibility would have been to use a variable to control for restrictions on interest rates. In Bouyon, such metrics have been developed. However, they were built only for 2008 and 2017, for a smaller sample of EU countries than in the present study. No other similar metrics that could be used in the present study have been identified.

dependent variables. Indeed, as higher rates should depress demand, this most likely should contribute to reducing the dependent variables.

Lastly, the dummies on legal origins are integrated in order to assess whether local rules regarding the rights of borrowers and lenders have an impact on access to credit. As shown by Djankov et al. (2007), legal origins for each country can be a strong determinant of creditor rights.[24]

4.2 Results (see Annex 3)

Figure 1. Impact of higher comprehensiveness on financial inclusion
(increase in percentage points resulting from an increase of 0.1 in the comprehensiveness index)



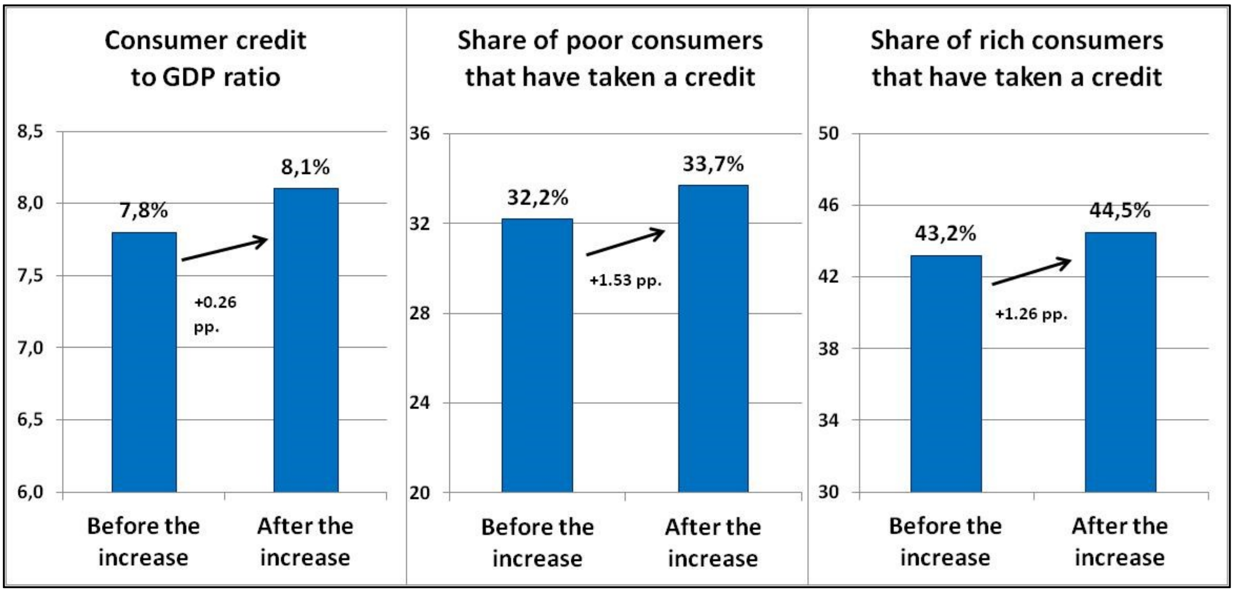
[24] One critique of this approach could be a priori that almost all countries of the sample have to follow EU rules. As such, diversity should not be significant. Nevertheless, a posteriori, many EU financial rules are based on minimal harmonisation. Therefore, member states still have the freedom to have more restrictive rules and develop local legal specificities, likely shaped by their original legal systems.

Notes:

-Only indexes that are robust and significant are considered. The values shown are the average between the two coefficients used for testing robustness.

-The values shown reflect the impact of an increase of 0.1 in a specific comprehensiveness index (for example, from 0.5 to 0.6). The impact shown is the change in percentage points.

Figure 2. Impact of an increase of 0.1 in the index of general breadth on the access to credit



Notes:

-The values chosen for each variable before the increase are the respective averages of the whole corresponding samples.

-The increase of 0.1 in the index of general breadth could for example be reflected by an increase from 0.5 to 0.6. One of the main limitations of this model is to consider a linear relationship between the degree of comprehensiveness and the variables for access to credit. It is likely that the impact might be stronger for lower values of comprehensiveness than for higher ones.

Findings in Annex 3 reveal that all comprehensiveness indexes that have a significant impact on financial inclusion contribute positively to this inclusion. As such, as a result of higher comprehensiveness in the data collected by CRAs, additional lending to safer borrowers more than offsets the lower lending to riskier borrowers.

In particular, higher general breadth and the collection of more structured data on non-loan services both contribute to increasing significantly financial inclusion (see Annex 3 and Figures 1 and 2) for both the poorer and the richer consumers. By considering the average values of the sample (see Table A2 in Annex 1 and Figure 2), an increase of 0.1 in the index for general breadth[25] would contribute to an increase of the consumer credit to GDP ratio, from 7.8% to 8.1% (an increase of 0.26 percentage points). The inclusion of poorer consumers would grow

[25] This increase would imply the collection of data for seven additional items.

on average from 32.2% to 33.7% (an increase of 1.53 percentage points) and the average inclusion of richer consumers would rise from 43.2% to 44.5% (growth of 1.26 percentage points). Considering the whole sample, this would mean that approximately more than 2.5 million additional consumers considered as poor and more than 3 million additional consumers considered as rich would have access to credit markets.[26]

Finally, the depth of negative data boosts all variables for inclusion. As regards positive data, the greater the depth of positive data, the higher impacts the consumer credit to GDP ratio and the greater the inclusion of poorer consumers.

5. Financial intermediation: strategy, limitations and results

5.1 Strategy and limitations

As analysed in Section 1.3, intermediation efficiency is measured through the loan to deposit ratio (LTD). The assumption is that low LTDs imply that banks under-exploit their resources (deposits) and/or use them for other types of activities (investments in capital markets, etc.). In economic theory, an LTD that is too high (often above 100%) is a signal of financial risk, as banks might be unreasonably exposed to certain risks in the real economy. In economic literature (see Beck et al., 2009), it is generally assumed that the higher the LTD, the higher the intermediation efficiency. This proxy could also reflect the degree of financial inclusion. The main limitation of this approach is that it does not capture banks that already had an excessively high LTD and wish to decrease it because of rising financial risks. The model implicitly assumed that the LTD is acceptable or too low. Subsequent research should be conducted to assess a possible optimum.

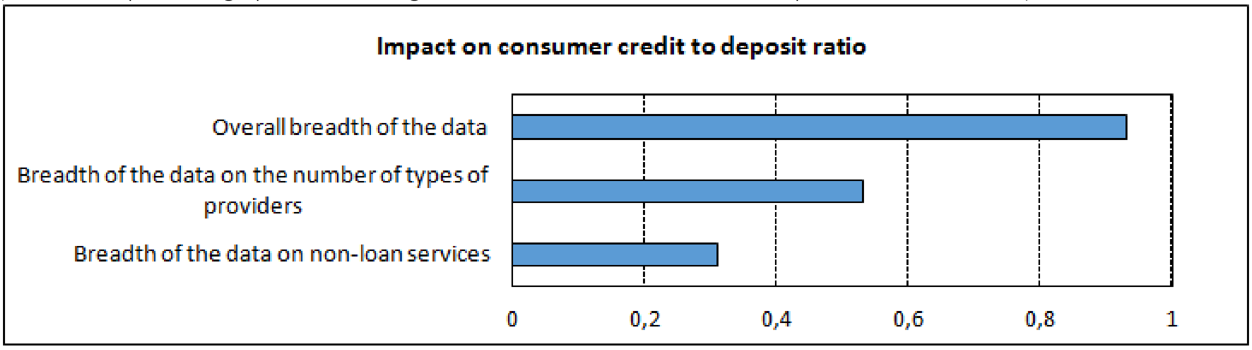
The control variables used are similar to those for the consumer credit to GDP ratio. The expected impact of GDP per capita and legal origin should be broadly in line with that for the consumer credit to GDP ratio. Regarding unemployment, this is a priori undetermined, as a rise in job losses should depress the supply of credit but in the meantime also decrease the volume of deposits. The effect captured by interest rates could be different than it should be for financial inclusion. Interest rates could indeed trigger opposite effects on the credit to deposit ratio. For example, increasing interest rates could depress the demand for credit but boost the demand for deposit. The sign of the coefficient of this variable would depend on whether the former effect more than offsets the latter effect.

[26] Considering the whole sample, this would mean that approximately more than 2.5 million additional consumers considered as poor and more than 3 million additional consumers considered as rich would have access to credit markets.

5.2 Results (see Annex 4)

Figure 3. Impact of higher comprehensiveness on financial intermediation

(increase in percentage points resulting from an increase of 0.1 in the comprehensiveness index)

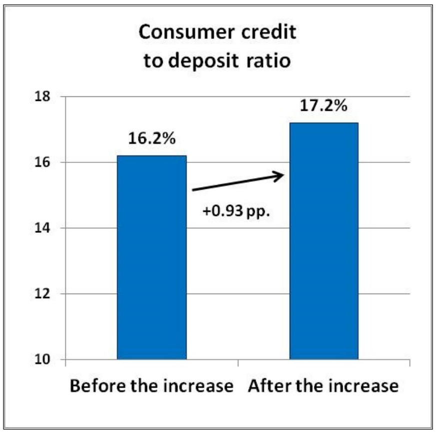


Notes:

-Only indexes that are robust and significant are considered. The values shown are the average between the two coefficients used for testing robustness.

-The values shown reflect the impact of an increase of 0.1 in a specific comprehensiveness index (for example, from 0.5 to 0.6). The impact shown is the change in percentage points.

Figure 4. Impact of an increase of 0.1 in the index of general breadth on the consumer credit to deposit ratio



Notes:

-The value chosen for the variable before the increase is the average of the whole sample.

-The increase of 0.1 in the index of general breadth could for example be reflected by an increase from 0.5 to 0.6. One of the main limitations of this model is to consider a linear relationship between the degree of comprehensiveness and the variable for consumer credit to deposit ratio. It is likely that the impact might be stronger for lower values of comprehensiveness than for higher ones.

The general breadth and the index on non-loan structured data have a robust positive impact on financial intermediation (see Annex 4 and Figures 3 and 4).[27] This shows that the sharing of more comprehensive data increases the share of lending directed to consumer credit compared to other financial activities.

[27] As a result of an increase of 0.1 in each of the two comprehensiveness indexes, the consumer credit to deposit ratio would respectively increase by 0.93 percentage point and 0.31 percentage point.

6. Financial risk: strategy, limitations and results

6.1 Strategy and limitations

Financial risk is proxied by the dependent variable non-performing loans (NPLs) ratio. As analysed in Section 1.1, no granular NPL ratios can be identified at the level of credit products. It is assumed that these general NPL ratios are positively correlated with NPL ratios at product level (consumer loans, housing loans, etc.).

Two outliers have been identified in the sample: Cyprus and Greece. The NPL ratios of these two countries reached tremendous levels in 2015 (47.7% and 36.6%, respectively) and 2017 (31.4% and 45.6%, respectively) as a result of the persistent effects triggered by the 2012 sovereign debt crisis. These figures were much above the average recorded in 2015 (9.9%) and 2017 (7.7%), and in other countries (for both years, Italy recorded the third highest value, at a distant 18.1% in 2015 and 14.4% in 2017). Given the small size of the sample and the risk of distortions that could be caused by these outliers, the two countries have been removed from the sample for the analyses on NPLs.

The core assumption to be tested is whether more comprehensive data can improve the accuracy of creditworthiness assessment and limit the risk of missed repayments in the future. As such, the methodology includes a time lag in order to ensure that the effect of more comprehensive data on NPLs is captured. It is assumed that higher data comprehensiveness in the period before should result in a lower ratio of NPLs, as the risk profile of the pool of borrowers had presumably been improved and the conditions of the credit contract better adapted to the characteristics of each borrower. One key limitation is the limited number of observations resulting from the removal of two outliers and the introduction of a lagged value for the comprehensiveness indexes. Findings should, therefore, be interpreted with caution. Given this limited number of observations, only the indicators of breadth are assessed.

Three control variables are included in the regressions. The first is the unemployment rate and aims to capture the effect of economic fluctuations on the level of NPLs. This should be the most adequate control variable for that purpose given that higher unemployment rates should directly affect the ability of the unemployed to reimburse their loans. Therefore, the expectation is that unemployment has a positive significant impact on NPLs.

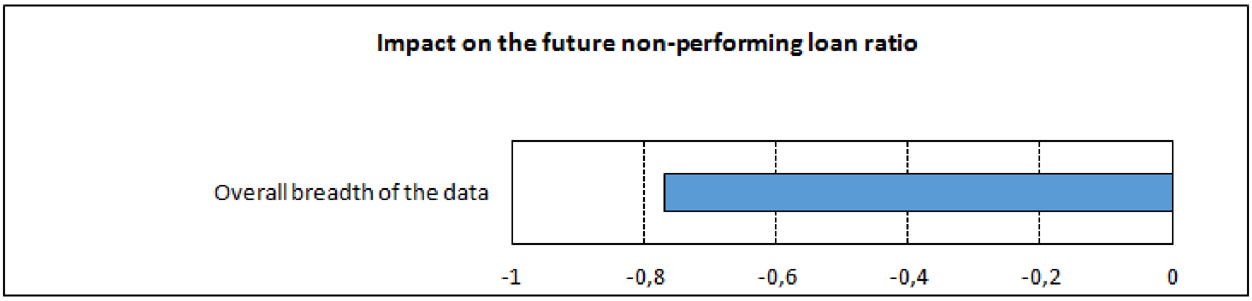
The second variable reveals the cumulative growth in the outstanding real amount of loans (for both households and non-financial corporations) between 2002 and 2012. The effect of past credit growth on NPLs has been used by Jakubik et al. (2013) for Central, Eastern and Southeastern Europe. The authors showed that higher credit growth in the past has had a significant positive impact on NPLs. Similar findings are expected for the present study.

Finally, there has been a long debate in the economic literature on the role of interest rates in the level of NPLs. In order to capture these particular effects, the policy rate of respective central banks is integrated into the regressions. It is reasonably assumed that growth in policy rates result in higher interest rates practised in credit markets. As a large share of loans is denominated with variable rates in many EU countries, it can be expected that rising policy rates tend to rapidly raise interest rates of many existing loans, thereby raising the risk of missed loan repayments and NPL ratios. Therefore, it is expected that the control variable "policy rate" has a significant positive impact on NPLs.

6.2 Results (see Annex 5)

Figure 5. Impact of higher comprehensiveness on the risk of missed repayment

(increase in percentage points resulting from an increase of 0.1 in the comprehensiveness index)

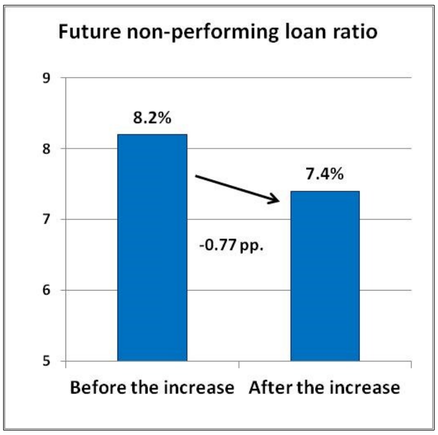


Notes:

-Only indexes that are robust and significant are considered. The values shown are the average between the two coefficients used for testing robustness.

-The values shown reflect the impact of an increase of 0.1 in a specific comprehensiveness index (for example, from 0.5 to 0.6). The impact shown is the change in percentage points. For example, according to the findings, the increase of 0.1 in the index mirroring the general breadth results in a decrease of 0.77 percentage points in the future non-performing loan ratio (for example from 8.2% to 7.45%).

Figure 6. Impact of an increase of 0.1 in the index of general breadth on the risk of missed repayment



Notes:

-The value chosen for the variable before the increase is the average of the whole sample.

-The increase of 0.1 in the index of general breadth could for example be reflected by an increase from 0.5 to 0.6. One of the main limitations of this model is to consider a linear relationship between the degree of comprehensiveness and the risk of missed repayment. It is likely that the impact might be stronger for lower values of comprehensiveness than for higher ones.

As revealed in Figures 5 and 6, and Annex 5, a higher general breadth in the data collected by CRAs reduces the future ratio for NPL. This result links higher comprehensiveness to the improvement of the risk profile of the pool of borrowers.

7. Competition, concentration and legal framework: strategy, limitations and results

7.1 Strategy and limitations

This time, the dependent variables concern the indexes of data comprehensiveness. Five drivers of these indexes have been identified: the barriers to entry proxied by the variable Tier 1 capital ratio; the degree of concentration controlled by the Herfindahl-Hirschman index; the domestic approach to privacy, which is controlled by the dummies on legal origin; the rules that restrict information sharing proxied by the Restriction_Index; and the rules that enable information sharing proxied by the Promotion_Index.

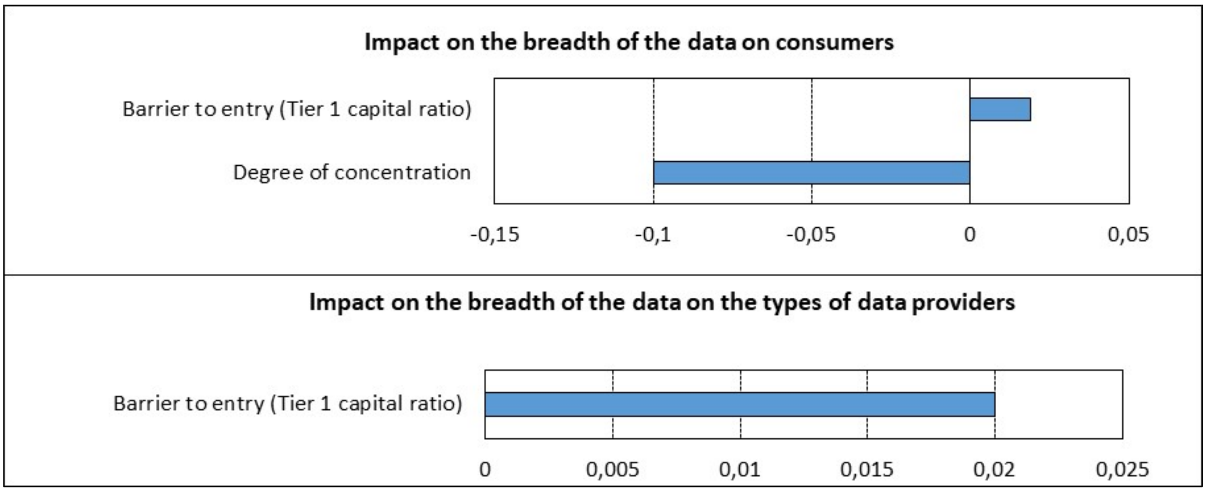
The two first drivers have already been analysed by Bruhn et al. (2013). Regarding the barriers to entry, the rationale is that if the Tier 1 ratio practised in a given country is higher, then the higher constraint on capital requirement makes it challenging for potential new firms to enter credit markets. As regards concentration, the assumption is that more concentrated credit markets tend to limit data sharing.

No relevant publication has been identified to analyse the relationship between privacy standards and legal origin. The intuition is that French and German origins are correlated with higher privacy rights when compared to Scandinavian and UK origins. Finally, as shown in Section 3, the Restriction_Index and Promotion_Index have been developed based on a set of five questions each. The assumption is that the higher the Restriction index, the lower the indexes of comprehensiveness. The reasoning is the opposite for the Promotion index.

The drivers of market concentration and barriers to entry are only related to credit markets. Therefore, their impact should be observed only on data that is available at lender level. Therefore, Breadth_Individual, Breadth_Loan-Services and Depth_Loan-Services should be directly affected by these two drivers. Credit providers might be able to retrieve data on non-loan services. However, it is very likely that only a few of these lenders were doing it and only for a tiny share of the data on non-loan services. Against this background, regressions with Breadth_Services and Breadth_Non-loan-services have little relevance.

7.2 Results (see Annex 6)

Figure 7. Drivers behind the degree of comprehensiveness



Notes:

-Only drivers that are robust and significant are considered. The values shown are the average between the two coefficients used for testing robustness.

-The values shown reflect the impact of an increase of 0.1 in the degree of concentration and of 1 point of percentage in the barrier to entry (mirrored by the Tier 1 capital ratio). The impact shown is the change in the specific index of comprehensiveness. For example, according to the findings, the increase of 1 percentage point in the Tier 1 capital ratio results in the increase of 0.02 in the breadth of the data on consumers. In addition, an increase of 0.1 in the degree of concentration results in a decrease of 0.01 in the breadth of the data on consumers.

Regressions in Annex 6 confirm some of the findings of Bruhn et al. (2013). First, it is revealed that established lenders are more willing to share data on their consumers and loan services when barriers of entry on credit markets are high and the risk of competing with new lenders is limited. In a context of high barriers of entry, the perception of existing lenders is that the sharing of their data with CRAs is less likely to facilitate the emergence of completely new credit providers.

In addition, a higher concentration in credit markets limits data sharing. The combination of high barriers of entry and low concentration in credit markets is, therefore, the optimal configuration for boosting data sharing overall.

Last but not least, laws promoting the sharing of data seem to have a positive significant impact on the breadth of the shared data.[28]

[28] Regressions on the role of restrictive laws did not reveal any specific mechanism. The absence of relevant findings might be explained by the ambiguity of some of the questions covered by the index on restrictive laws (see for example in Table 2 the questions a, b and c of the Restriction_Index).

8. Conclusion

Empirical findings reveal that the general breadth in the data collected by CRAs contributes to:

- increasing financial inclusion for consumers with both lower and higher incomes
- increasing the share of lending directed to consumer credit compared to other financial activities
- reducing the risk of consumers missing future repayments.

The collection of more comprehensive structured data on non-loan services significantly increases financial inclusion for rich and poor consumers, and the attractiveness of consumer credit activities. However, this last finding concerns a relatively small share of non-loan data. The other types of non-loan data will be analysed in Chapter 2.

Finally, the collection of more granular positive data boosts the inclusion of poorer consumers.

As regards the drivers behind the degree of comprehensiveness, results show that a lower degree of concentration, higher barriers to entry and a higher degree of laws promoting data sharing all contribute positively and significantly to the collection of more comprehensive data by CRAs.

Chapter 2.
Non-traditional data and its potential use for creditworthiness assessment

Introduction

The previous chapter investigated the level of detail of the data currently held by the European Credit Reference Agencies (CRAs) and its associations with macroeconomic indicators of financial inclusion, intermediation and risk. This chapter investigates the definitions and types of non-traditional data, its benefits and challenges within the context of the decision-making process of whether to give credit to an applicant.

The use of such data for credit risk assessment is relatively limited, and the definition of what is non-traditional data varies from country to country.[29]

The main focus of this chapter is on financial inclusion, understood as the chance of an individual or a group of individuals being accepted for credit. The chances of any given customer being granted credit depend directly on the ability of a lender to correctly predict the future credit performance of borrowers. Therefore, predictive accuracy is fundamental for increasing chances of being granted credit and improving financial inclusion. One of the aspects underlying predictive accuracy is the data that is used – its quality, strength of association with the credit performance, diversity in capturing various facets of behaviour related to credit. Traditional data may not capture all relevant details of a credit applicant's behaviour, which may be contained in the non-traditional data. This data may be particularly helpful to individuals with no or little previous credit history, such as underbanked and non-banked segments, young people, migrants (the so-called 'thin file' population). These individuals may be rejected for credit not because they are bad credit risks, but simply because a lender may not have the data to assess their risk level.

Despite huge interest in the potential use of non-traditional data, little is known about it, and the current chapter fills in this gap by reviewing the academic studies and public reports and by conducting qualitative interviews with relevant stakeholders.

The findings suggest that non-traditional data is seen as a potentially useful resource for improving the accuracy of credit risk assessment and increasing financial inclusion, especially for 'thin file' borrowers. Open banking data, in particular, is perceived as most valuable. However, there are challenges that need to be addressed before the use of non-traditional data becomes widespread, the most notable being legal compliance.

The rest of the chapter is structured as follows. Section 9 introduces the main concept and literature review. Then, Section 10 presents the views of different stakeholders on the use of non-traditional data types. Finally, Section 11 explores the requirements, benefits and challenges associated with the use of non-traditional data.

---

[29] The definitions and types of non-traditional data will be discussed later. As noted above certain types of credit data may be regarded as traditional and be used in some countries, whereas the same data will be novel in other countries. Section 10 provides more details.

9. Main concepts and literature review

9.1 Methodology

Given the lack of publicly available quantitative data that could be analysed, this chapter employs the methodology of literature review and qualitative semi-structured interviews with relevant stakeholders.

Qualitative research (in contrast to quantitative methods used in Chapter 1) is concerned with exploring and understanding phenomena rather than measuring them; the research findings are not subject to quantification or quantitative analysis. Qualitative methods of enquiry are best suited to clarifying the issues under investigation, generating hypotheses, exploring and explaining a range of different opinions.

Qualitative interviews, unlike quantitative techniques such as surveys, focus on an individual rather than populations and groups. It should be noted that the qualitative interviews are conducted on small samples that do not allow for generalisations.

Eight qualitative interviews were conducted with representatives of different stakeholders to ensure that a wide range of opinions was covered. The respondents include a mix of financial institutions and representatives of consumer organisations so that different points of view could be captured. The table below gives the types of organisations and positions of the respondents together with their ID codes that will be used in the subsequent analysis. The names and countries cannot be used for confidentiality reasons.

Table 5. Summary of respondents' backgrounds and ID codes ("C" stands for Consumer Association; "F" for Financial Institution).

ID Code	Geography	Organisation	Role
C1	National	Consumer Association	Research & Advocacy Advisor
F2	National	Fintech Start-up	Head of Credit Risk Unit
C3	National	Consumer Association	Administrator
F4	International	Retail Bank	Retail Credit Risk Manager
C5	EU	Consumer Association	Senior Policy and Advocacy Manager
F6	National	CRA	Senior Manager
F7	International	CRA	Strategic Programme Consultant
F8	International	Retail Bank	Director of Analytics

9.2 Creditworthiness, credit scoring and the data it uses

As defined in the Glossary, creditworthiness is usually assumed to be a broad concept referring to whether a particular individual should be granted credit or can be regarded as an attractive/acceptable customer by a particular lender (Thomas et al., 2017). Assessing creditworthiness before granting credit is seen as a key component of responsible lending. Assessment typically evaluates the risk a borrower will not make repayments (credit risk), and the risk a borrower will not have sufficient disposable income to make repayments (affordability risk).

This chapter also refers to 'credit risk assessment', which is part of establishing creditworthiness and consists of estimating the probability of default (or how likely it is that a credit applicant will repay the debt on time under the agreed terms). However, since almost all accept/reject credit decisions are made on the basis of credit risk assessments, in this chapter these two terms are used interchangeably.

Risk assessment in retail credit is almost exclusively done with credit scoring (Andreeva & Matuszyk, 2019). In a nutshell, a credit scoring model is developed on a sample of previous borrowers, for whom the repayment performance is already known. The model links the performance with the borrower's characteristics that can be observed at the point of credit decision (e.g. whether to accept an applicant). The model is then applied to future customers and produces an estimate of their performance (e.g. probability of default).[30] The estimate is compared to the predetermined threshold or cut-off which corresponds to the level of risk acceptable of a lender. If the estimated risk is higher than the acceptable level, the application is rejected; otherwise, it is accepted. The process is relatively easy to automate and has been used widely.

The literature on credit scoring or risk assessment in retail credit is less extensive compared to the number of studies on corporate credit risk. However, an interested reader can refer to a brief overview of the credit scoring history and techniques by Thomas (2000) and for more detailed comprehensive texts to Anderson (2007) and Thomas et al. (2017).

Based on the literature review, the data traditionally used for credit risk assessment by lenders includes (Thomas et al., 2017; FICO,[31] 2018):

- application data (supplied from the application form, e.g. residential status, occupation, age);
- behavioural data (monthly data from statements, e.g. amount repaid, outstanding balance);
- credit bureau/CRA data (performance over all credit products held, linked/financially associated accounts).

Previous credit history is believed to be highly predictive of future credit performance (Anderson, 2007; Thomas et al., 2017), and it is this type of data which is traditionally collected by CRAs and which is the subject of this report, particularly of Chapter 1, which

[30] CRAs typically estimate the opposite – probability of repaying on time – therefore higher CRA or credit bureau scores correspond to better creditworthiness.

[31] FICO (Fair Isaac Corporation) is an international credit data analytics company, the first developer of credit scores.

provided a detailed overview of different levels and types of this data. The three studies (two from the US and another from the EU context) are reviewed in Appendix 1 and support the results of Chapter 1, which generally demonstrate that the more detailed the CRA data, the greater the financial inclusion.

The first US study (Chandler & Johnson, 1992) shows that up to 57% of predictive accuracy of scoring models comes from the CRA data, and the accuracy improves with greater level of detail. The second US study (Barron & Staten, 2003) confirms that the level of detail of CRA data lowers the level of risk and increases financial inclusion. It also demonstrates the value of positive CRA data, where an up to 47.5% increase in financial inclusion is estimated as a result of adding positive data to negative CRA data. Similarly, in the European context, Andreeva (2004) and Andreeva et al. (2008) demonstrate that the best predictive accuracy is achieved when CRA data is included in credit scoring models.

Prior research has also reported on the importance of credit and non-credit account behavioural data, which is available within the same bank from different accounts. Mester et al. (2004) conduct such an investigation for small firms in Canada, whilst Norden & Weber (2010) and Puri et al. (2017) explore a similar type of data for individuals in Germany. These studies conclude that behavioural data is statistically significant in explaining the default probability, but they do not measure its contribution to predictive accuracy.

More recently the focus has shifted to FinTech and BigTech lending, which is predominantly conducted via online platforms with heavy use of technology. Using the data from Prosper, one of the largest US platforms, the following studies explore the statistical association of the non-credit types of data with funding success and default risk. Lin et al. (2013), Freedman & Jin (2014), and Liu et al. (2015) find that borrowers' online friendships and social networks have significant associations with lower default risk. Duarte et al. (2012) analyse borrowers' profiles and find that 'trustworthy' or 'attractive' ones have a better chance of being funded and a better chance of avoiding default.

Larrimore et al. (2011) show the effect of the quality and quantity of a loan's textual description on funding success. Iyer et al. (2016) find that soft or non-standard information, e.g. appearance, acceptable maximum interest rate and textual description, are significantly associated with default and improve prediction of default probability as compared to predictions based only on borrower's credit rating. Gao et al. (2017) also confirm the importance of textual information by showing that readability, positivity, and objectivity of loan description texts are related to loan defaults. However, in a European context, Dorfleitner et al. (2016) do not find a significant association of textual information with defaults, only with the probability of getting funded.

In European BigTech, Berg et al. (2018) explore the value of web browsing/digital footprint using account-level data from an e-commerce company located in Germany. They claim that web browsing/digital footprint variables could be equally or even more predictive compared to traditional credit bureau scores. The authors also explore the predictive value of variables from web browsing/digital footprint for customers with no bureau scores ('unscorable' customers). The authors state that the predictive power of these variables for 'unscorable' is comparable to that of the credit score for 'scorable' customers. However, it should be noted that the authors only use a generic bureau score and do not use detailed CRA information, which is known to be more predictive than a single credit score (Chandler and Johnson (1992) reviewed in Appendix 1). Besides, the authors do not use an out-of-time model validation, which in the context of their study would ensure a fair comparison of the credit score and digital footprint. Nevertheless, these results show the potential of web browsing data in increasing the access to credit, especially for the 'thin file' customers.

Not only FinTech and BigTech lenders experiment with novel data. There are studies reporting experiments based on traditional banking data. Oskarsdottir et al. (2019), using the data from an undisclosed country (for confidentiality reasons), combines credit and debit account information of customers with information from their mobile calls. Call-detail records are used as separate characteristics and as inputs into social network analytics techniques to track the impact from prior defaulters to produce influence scores. The best predictive accuracy is achieved when all measures are combined. Liberati & Camillo (2018) consider financial behaviour and psychological traits of customers of an Italian bank to predict the credit performance. The results demonstrate a notable improvement in predictive accuracy when personality traits are included in models.

There are also studies that explore the value of non-traditional information in developing economies that are not reviewed here, because these results may not be directly applicable to the European context with mature banking and credit reference systems. Indeed, Frost et al. (2019) using cross-sectional international macroeconomic data demonstrate that the share of BigTech lending is higher in countries with less competitive banking sectors and less stringent regulation. Focusing on SMEs in Argentina, the authors conduct an account-level analysis using the data from a BigTech lender and conclude that BigTech lenders have an information advantage in credit risk assessment relative to a traditional credit bureau in Argentina.

In the majority of these studies, the best results are achieved when different types of data are included in risk assessment. However, not all academic studies reviewed in this section evaluate the contribution of different types of data to predictive accuracy. Therefore, academic research is complemented with industry reports and practical examples in subsequent sections that combine results from the interviews with some observations from the literature on non-traditional data in credit risk assessment.

9.3 Definitions

In general, the term 'non-traditional data' is used interchangeably with 'alternative data', especially by international organisations, such as the World Bank or International Committee on Credit Reporting (ICCR, 2018), which prefer the term 'alternative'. In order to better understand what constitutes 'non-traditional', first, it makes sense to define 'traditional' data.

In addition to the definition given in the previous section, Table 6 summarises the definitions provided by respondents in qualitative interviews. Throughout this chapter, the authors emphasise in bold the most common and important themes.

Structure comes out as the most important distinctive feature of this type of data, mentioned by almost all respondents. It refers to the shape of the data, how they are stored and the degree to which they can be analysed with existing software. The respondents emphasised accuracy, precision, and the impossibility of distortion, i.e. data subjects cannot manipulate their records. Another distinctive feature refers to the relatively limited and manageable amount of data.

As for the different types of traditional data, the most important aspects fall under the financial or credit domain (payments, income, various aspects of credit behaviour). Personal characteristics are also mentioned.

Table 6. Definition and types of traditional data

Definition	*Types*
Structured (C1, F2, C3, F4, F7, F8) - can be analysed by traditional software and techniques (F2) - **well-defined**, stored in the company's database (F4) - a **limited** amount of… (F4) - **predictive, accurate, comprehensive** but **proportionate** data (F7) - any data that resides in a **fixed** field within a record or file **organised** in rows and columns (F7) - **official data** that cannot be **distorted (C1**) - depends on the **nationa**l context (C5) - may be different for creditworthiness vs. risk assessment (C1) - minimum **data necessary to grant a loan (C5)**	- **personal** data/socioeconomic characteristics, **behavioural** data, interest bearing assets, **financial statement** - **credit** bureau/previous **credit** history/external **credit** scores (positive data e.g. **credit** use, length of **credit** history, **credit** activity, settled **payment** disruptions; in certain countries: utilities, phone contracts/obligations; negative data on consumer payment disruptions/adverse information) - answers to **credit** questionnaire - analysis of **income**, **spending** on current accounts - identity, etc. - composition of households - **arrears or defaults** (negative data) - proof of **income**

For non-traditional/alternative data, most publicly available information refers to developing countries and the US. The interest in this topic in developing countries has arisen because of the scarcity of traditional credit data, therefore alternative data have been regarded as a potential replacement. In general, the motivation behind the existing research into alternative data has been prompted by financial inclusion. There is no universal definition, since the state of development of CRAs and financial inclusion differs from country to country. However, it is possible to draw on some examples proposed by international bodies and commercial organisations:

- "A generic term that designates the massive volume of data that is generated by the increasing use of digital tools and information systems." (GPFI: Global Partnership for Financial Inclusion, Source: GPFI, 2018).
- "Data serving to describe ways to collect and analyse data on creditworthiness, which are "alternative" to conventional methods such as documented credit history. Alternative data is also considered to be information readily available in digitized form that is collected through technological/electronic platforms." (ICCR: International Committee on Credit Reporting, Source: ICCR, 2018).
- "Information used to evaluate creditworthiness that is not usually part of a credit report." (CFPB: Consumer Federal Protection Bureau, US).[32]
- Information that is not found on traditional credit reports (LexisNexis).[33]
- "... Additional financial payment information on consumers or otherwise information with predictive power." Utility, telecommunication, rentals, asset records, alternative lending payments and demand deposit account information. (Oliver Wyman, Source: Carroll & Rehmani, 2017).

Apart from the most common 'non-traditional' descriptor, there are some other key words that are often repeated and that emphasise different aspects of this data – additional, massive, digital.

Table 7 summarises the definitions from the interviews that echo the key themes arising from the documents above. Again, the most prominent definition is anything that goes beyond traditional data, i.e. 'all the rest', with additional aspects of being unlimited and unstructured.

Table 7. Definition of non-traditional data

All the rest (**C1, F2, C3**)
Unlimited, **unstructured**, minimising the risks associated with human intervention (F4). **Going beyond** checking the existing financial commitments in terms of credit (positive data), arrears/defaults (negative data) and proof of income (pay slips, employment contracts) (C5). **Additional** data to that listed above that provides additional **predictiveness** when used in combination with traditional data, or some predictiveness when used without traditional data. The data must also be **accurate** and may be structured (e.g. telephone usage data) or **unstructured** (e.g. social media, transactional data, web data, text, audio, video and images) (F7).
Regional data, macroeconomic data, social media, digital footprint, textual data (F6).

32 https://www.consumerfinance.gov/about-us/blog/using-alternative-data-evaluate-creditworthiness/

33 https://risk.lexisnexis.com/insights-resources/white-paper/modeling-blending-traditional-and-alternative-data

9.4 Types of non-traditional data

Similar to the definitions of non-traditional data, the typology also varies from country to country. In some countries, CRAs collect certain types of data (e.g. utilities and mobile phone obligations/contracts/payments) that would be regarded as non-traditional in other countries. In this report, we follow the typology proposed by international organisations (in particular, GPFI and ICCR), in order to make sure that all countries with different levels of data comprehensiveness are taken into account. As a robustness check, we also verify this typology against the one used by FICO.

Although non-traditional data is generally perceived as unstructured, ICCR (2018) distinguishes the two main categories within it: structured and unstructured non-traditional types.

Structured data:

- "Information with a high degree of organization, such that inclusion in a relational database is seamless and readily searchable by simple, straightforward search-engine algorithms or other search operations" (ICCR, 2018);
- Relates to the "ability to repay" debt and is usually the most useful.

The types of structured non-traditional data include:

- Data on payments (e.g. utilities, mobile phone, and certain other obligations like rental information, taxes, etc.);
- Crowdfunding transactions, factoring, leasing and credit insurance;
- Transactions from P2P lending platforms, invoice, accounts payable, sales volume, merchant transactional data, mobile/e-money, procurement, historical business cash flows, shipping history, bills of lading, and data from online accounting platforms;
- Data associated with assets (movables and fixed);
- Payment flows received by disadvantaged individuals (e.g. subsidies, pensions, domestic and cross-border remittances, etc.).

Unstructured data:

- "[I]nformation that either does not have a pre-defined data model and/or is not organized in a predefined manner" (ICCR, 2018);
- Can be useful in cases of first-time borrowers with no or thin credit histories.

The types of unstructured non-traditional data include:

- Social media and internet usage;
- Emails;
- Text and messaging files;
- Audio files;
- Digital pictures and images;
- GPS data;
- Mobile usage (how many calls to the same number, peak usage, etc.);

- Other meta data;
- Psychographic, psychometric and other non-financial behavioural data.

The definitions used by the credit industry for non-traditional data are very close to the point of view of the international bodies above, e.g. FICO (2018) lists the following types of non-traditional data:

- Transactional data (e.g. disaggregated detailed transactions for card payments or from current accounts);
- Telecom/rent/utility data (;
- Mobile data (information in term of calls, messages, apps used, etc.);
- Social profile data (from social media, using only information on a particular individual);
- Social network data (from social media connections/friends of the individual);
- Survey data;
- Psychometrics;
- Clickstream data (web browsing, navigation of the lender's website);
- Audio/text data (from application notes, customer service calls, collections, etc.).

10. Views on the potential use of non-traditional data types

It should be noted that several types of structured non-traditional (or non-bank) data (such as utility or mobile contract payments) are already being collected in certain countries and therefore appeared in Chapter 1. Yet the typology adopted in this chapter is broader, since Chapter 1 refers to what is being collected, whilst Chapter 2 is forward-looking and explores what may or may not be collected in future. Even if some types can be considered semi-alternative data, they are still included here for completeness, because they are not collected in all EU countries.

The types of non-traditional data listed above have been used as an input for the interviews, where respondents were asked to indicate which types could be useful for credit risk assessment. The answers are summarised in Table 8.

'Open banking' is an undisputed leader, with all respondents believing it will be useful. However, there are still unbanked or underbanked consumers even in developed EU countries. According to the recent study based on information from Global Findex by the World Bank, the estimated number of unbanked adults in the EU is over 37 million, with the percentage of total adult population ranging from 0% in Denmark and Finland to 39.2% in Romania.[34] It is possible that for such customers with limited banking transactions unstructured data will be useful, although unstructured softer measures are generally perceived as less useful, compared to structured (or payment) types.

[34] www.wsbi-esbg.org/press/latest-news/Pages/Close-to-40-million-EU-citizens-outside-banking-mainstream.aspx.

Table 8. Views on usefulness of non-traditional data for credit risk assessment

Type of non-traditional data	Examples	Can be useful for credit risk assessment
Open banking data	Transactions from different accounts and financial institutions	8
Non-loan payments	Utilities, mobile phone, and certain other obligations like rental information, taxes, etc.	6
Assets and payment flows received by disadvantaged individuals	Subsidies, pensions, domestic and cross-border remittances, etc.	6
Transactions from P2P lending platforms	Information on loans from P2P alternative lenders	5
Crowdfunding transactions	Information on loans from crowdfunding alternative lenders, where funds are collected from many investors	4 + 1 may be
GPS data	Geo-location data, places visited	4+1 may be
Social media	Posts, likes, connections on Facebook, twitter, LinkedIn	4 +1 may be
Emails	Text and patterns of writing/sending	4+1 may be
Text and messaging files	Text and patterns of writing/sending	3+1 may be
Mobile usage	How many calls to the same number, peak usage, connections to other contacts	3 +1 may be
Audio files	Recordings of telephone calls	2+2 may be
Digital pictures and images	Images from various sources	2+2 may be
Psychographic, psychometric	Special measures of personality	3+1 may be
Internet usage	Clickstream data, browsing behaviour	3+1 may be
Data from wearables, Internet of things	Data from sensors	3+1 may be

Note: The numbers indicate how many respondents supported a potential use of a certain type of data. These numbers should be interpreted as relative preferences, not as representing a general level of support.

When asked whether anything else (not listed in the table) may be useful for credit risk assessment, the following suggestions were made:

- credit files of neighbours for consumers with no credit (noting a possible risk of discrimination);
- 'administrative status' of the consumers (are they under administrative supervision, as not responsible for taking some decisions?).

One respondent expressed strong concerns about the use of all types of non-traditional data (except for open banking) because from her/his point of view they are unlikely to add any value:

"No ..., it is simply because an analysis of the consumers' budget management is largely sufficient to decide whether or not he/she is creditworthy" (C5).

In addition to the views above on the potential value, financial institutions were asked if they already use any non-traditional data. As mentioned previously, payments data is already being used, and all respondents confirmed this. Similar to views on usefulness, structured non-traditional data is preferred, possibly because of the relative ease of processing and logical connection to affordability and credit management. Some respondents have also reported using open banking data and P2P transactions. Unstructured data is not being used; some of its types are only considered (or have been considered) for potential use by one respondent. Two respondents are testing social media for firms only, which is legal because such information refers to business entities, not individuals.

11. Requirements, benefits and challenges of non-traditional data

11.1 Requirements for non-traditional data

ICCR (2018) identified the following criteria or requirements for sources of non-traditional data, which can be equally applied to the data itself:

- Coverage: the data must be available to the majority of the population;
- Regulatory compliance: the data should conform to the existing regulations;
- Predictive power: the data must be predictive of credit behaviour, i.e. there should be a high degree of association;
- Orthogonality: the data should provide novel information that is not contained in existing traditional data;
- Accuracy and timeliness: the data should be accurate, and should its accuracy be validated, it should be frequently updated;
- Depth of information: the data should provide sufficient detail about an individual.

These requirements are echoed by responses from the interviews, with legal compliance issues being given most attention.

Table 9. Requirements for non-traditional data, opinions from interviews

Legality (consent, notification, purpose of use) (F7); **compliance** (GDPR); depending on the country; beyond compliance stricto sensu, to prove that the use of any data is **legitimate** (to remove fear) (F8)
Using the filter of **GDPR** (C1)
Availability (F6, F7)/historical availability (F6) **Relevance** (F6, F7)/personal reference (F6) **Predictive**, fairness/accuracy/coverage (F7)
Technical requirements: Software that can handle data storage and analysis (F2) A highly scalable analytics process, without a degradation in performance Increased robustness Real time results Flexibility Scale-out storage systems (F4)
Detecting fraud (C3)
No requirements as it should not be used (C5)

11.2 Predictive value

The most fundamental question when it comes to the value of non-traditional data is its predictiveness – if it is not predictive, there is no point in using it. The importance of predictiveness is also confirmed in interviews touching on benefits later in this chapter (Table 12), which is why this criterion is discussed in a separate section. A related concept is orthogonality. The data may be predictive on its own but its informational scope may largely overlap with traditional data.

Due to the confidential and sensitive nature of credit and debt, the evidence related to the predictive value of non-traditional data is patchy and incomplete. This can be explained by the fact that it requires analysis at account level, i.e. the model should be built on real credit accounts of individuals, which would include their repayment performance and different types of data that could be used to predict this behaviour. Building such a model requires access to credit accounts that are not public, hence the scarcity of studies in this area.

Given that such a model can be built, the contribution of different types of information can be measured. One of the most common measures is the Gini coefficient that ranges from zero (no predictive value) to one (perfect predictive performance).

Below is the information from the FICO analysis of the predictive power of different types of data (FICO, 2018). The results suggest that traditional data is still a more powerful predictor, but non-traditional data can improve predictive performance when combined with traditional characteristics. Values in the table below are relative indicators, as many parameters will affect them, such as the predictive power of existing models, the strength of the customer relationship with the lender, amount of traditional data available, etc.

Table 10. Expected improvement from non-traditional data for credit risk assessment

Types of data	Strength of relationship with the customer/ amount of traditional information		
	New	Weak	Strong
Traditional application data	Baseline	Baseline	Baseline
Transactions on debit/credit cards	N/A	0-5%	5-10%
Utility & rental payments	5-10%	5-10%	0-5%
Social media profile data	0-15%	0-15%	0-5%
Social network / friends data	10-20%	10-20%	5-15%
Clickstream data/ website navigation	0-10%	0-10%	0-5%
Text data	N/A	0-5%	0-5%

Source: Adapted from FICO (2018)

In our interviews, respondents who are potential data users were also asked if the non-traditional data should be used together with the traditional or as a sole input, and what increase in predictive accuracy could be expected. Most financial institutions indicated that traditional is still more powerful, and therefore non-traditional data should be used in combination with it. However, it depends on the country and portfolio. In cases where traditional data is less powerful, the expected increase should be higher. In cases where traditional data is limited or not available, the non-traditional data may be used on its own.

Question: Expected increase in predictive accuracy?

- Currently, traditional data have stronger predictive accuracy in the models (around 70%). Adding non-traditional data increase accuracy by 3-4 percentage points (F2).
- In case of low performance of the model, >10%, whilst in case of medium-high performance of the model, <10% (F4), up to 10 Gini points (F6).
- Depends on the scenario (types of data, country, the profile of consumer etc.) (F7).

11.3 Benefits and challenges

We also asked all respondents to list the three most predictive types and three most orthogonal types, starting with the most predictive/orthogonal. In terms of most predictive, financial and structured 'harder' information (open banking, payments, assets) is perceived as more powerful compared to unstructured 'softer' types (social media). However, in term of orthogonality, i.e. information which is not contained in traditional data, the order reverses, with social media and web browsing behaviour being mentioned as most valuable. This distinction is important for use in combination with traditional data; most novel and unexpected insights into the customer's behaviour are likely to come from the unstructured data.

Table 11. Which three types of non-traditional data are …

most predictive?
1. Open banking | Cash flow | Transactional data | Payments | Assets | None 2. Payment Cards | Q-refilling patterns | Regional data | Assets and payment flows received by disadvantaged individuals 3. Social media | Social media data | Positive data from non-banks
…most orthogonal to traditional data?
1. Web browsing | social media | None 2. Social media activity | digital footprint 3. Call data, SMS data | open banking data
…most problematic from legal standpoint?
1. SMS and call data | Social media | Images | Psychometric data | All of them. The most problematic issue: they are not accurate. People misrepresent who they are online, and if they know that this data is used to sell to them, they will openly learn to manipulate it (it is much easier to manipulate GPS data or social media behaviour than to manipulate your spending patterns on your bank account) 2. Web browsing | Open banking data (banks are not really willing to share this data with other banks) | Ethnicity | Audio 3. Location | Assets and payment flows received by disadvantaged individuals
…most difficult to collect?
1. SMS and call data | Social media | positive data from non-banks | Images | Open banking | Payments 2. Apps installed | Open banking | Audio 3. Web browsing | Assets and payment flows received by disadvantaged individuals

As for the benefits associated with non-traditional data, predictive accuracy and quality of credit decisions are the most prominent ones as seen by respondents. Holistic analysis or review of the customer (or KYC, 'know your customer') is also emphasised as one of the benefits arising from non-traditional data.

Table 12. Benefits of the non-traditional data

What benefits are provided by the non-traditional data?
The main benefit is that more data is always better than less data (F2)
1) Ability to conduct **holistic** analysis of different data sources (F4) 2) Increase model **accuracy** (F8) 3) More **efficiency** in model development process (increased speed, leaner processes) 4) Increase the financial **inclusion**, i.e. the number of segments of population covered by the PD models (e.g. new clients) 5) Increase the loans to consumers with **little credit history** or to those that banks formerly would have overlooked (F4)
Robustness; increase of predictive **accuracy; holistic** view (F6)
Improved credit risk assessment (including fraud prevention) when limited or no traditional data is available. This can lead to **increased credit availability** for the lower risk consumers. This in turn could lead to **increased financial inclusion** (F7)
Cross border data availability for some non-traditional data (F7)
Open banking data can raise markedly the **quality and accuracy** of creditworthiness assessments (C1)
Excluding open banking, this is mostly about **KYC** (C3)

Comments are also made that benefits will be more pronounced for specific segments, in particular, new clients and customers with little credit history. This is consistent with the published research that reports the results from developing countries, where the majority of customers do not have much credit history due to a low level of development of the banking system and CRA (e.g., Frost et al., 2019). It is also consistent with the German research by Berg et al. (2018), both studies have been reviewed earlier in section 9.3. The results show the potential of web browsing data for access to credit for the unbanked. The best prediction, however, is achieved by combining the traditional scores with non-traditional variables, which confirms findings from the interviews on the combination of traditional and non-traditional data in Section 10.

Further, in the US context, a study conducted by Experian (2018) concluded that alternative data is crucial for 'thin file' consumers that are estimated to constitute 25% of the American adult population. In Experian's lender survey from 2019, two-thirds (66%) of lenders reported declining more than 5% of applicants due to insufficient credit history. In the consumer survey conducted by the same company, 61% of respondents believe that adding payment history would have a positive impact on their credit score.

In our interviews, the respondents from financial institutions confirm that non-traditional data will improve the situation for those with bad/low scores, those excluded from the traditional financial system, and small businesses. There is a view that such data will help all customers, since it will provide better, more complete ('holistic') information about them, and this will open access to more services (Table 13).

At the same time, some respondents from consumer organisations do not see value in non-traditional unstructured data, and are concerned that the use of such data will only reinforce existing inequality. It should be noted that the quotation referring to Lenddo is not supported by the information provided on the company's website – Lenddo is using non-traditional data for scoring.[35]

Table 13. Benefits to specific segments of customers

What type/segments of customers/population would benefit most from the non-traditional data? In what way?
Those who would receive a **bad credit score** when using only traditional data *(* (F2)
SME Retail segment (F4)
Increase of predictive accuracy for **thin data, currently low-scored** persons can increase their score due to the **holistic** view (F6)
All types (C1)| **All** segments (F8)
Different needs across segments (F8)
In particular, the ones who have been **excluded** from mainstream credit so far (C1)
'**No file**' or '**thin file**' consumers. The **unbanked** population (F7)
Lenddo in the Philippines tried to use same [non-traditional] type of data to conduct creditworthiness assessment but it did not work (C3)
None. But all customers except the very wealthy will suffer from using non-traditional data … this will help extract more money out of the most vulnerable and ensure the best prices for those who do not need credit. It is a way to completely kill a minimum of risk socialization/mutualization (C5)

The theme of potential risks and challenges associated with non-traditional data continued in the next question (Table 14). Legal compliance stands out as the topic that causes most concerns. Additional issues include data quality and potential bias, because of potentially inaccurate, incomplete data that may be even manipulated by data subjects in some cases. It is interesting that IT problems are being mentioned, but they are overshadowed by themes of legal compliance and social acceptance.

A topic that deserves a special attention is the increasing power of GAFA (Google, Amazon, Facebook, Apple). Big Tech companies continue to accumulate a lot of non-traditional data, thus gaining perhaps disproportionate control over access to such data. Some of them (e.g. Amazon) offer credit themselves, taking advantage of the unique data and knowledge of their users. Concerns have been expressed that these companies do not have a sufficient level of regulatory scrutiny.

[35] www.lenddo.com/products.html#creditscore.

Table 14. Challenges and risks

What are main challenges/risks associated with the non-traditional data (overall, and for any specific types)?
The legal issues like **GDPR** (F2)| Not to be compliant with **GDPR** provisions (C1)
1) **Biased** data; 2) **IT** infrastructure-related problems; 3) Data **quality** (potentially) (F4)
Possible **biases** in the use of the data and the way models are shaped (F8)
Difficulty of **validating** results (potentially) (F4)
Data protection; **stability** over time; **social acceptance** (the more transparency will improve on processes and intentions – i.e. **GDPR** – the more likely it will be accepted socially); **legal/regulatory** aspects (F6)
Legal access to the data for the defined purpose of use; managing the **consent**/notification channel; data **inaccuracies**; **potential for discrimination**; data **security**; availability (F7).
There is a risk that **GAFA** gets all the data and that little supervision will be possible. In addition, there is a problem of European sovereignty regarding this data as GAFAs are all from the US (C3)

When asked about the required improvements in the regulatory environment (Table 15), a very positive support was expressed for the GDPR. Nevertheless, calls for more discussion, clarification and consistency were made, especially for transnational consistency. Similar to previous questions, the views ranged from very optimistic – 'no problems' – to very pessimistic – unstructured non-traditional data should not be used for creditworthiness assessment. It should be clarified that the negative view refers to unstructured data only; structured non-traditional data received the full support, in particular, open-banking transactions.

Table 15. Required improvements in regulation/infrastructure

Are there any gaps in regulations/infrastructure that should be addressed in order to overcome the above-mentioned challenges?
If one is careful enough, then there should not be any problems (F2)
At this stage, given the fact that Big Data Technologies are at their early stage and the new forthcoming regulation frameworks (such as the GDPR) are in their implementation phase, **we encourage a dialogue at EU level in order to unlock the full potential of the usage of non- traditional data**, while being compliant with the GDPR (F4)
Transnationally **consistent regulation** (F6)
GDPR has helped to improve **transparency** regarding data usage and consent/notification. It also has the concept of **fairness** in legal decisions such as the application for credit (F7)
Consistency between different regulatory frameworks relating to consumer protection, financial services (and their providers) and data protection is needed to provide a level playing field across the EU. A general principle of same services = same rules should apply (F7)
Unfair regulatory treatment (compared with Fintech) (F8)
The challenges should not be overcome. Non-traditional data **should not be used for creditworthiness** as it's not fit for purpose and will do more damage than good. It would create an impossibly complex prudential compliance since risk assessments will be carried out on a personal basis rather than on larger risk pools. The risk is to find ourselves in the same situation as the subprime loan crisis, where there is an illusion of risk hedging thanks to "sophisticated" new, innovative hedging and risk assessment practices, which end up blowing up in the face of banks (C5)

12. Conclusions

Using public reports and qualitative interviews, this chapter reviewed the potential for the use of non-traditional data in lending decisions, and discussed the benefits and challenges associated with it.

Structured non-traditional data (payments and open banking transactions) are perceived as the most valuable (especially for affordability assessment) and the most socially/ethically acceptable type of data. Unstructured non-traditional data attracts mixed opinions with some stakeholders seeing its potential value, whilst others are completely opposed to its use.

The main benefits are seen in increasing financial inclusion, increasing the predictive accuracy of credit risk models, and offering a holistic view of a customer. The use of non-traditional data should be complementary to traditional types. Customers who are expected to benefit most from the use of non-traditional data include those with 'thin files' (with no or little credit history). This is supported with the evidence from studies mainly conducted in developing countries and the US.

Legal compliance is seen by almost all respondents as the main challenge/risk associated with the use of non-traditional data in credit granting. Other challenges include data quality, IT/technological difficulties, and social acceptance/ethics.

Overall policy recommendations

The results presented in Chapters 1 and 2 allow for the formulation of specific policy recommendations.

- To boost financial inclusion, policy-makers should facilitate the availability of more comprehensive credit data, notably structured non-loan data.
- More attention should be paid to the extent of credit data sharing in the analysis of solvency and the availability of credit.
- Policy changes should be introduced to enable wider data sharing in credit markets, particularly those markets with a high degree of concentration and relatively low barriers to entry.
- Legislation should not promote information that is only "up to date" and "accurate", as mandated by data protection legislation, but also "comprehensive".
- If unstructured, non-traditional information, such as social media data, is to be used as an input to determine creditworthiness, it should be done within a clearer ethical framework and with a better understanding of societal preferences.
- Further research should investigate the impact of non-loan data for specific segments of the population.

Annex 1. Definitions of the variables used in Chapter 1 and related statistics

Table A 1. Variables used in the regressions

VARIABLES	MEASURE	DESCRIPTION	SOURCE
Access_Poorest	%	Denotes the percentage of respondents in the poorest 40% of households, ages 15+, who report borrowing any money from a bank or another type of financial institution in the past 12 months	World Bank – Financial inclusion indicators
Access_Richest	%	Denotes the percentage of respondents in the richest 60% of households, ages 15+, who report borrowing any money from a bank or another type of financial institution in the past 12 months	World Bank – Financial inclusion indicators
Account_Poorest	%	Denotes the percentage of respondents in the poorest 40% of households, ages 15+, who report having an account (by themselves or together with someone else) at a bank or another type of financial institution or personally using a mobile money service in the past 12 months	World Bank – Financial inclusion indicators
Account_Richest	%	Denotes the percentage of respondents in the richest 60% of households, ages 15+, who report having an account (by themselves or together with someone else) at a bank or another type of financial institution or personally using a mobile money service in the past 12 months	World Bank – Financial inclusion indicators
Saved_Poorest	%	Denotes the percentage of respondents in the poorest 40% of households, ages 15+, who report saving or setting aside any money by using an account at a bank or another type of financial institution in the past 12 months	World Bank – Financial inclusion indicators
Saved_Richest	%	Denotes the percentage of respondents in the richest 60% of households, ages 15+, who report saving or setting aside any money by using an account at a bank or another type of financial institution in the past 12 months.	World Bank – Financial inclusion indicators
CB rate	%	Central bank policy rate	Bank for International Settlements
Consumer credit to deposits	%	Outstanding consumer credit to household deposits (%)	ECRI Statistical Package
Consumer credit to GDP	%	Outstanding consumer credit to GDP (%)	ECRI Statistical Package
GDP/capita	Current $/1000	GDP per capita (Purchase Power Parities)	World Bank
Concentration	Index	Herfindahl index for total credit	European Central Bank

Legal origin	Dummy	Dummy that indicates the origin of each country's company law or commercial code, which may be of English (Leg_UK), French (Leg_FR), German (Leg_DE) or Scandinavian (Leg_Scan)	Djankov et al. (2007)
Non-performing loans to total gross loans	%	Non-performing loans to total gross loans (%)	IMF – Financial Soundness Indicators
Interest rate	%	Bank interest rates – loans to households for consumption & other purposes with an original maturity of over one & up to five years (outstanding amounts)	European Central Bank
Tier1 capital ratio	%	Tier 1 capital (according to Basel III) – Tier 1 capital ratio is the ratio of a bank's core equity capital to its total risk-weighted assets (RWA)	European Central Bank
Total credit growth	%	Total credit to household and non-financial corporation 10-year growth rate (%)	ECRI Statistical Package
Unemployment	%	Unemployment rate (%)	Eurostat

Table A 2. Summary statistics

Panel 1: Comprehensiveness and Regulation Indexes	Average	Standard deviation	Minimum	Maximum	Observations
Breadth_General	0,43	0,15	0,13	0,77	85
Breadth_Individual	0,42	0,14	0	0,71	85
Breadth_Providers	0,36	0,20	0,04	0,81	83
Breadth_Services	0,56	0,28	0,16	1	83
Breadth_Loan services	0,79	0,19	0,29	1	83
Breadth_Non-loan services	0,43	0,39	0	1	83
Depth_General	0,27	0,15	0,07	0,75	50
Depth_Positive	0,24	0,17	0	0,82	50
Depth_Negative	0,28	0,15	0,06	0,76	50
Depth_Loan services	0,48	0,24	0,05	0,91	50
Depth_Non-loan services	0,14	0,16	0	0,73	50
Restriction_Index	0,59	0,24	0	1	55
Promotion_Index	0,41	0,24	0	1	55
Panel 2: Dependent variables	**Average**	**Standard deviation**	**Minimum**	**Maximum**	**Observations**
Access_poorest	32,2	12,4	5,7	58,5	40
Access_richest	43,2	15,5	14,8	68,9	40
Consumer credit to deposit	16,2	9,4	4,1	47,4	57
Consumer credit to GDP	7,8	3,8	1,1	17,7	59
Non-performing loans to total gross loans	8,2	10,3	0,6	47,7	61
Panle 3: control variables	**Average**	**Standard deviation**	**Minimum**	**Maximum**	**Observations**
CB rate	0,63	1,27	-0,75	5,75	62
GDP/capita	38,7	11,5	18,9	65,4	62
Interest rate	7,4	3,8	1,9	22,2	57
Concentration	0,11	0,06	0,01	0,27	53
Leg_UK	0,08	0,28	0	1	60
Leg_FR	0,33	0,48	0	1	60
Leg_Scan	0,15	0,36	0	1	60
Leg_DE	0,43	0,50	0	1	60
Tier 1 capital ratio	15,3	3,5	4,5	23,0	56
Total credit growth	57,2	90,6	-40,1	560,8	58
Unemployment	9,2	4,9	2,8	26,3	62
Account_poorest	87,5	15,0	37,8	100,0	41
Account_richest	92,9	8,1	71,0	100,0	41
Saved_poorest	35,4	17,9	5,0	71,3	41
Saved_richest	51,9	19,2	17,2	84,7	41

Source: ACCIS Membership Survey and authors' calculations.

Note:

- Standard deviation is the average of this variable over time.
- Contrary to the analyses in Section 2the Table A2 includes Iceland, Norway and Switzerland.

Annex 2. Econometric model: systematic testing and interpretation of the findings

The empirical part uses tools of panel econometrics. The objective is to assess the impact of different indexes of data comprehensiveness on a set of dependent variables which proxy specific mechanisms that can be of interest for the policy-making process: credit to GDP ratio for financial inclusion, credit to deposit ratio for an effective financial intermediation, the ratio of NPLs for financial risk and the Herfindahl-Hirschman index for the degree of competition. In addition, the Herfindahl-Hirschman index and the Tier1 ratio, which provides a proxy for the barrier to entry on credit markets, are regressed on different indicators of data comprehensiveness.

As regards the composition of the databases, the indexes of comprehensiveness are available for 33 CRAs which represent 23 countries. Both control variables and dependent variables provide data at country level. As several countries are represented by more than one CRA, there is a mismatch between variables at CRA level and variables at country level. Regressing variables at CRA level on variables at country level would entail some correlation between dependent variables, hereby distorting the model. Therefore, indexes of data comprehensiveness have been reshaped to provide data at country level by simply considering whether at least one CRA in a given country collects data on a specific item. If the answer is “yes”, then the value is “1”. Otherwise, the value is “0”. As no large gaps have been generally identified between the comprehensiveness indexes of CRAs within a given country, this restrictive approach could have a limited impact on the findings.

As regards the coverage of the respective CRAs, an interesting metric that is provided by the World Bank concerns the coverage of the population by credit bureau and credit registers in each country. However, this index does not provide any information on the share of that population which actually uses credit. Therefore, in line with most of the economic literature on credit data sharing, no metric has been included to appreciate the real coverage of CRAs within a given country. It is nonetheless assumed that these CRAs are related to a significant share of credit granted every year, especially within consumer credit markets.

The impact of the different indexes for comprehensiveness is systematically tested. This approach has different justifications. First, on the technical side, the testing of a general index of comprehensiveness and its different components can help elucidate the degree of robustness of findings and the extent to which they can matter.

Secondly, on the economic and financial side, these different tests can provide some insights on which specific type of data makes a difference. For instance, is positive data more powerful than negative data? Does the metric on the number of types of providers have a stronger impact than the number of types of services? Overall, the general indicators can give a first appreciation of specific synergies between different sets of data. Should, for instance, Breadth_General have a significant effect when all of its sub-indexes do not, then it could be assumed that the combination of different sub-indexes generates synergies and the ‘together approach’ has a greater impact than the ‘single approach’.

For example, a high number of data providers (Breadth_Providers) could have an impact only if this sub-index is combined with data on many services (Breadth _Services) and on income/assets of borrowers (Breadth_Individual). This would imply that Breadth_General would have a significant impact, whereas each sub-index taken separately would have only a non-significant impact.

Different tests are conducted for each regression to find the best tools to control for cross-country heterogeneity. The conduct of the technique of Pooled Ordinary Least Square (OLS) with panel data ignores individually specific effects. As such, specific assumptions such as the orthogonality of the error term can be violated.

One common possibility to address these issues is to use a model of random effects or fixed effects. The Breusch-Pagan Lagrangian multiplier test for random effects indicates that random effects are preferred over Pooled OLS for all the regressions conducted. The Hausman test indicates that random effects are preferred over fixed effects. Models integrating random effects are therefore systematically used in the regressions by implementing an individual specific intercept which is assumed to be random. This approach can significantly strengthen the robustness and consistency of the estimates. It should also help reinforce consistency over time and limit heterogeneity in the dynamics observed across the different years. This approach should therefore help limit specific effects exacerbated in certain years.

Several additional tests have been systematically conducted to assess potential problems of collinearity and autocorrelation. For this purpose, the Variance inflation factor (VIF) analysis and the Breusch-Godfrey test have been run and have not revealed significant issues in the regressions published in Annex 1.

Despite the use of all of these parameters to control for different possible distortions, the limited number of observations (between 34 and 60, depending on the regressions) entails that findings should be interpreted with caution. The use of further data, for example during a next wave of the survey, could help reinforce the robustness of the findings. Nevertheless, as analysed in Section 2.2, several relevant papers in the economic literature on credit data sharing conducted their regressions with a broadly similar number of observations and often with more heterogeneous countries within the same sample.

More specifically on the use of control variables, the risk of multi-collinearity between the four dummies for legal origin is controlled by using the dummy for German origin as a benchmark. This allows avoiding the risk of the dummy variable trap. As a consequence, the coefficients of the other dummies for legal origins need to be interpreted in comparison with that for German origin (in relative terms).

Last but not least, coefficients deemed to be considered significant are those whose P-value is below 10%. Four possibilities are shown in the regressions for each coefficient: no significance at all (no asterisk), moderate significance (one asterisk: the P-value is between 5% and 10%), high significance (two asterisks: the P-value is between 1% and 5%) and very high significance (three asterisks: the P-value is below 1%).

For most regressions, there is a mismatch between the measures of certain dependent variables and the indexes of comprehensiveness. Most of the dependent variables provide values in per cent whereas the indexes of comprehensiveness range from zero to one. Therefore, coefficients can be interpreted as follows: when the dependent variable A is in per cent, then a significant coefficient of five would imply that if the index of comprehensiveness increases by 0.1 (for example from 0.5 to 0.6), then A will increase by 0.1X5=0.5 percentage point (for example, from 10% to 10.5%). For these specific figures, this would mean that an increase of 20% in the index of comprehensiveness would result in an increase of 5% in the dependent variable A. Should the index of comprehensiveness be Breadth_General, this would entail that data should be collected for an additional number of broadly seven new items (for instance one item among consumer/borrower identity data, two new types of providers of data and four new types of services) in order to have an increase from 10% to 10.5% of the dependent variable.

Annex 3. Econometric findings on financial inclusion

Table A 3. The impact on consumer credit to GDP

	Consumer credit to GDP						
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Breadth_General	2.599**	2.502*					
	(1.124)	(1.339)					
Breadth_Non-loan services			1.132**	1.106*			
			(0.564)	(0.637)			
Depth_Positive					5.592***	4.539***	
					(1.710)	(1.364)	
Depth_Non-loan services							4.098*
							(2.517)
GDP/capita	-0.234***	-0.186*	-0.220***	-0.167*	-0.098***	-0.154**	-0.124***
	(0.041)	(0.096)	(0.041)	(0.087)	(0.029)	(0.063)	(0.038)
Unemployment	0.003	-0.071	0.024	-0.051	0.261**	0.317***	0.222*
	(0.123)	(0.084)	(0.119)	(0.084)	(0.106)	(0.080)	(0.131)
Leg_UK	3.131***	2.605***	3.340***	2.715***	1.156	2.182**	1.788
	(0.856)	(1.007)	(0.816)	(1.023)	(1.168)	(0.952)	(1.666)
Leg_FR	1.078	1.467	1.114	1.455	0.089	-0.594	0.064
	(1.647)	(1.889)	(1.652)	(1.919)	(1.307)	(1.256)	(1.389)
Leg_Scan	0.963	1.021	0.711	0.681	1.387	0.496	0.789
	(0.914)	(1.263)	(0.870)	(1.207)	(1.045)	(1.371)	(0.985)
Interest rate		0.173		0.177		-0.328	
		(0.218)		(0.212)		(0.233)	
Constant	14.254***	11.840**	14.222***	11.627**	6.819***	11.271**	8.941***
	(2.048)	(5.211)	(1.995)	(4.953)	(2.112)	(4.415)	(2.415)
Observations	56	54	56	54	37	36	37
Number of countries	20	19	20	19	20	19	20

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table A 4. Breadth and depth for the poorest and richest consumers

	Access_poorest				Access_richest			
VARIABLES	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Breadth_General	15.329*	15.252*			12.143*	13.057*		
	(8.547)	(8.759)			(7.968)	(7.891)		
Depth_General			21.053***	21.068***			23.978**	25.239***
			(6.406)	(6.486)			(9.751)	(9.689)
GDP/capita	0.088	0.086	0.338*	0.330**	0.161	0.255	0.441*	0.570**
	(0.188)	(0.185)	(0.177)	(0.160)	(0.282)	(0.293)	(0.254)	(0.231)
Account_poorest	0.225	0.224	0.274**	0.269*				
	(0.145)	(0.162)	(0.134)	(0.149)				
Saved_poorest	0.271**	0.269**	0.153	0.153				
	(0.128)	(0.126)	(0.121)	(0.118)				
Account_richest					0.751***	0.885***	0.901***	1.025***
					(0.278)	(0.291)	(0.347)	(0.372)
Saved_richest					0.155	0.138	-0.062	-0.093
					(0.164)	(0.145)	(0.197)	(0.183)
Leg_UK	5.716	5.710	1.531	1.529	5.833	5.580	1.091	0.684
	(8.092)	(8.278)	(6.495)	(6.678)	(8.419)	(7.557)	(8.325)	(7.580)
Leg_FR	-1.336	-1.349	-0.864	-0.863	1.454	1.472	2.622	2.685
	(3.602)	(3.658)	(4.210)	(4.263)	(4.288)	(4.250)	(5.022)	(5.076)
Leg_Scan	-1.150	-1.145	3.278	3.278	2.606	2.553	9.698	10.042
	(4.521)	(4.565)	(4.568)	(4.640)	(6.987)	(6.682)	(7.014)	(6.876)
CB rate		-0.112		-0.247		2.862*		2.865*
		(1.979)		(1.952)		(1.485)		(1.574)
Constant	-7.764	-7.460	-15.709*	-14.963	-47.893***	-64.349***	-62.594**	-78.683***
	(8.797)	(11.891)	(9.128)	(12.071)	(17.389)	(18.958)	(24.319)	(27.169)
Observations	40	40	40	40	40	40	40	40
Number of countries	21	21	21	21	21	21	21	21

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1 (the coefficient of 12.143 for the Breadth_General in regressions on Access_richest has a significance of 12%)

Table A 5. The depth of positive and negative data for the poorest and richest consumers

	Access_poorest				Access_richest			
VARIABLES	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Depth_Negative	14.507***	14.500***			19.925***	20.956***		
	(5.441)	(5.529)			(5.930)	(5.514)		
Depth_Positive			18.872***	18.921***			10.868	11.130
			(6.235)	(6.241)			(10.089)	(10.412)
GDP/capita	0.297*	0.293*	0.370*	0.359**	0.433*	0.565**	0.378	0.479*
	(0.168)	(0.153)	(0.189)	(0.174)	(0.242)	(0.222)	(0.284)	(0.282)
Account_poorest	0.235*	0.233*	0.299**	0.294*				
	(0.126)	(0.140)	(0.145)	(0.160)				
Saved_poorest	0.189	0.189	0.126	0.126				
	(0.126)	(0.123)	(0.123)	(0.122)				
Account_richest					0.867***	1.001***	0.757**	0.866***
					(0.299)	(0.299)	(0.314)	(0.328)
Saved_richest					-0.032	-0.062	0.070	0.057
					(0.184)	(0.162)	(0.190)	(0.183)
Leg_UK	4.350	4.352	0.570	0.555	3.549	3.249	3.645	3.437
	(7.598)	(7.756)	(7.314)	(7.534)	(7.932)	(7.025)	(9.749)	(9.088)
Leg_FR	-1.407	-1.408	-1.178	-1.173	2.191	2.227	1.525	1.503
	(4.084)	(4.142)	(4.336)	(4.386)	(4.688)	(4.725)	(4.646)	(4.661)
Leg_Scan	1.870	1.866	2.657	2.666	8.315	8.564	5.146	5.012
	(4.228)	(4.285)	(4.461)	(4.539)	(5.915)	(5.736)	(6.855)	(6.766)
CB rate		-0.118		-0.314		3.077*		2.515
		(1.977)		(1.938)		(1.593)		(1.541)
Constant	-10.587	-10.226	-17.024	-16.087	-60.055***	-77.304***	-49.255**	-63.294***
	(7.993)	(11.273)	(10.474)	(13.351)	(19.521)	(19.394)	(22.074)	(24.115)
Observations	40	40	40	40	40	40	40	40
Number of countries	21	21	21	21	21	21	21	21

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table A 6. The depth of loan and non-loan data for the poorest and richest consumers

	Access_poorest				Access_richest			
VARIABLES	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Depth_Loan services	9.921**	9.993**			11.966**	13.396**		
	(4.521)	(4.809)			(5.915)	(5.417)		
Depth_Non-loan services			25.771***	26.513***			28.644**	27.997**
			(6.064)	(6.441)			(11.219)	(11.442)
GDP/capita	0.343*	0.352**	0.311*	0.279*	0.398	0.550**	0.429*	0.511**
	(0.179)	(0.160)	(0.175)	(0.161)	(0.272)	(0.240)	(0.241)	(0.238)
Account_poorest	0.278**	0.283*	0.247*	0.224				
	(0.133)	(0.150)	(0.132)	(0.145)				
Saved_poorest	0.205	0.204	0.088	0.090				
	(0.131)	(0.127)	(0.130)	(0.130)				
Account_richest					0.784***	0.919***	0.922**	1.008***
					(0.282)	(0.313)	(0.370)	(0.365)
Saved_richest					0.055	0.028	-0.110	-0.120
					(0.160)	(0.145)	(0.232)	(0.218)
Leg_UK	4.240	4.220	0.150	0.017	4.284	3.860	-0.335	-0.280
	(9.508)	(9.560)	(4.385)	(4.644)	(10.500)	(9.522)	(6.282)	(6.107)
Leg_FR	-1.667	-1.660	-0.823	-0.756	1.820	1.920	2.720	2.634
	(4.261)	(4.351)	(4.093)	(4.105)	(4.940)	(5.001)	(4.850)	(4.897)
Leg_Scan	1.235	1.265	3.595	3.703	7.640	8.055	8.761	8.588
	(4.963)	(5.029)	(3.948)	(4.037)	(7.300)	(7.206)	(6.892)	(6.784)
CB rate		0.247		-1.058		3.164*		2.182
		(2.108)		(1.936)		(1.701)		(1.550)
Constant	-17.012*	-17.876	-8.222	-4.854	-55.148***	-73.804***	-59.145**	-70.349***
	(9.622)	(13.649)	(8.719)	(11.763)	(20.050)	(23.746)	(24.027)	(23.777)
Observations	40	40	40	40	40	40	40	40
Number of countries	21	21	21	21	21	21	21	21

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Annex 4. Econometric findings on financial intermediation

Table A 7. Impact on the consumer credit to deposits ratio

	Consumer credit to deposits							
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Breadth_General	9.566***	9.109***						
	(3.149)	(3.375)						
Breadth_Individual			5.886**					
			(2.796)					
Breadth_Providers				6.124**	4.519**			
				(2.661)	(2.147)			
Breadth_Services						3.303*		
						(1.840)		
Breadth_Non-loan services							3.077**	3.087*
							(1.516)	(1.620)
GDP/capita	-0.890***	-0.612***	-0.811***	-0.815***	-0.515***	-0.791***	-0.797***	-0.510***
	(0.253)	(0.194)	(0.231)	(0.238)	(0.176)	(0.245)	(0.250)	(0.188)
Unemployment	-0.167	-0.311	-0.199	-0.184	-0.313	-0.121	-0.105	-0.257
	(0.324)	(0.329)	(0.347)	(0.344)	(0.359)	(0.327)	(0.327)	(0.336)
Leg_UK	-1.175	-1.745	-1.022	-1.845	-2.395	-1.533	-1.510	-2.119
	(2.675)	(2.431)	(3.286)	(2.689)	(2.348)	(3.337)	(3.211)	(2.563)
Leg_FR	1.064	2.076	0.776	0.267	1.175	0.643	0.734	1.808
	(4.342)	(4.174)	(4.286)	(4.359)	(4.245)	(4.480)	(4.532)	(4.393)
Leg_Scan	7.024**	6.790*	6.743*	6.210*	5.623	5.842*	5.617*	5.338
	(3.581)	(3.829)	(3.496)	(3.548)	(3.740)	(3.392)	(3.316)	(3.596)
Interest rate		0.792**			0.772*			0.817**
		(0.402)			(0.439)			(0.388)
Constant	44.978***	30.220***	44.151***	44.785***	29.636***	43.638***	44.276***	28.835***
	(9.525)	(7.285)	(9.460)	(9.896)	(7.727)	(9.660)	(9.936)	(7.349)
Observations	55	55	55	55	55	55	55	55
Number of countries	20	20	20	20	20	20	20	20

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Annex 5. Econometric findings on the risk of future missed repayments

Table A 8. Impact on the non-performing loan ratio

	Non-performing loans to total gross loans				
VARIABLES	(1)	(2)	(3)	(4)	(5)
L.Breadth_General	-11.163***	-4.273**			
	(3.306)	(1.867)			
L.Breadth_Individual			-6.716**		
			(2.626)		
L.Breadth_Providers				-9.228***	
				(3.387)	
L.Breadth_Services					-3.295*
					(1.886)
CB rate	3.171***	2.737***	3.451***	3.037***	3.041***
	(0.895)	(0.849)	(1.293)	(1.027)	(0.991)
Unemployment		0.972***			
		(0.292)			
Total credit growth		0.001			
		(0.020)			
Constant	8.957***	-1.406	7.142***	7.690***	6.040***
	(1.928)	(1.877)	(1.704)	(1.850)	(1.504)
Observations	34	34	34	34	34
Number of countries	18	18	18	18	18

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Annex 6. Drivers behind the degree of comprehensiveness

Table A 9. Impact of the degree of competition and concentration

	Breadth_Individual			Breadth_Loan services		
VARIABLES	(1)	(2)	(3)	(1)	(2)	(3)
Tier1 capital ratio	0.020***		0.018***	0.023***		0.021***
	(0.004)		(0.005)	(0.006)		(0.007)
Concentration		-0.871***	-1.134***		0.792	0.479
		(0.255)	(0.328)		(0.526)	(0.506)
Leg_UK	0.038	-0.250***	-0.051	0.136***	-0.132***	0.094
	(0.073)	(0.029)	(0.064)	(0.042)	(0.038)	(0.089)
Leg_FR	-0.050	-0.018	-0.006	-0.065	-0.101*	-0.085
	(0.055)	(0.038)	(0.049)	(0.056)	(0.055)	(0.059)
Leg_Scan	-0.155	-0.020	-0.083	-0.226**	-0.177*	-0.240**
	(0.098)	(0.070)	(0.073)	(0.104)	(0.091)	(0.096)
Constant	0.156***	0.526***	0.275***	0.469***	0.756***	0.474***
	(0.056)	(0.037)	(0.070)	(0.097)	(0.060)	(0.118)
Observations	54	51	51	54	51	51
Number of countries	20	19	19	20	19	19

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A 10. Impact of the laws promoting data sharing

	Breadth_General	Breadth_Individual	Breadth_Providers	Breadth_Services	Breadth_Non-loan services
VARIABLES	(1)	(2)	(3)	(4)	(5)
Promotion_Iindex	0.284***	0.106**	0.171**	0.483**	0.622***
	(0.051)	(0.052)	(0.070)	(0.220)	(0.233)
Leg_UK	0.150*	0.085**	0.230	0.134	0.182
	(0.089)	(0.041)	(0.141)	(0.114)	(0.161)
Leg_FR	-0.086	-0.079	-0.009	-0.245**	-0.346*
	(0.092)	(0.060)	(0.108)	(0.124)	(0.186)
Leg_Scan	-0.036	-0.120	-0.036	-0.084	-0.011
	(0.067)	(0.090)	(0.154)	(0.229)	(0.270)
Constant	0.401***	0.478***	0.329***	0.509***	0.339**
	(0.061)	(0.031)	(0.082)	(0.124)	(0.160)
Observations	37	37	37	37	37
Number of countries	20	20	20	20	20

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

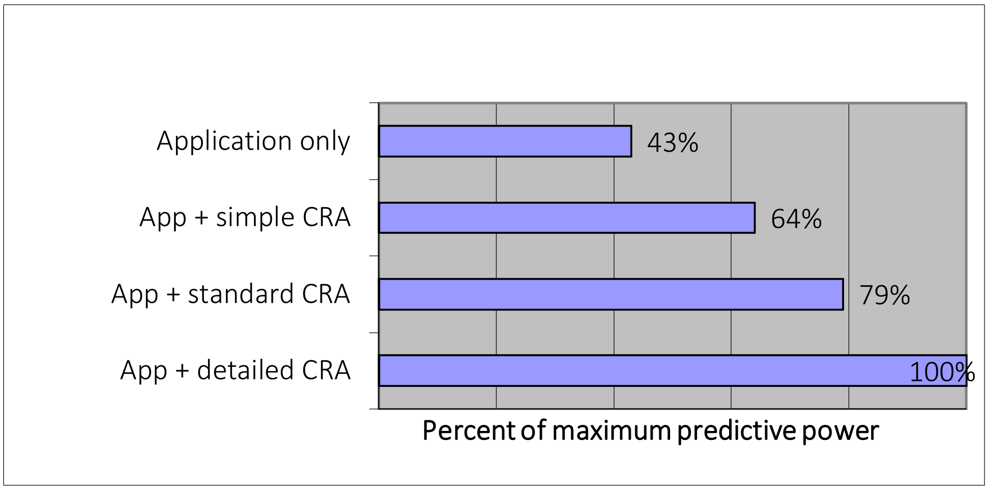
Appendix 1. The role of traditional data in improving predictive accuracy of credit scoring models and widening access to credit: the account-level analysis

Improvement in predictive accuracy (US evidence)

There is little empirical evidence at the individual account level to support the importance of CRAs in lending in terms of the quality of prediction given by credit scoring models. Chandler & Johnson (1992) tested the contribution of data contained in US credit reports to the predictive ability of scoring models. The relationship between the level of predictive power and the level of CRA data was studied under four scenarios with different levels of detail of CRA information. The predictive power of each model was measured by the Kolmogorov-Smirnov statistic that measures the ability of the scoring model to separate 'good' and 'bad' customers. Figure A1 presents the predictive power for different levels of CRA data that are compared on a percentage basis, with the highest Kolmogorov-Smirnov score scaled to 100%.

Their results suggested that the more detailed the CRA data, the better is prediction and the more likely it is that good customers will receive credit and that bad customers will be denied the use of credit. They concluded that placing limits on data retained and communicated by CRAs is 'anti-consumer'.

Figure A 1. Predictive power of credit scoring models depending on the level of CRA report detail



Source: Chandler & Johnson (1992).

Another study by Barron & Staten (2003) addressed the value of the positive CRA data ('Full Model', i.e. using data on transactions, usage of different credit products) in comparison to negative CRA data only ('Negative-only Model', i.e. using data on delinquencies/defaults only). In addition, the authors investigated the impact of using the information about all possible credit products from multiple lenders ('Full Model') as compared to the information from only one product of the incumbent lender ('Reduced CRA Model'). The impact was considered in two dimensions: (1) the difference in default rates (i.e. the level of risk) for the same acceptance rate; (2) the difference in the percentage of customers accepted for credit (credit availability) for the same default rate. The authors observed superior results for the Full Model in both scenarios and both impact measures. Table A11 reports results on credit availability for the Negative-only scenario, and Table A12 for the reduced CRA model. Whilst the Full Model outperforms the other models in both cases, the most striking difference is observed when only negative CRA data is used: up to 47.5% more customers would receive credit if positive CRA were included in the model.

Table A 11. Effects on credit availability of adopting a negative-only credit scoring model for various default rates

Target default rate	*Percent of consumers who obtain a loan Training sample*			*Percent of consumers who obtain a loan Test sample*		
	Full Model	*Negative-only Model*	*Percent decrease in consumers who obtain a loan with Negative-only Model*	*Full Model*	*Negative-only Model*	*Percent decrease in consumers who obtain a loan with Negative-only Model*
3%	74.8%	39.8%	46.8%	74.3%	39.0%	47.5%
4%	83.2%	73.7%	11.4%	82.9%	73.7%	11.1%
5%	88.9%	84.6%	4.8%	88.9%	84.2%	5.3%
6%	93.1%	90.8%	2.5%	92.8%	90.6%	2.4%
7%	95.5%	95.0%	0.5%	95.6%	94.6%	1.0%

Table A 12. Effects on credit availability of a reduced CRA credit scoring model for various bankcard loan default rates

Target default rate	*Percent of consumers who obtain a loan Training sample*			*Percent of consumers who obtain a loan Test sample*		
	Full Model	*Reduced CRA Model*	*Percent decrease in consumers who obtain a loan with Reduced CRA Model*	*Full Model*	*Reduced CRA Model*	*Percent decrease in consumers who obtain a loan with Reduced CRA Model*
2%	79.6%	75.6%	5.0%	77.8%	74.6%	4.1%
3%	89.3%	85.6%	4.1%	88.3%	85.3%	3.4%
4%	95.8%	93.4%	2.5%	95.0%	93.4%	1.7%

<u>Improvement in predictive accuracy (European evidence) and implications for customers from the application of different models</u>

Since the level of the detail and the amount of data held by CRAs varies from country to country, the significance of CRA data in credit scoring models will also be different. In countries with less detailed CRA information, the contribution of application information to predictive power would be expected to be greater than in countries with more detailed CRA information. Such asymmetries create additional difficulties in credit risk assessment at the European level.

In the European context Andreeva (2004), Andreeva et al. (2008) compared credit scoring models for the same credit product (store card) in Belgium, Germany and the Netherlands. The objective was to build a generic model and measure how its predictive accuracy differed across the countries

The generic model did not include credit bureau characteristics, since they were too different across the three countries. Only 16 characteristics were common across three countries and were used in 'generic' and 'national' models.For each country, 'full-information' national models were built on the full range of variables available for that country. They resulted in more statistically significant variables: Belgium had eight additional statistically significant variables (no CRAs); the Netherlands, 13 (six CRAs); Germany, 23 (eight CRAs).

Table A 13. Datasets used in a comparison of three EU countries

	Belgium	The Netherlands	Germany
No. of applications	108,517	566,960	894,251
No. of months	26 from 01/10/1998 to 30/11/2000	35 from 01/01/1998 to 30/11/2000	68 from 01/04/1995 to 30/11/2000
Characteristics	38 (incl. 4 CRA)	60 (incl. 13 CRA)	75 (incl. 17 CRA)

The research demonstrated that additional information substantially increases the predictive ability of models, whilst the difference between generic and national models developed on the same set of common characteristics is marginal.

Table A 14. Area under the ROC curve (AUC), the higher the better

	Belgium	The Netherlands	Germany
Generic	0.701	0.777	0.731
National	0.707	0.780	0.739
Full-info	0.722	0.800	0.789

Table A 15. % of misclassified observations, the lower the better

	Belgium	The Netherlands	Germany
Generic	17.00%	16.58%	16.04%
National	16.96%	16.48%	15.74%
Full-info	16.72%	15.94%	14.68%

Apart from predictive accuracy, there are differences in the types of customers accepted/rejected by each model. For customers, it matters which model is applied.

For each country hold-out sample, the proportions of applicants accepted by the generic model but rejected by the national model and vice versa were calculated. For Belgium, these proportions constituted 3.83% each from the total hold-out sample, for the Netherlands 2.75%, and for Germany 2.82%. In other words, the decision whether to grant credit or not would be identical (irrespective of whether the generic or national decision rule is used) for 92.34% of Belgian applicants, 94.50% of the Dutch, and 94.36% of German applicants.

However, it would be of interest to investigate what characteristics distinguish the two groups:

1. the applicants predicted 'good' (and therefore accepted) by the generic model and predicted 'bad' (and rejected) by the country model;
2. the applicants predicted 'good' by the country model and 'bad' by the generic model.

So, for each country the two groups were cross-checked by looking at frequency distributions for categorical characteristics; an example for Belgium is shown in Table A16 for characteristics that demonstrated the most striking differences. For categorical characteristics only some attributes are selected for the purpose of illustration, that is why the percentages within one characteristic do not always sum up to 100%.

Table A 16. Differences between the applicants accepted by one model but rejected by another one. Belgium: categorical characteristics

Characteristic	Attribute	Total	Accepted by generic, rejected by country (3.83%)	Accepted by country, rejected by generic (3.83%)
		% cases		
Marital status	Single	27.09%	45.18%	64.12%
	Divorced	12.54%	16.61%	6.98%
Occupation	Part-time	5.10%	8.31%	2.66%
	Self-employed	3.96%	6.64%	11.96%
Residential status	Rented flat	18.70%	28.90%	16.61%
	Living with parents	12.94%	26.25%	35.55%

References

ACCIS, 2012 Survey of Members. Analysis of credit bureaus in Europe. June 2013.

ACCIS, 2015 Survey of Members. Analysis of credit reporting in Europe. November 2015.

ACCIS, 2017 Survey of Members. Analysis of credit reporting in Europe. March 2018.

Andreeva G. (2004) 'Credit Risk in the Context of European Integration: Assessing the Possibility of Pan-European Scoring'. *PhD Thesis,* University of Edinburgh.

Andreeva G., Ansell J., Crook J.N. (2008) 'Credit scoring in the context of European integration: is there a future for generic models? *Journal of Financial Transformation,* 23, 129-134.

Andreeva G., Matuszyk A. (2019) 'The Law of Equal Opportunities or Unintended Consequences: the impact of unisex risk assessment in consumer credit', *Journal of Royal Statistical Society, Series A.* https://rss.onlinelibrary.wiley.com/doi/full/10.1111/rssa.12494

Barron, J.M., & M. Staten, M. (2003). The value of comprehensive credit reports: Lessons from the US experience, in Credit Reporting Systems and International Economy, ed. by M. Miller.

Beck, T., & Demirguc-Kunt, A. (2009). Financial institutions and markets across countries and over time-data and analysis. The World Bank.

Berg T., Burg V., Gombović A., Puri M. (2018): 'On the rise of fintechs – credit scoring using digital footprints', NBER Working Papers, no 24551, April.

Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models. Australian Economic Papers, 17(31), 334-355.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica: Journal of the Econometric Society, 1287-1294.

Brown, M., Jappelli, T., & Pagano, M. (2006). Information sharing and credit market performance: firm-level evidence from transition countries. CEPR Working Paper.

Bruhn, M., Farazi, S., & Kanz, M. (2013). Bank competition, concentration, and credit reporting. The World Bank.

Carroll P., Rehmani S. (2017) 'Alternative data and the underbanked'. *Point of view*, Oliver Wyman (www.oliverwyman.com/content/dam/oliver-wyman/v2/publications/2017/may/Alternative_Data_And_The_%20Unbanked.pdf).

Chandler, G.G., Johnson R.W. (1992) 'The benefit to consumers from generic scoring models based on credit reports', *IMA Journal of Mathematics Applied in Business & Industry*, 4, 61-72.

Djankov, S., McLiesh, C., & Shleifer, A. (2007). Private credit in 129 countries. Journal of financial Economics, 84(2), 299-329.

Doblas-Madrid, A., & Minetti, R. (2013). Sharing information in the credit market: Contract-level evidence from US firms. Journal of financial Economics, 109(1), 198-223.

Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., & J. Kammler, J. (2016). Description- text related soft information in peer-to-peer lending – Evidence from two leading European platforms. *Journal of Banking & Finance* 64:169-187.

Duarte, J., Siegel, S. & Young, L. (2012). Trust and credit: the role of appearance in peer-to-peer lending. *The Review of Financial Studies* 25(8): 2455–2484.

Experian Information Solutions (2018)' The state of alternative credit data' (www.experian.com/consumer-information/alternative-data.html).

FICO (2018) 'Using alternative data in credit risk modelling' (www.fico.com/blogs/analytics-optimization/using-alternative-data-in-credit-risk-modeling/).

Finextra (2017) 'CFPB investigates alternative data for credit scores' (www.finextra.com/pressarticle/68074/cfpb-investigates-alternative-data-for-credit-scores).

Freedman, S. & Jin, G. (2014). The Information Value of online social networks: lessons from peer-to-peer lending. *NBER Working Paper 19820* (www.nber.org/papers/w19820).

Frost J., Gambacorta L., Huang Y., Shin H.S., Zbinden P. (2019) 'BigTech and the changing structure of financial intermediation'. Bank for International Settlements, Working Paper 779.

Gao, Q., M. Lin, & R. Sias (2017). Word Matters: The Role of Texts in Online Credit Markets, Working Paper.

Giannetti, C., & Jentzsch, N. (2011). Credit Reporting, Access to Finance and Identification Systems: International Evidence. Jena economic research papers, No. 2011,031Global Partnership for Financial Inclusion, GPFI (2018). 'Argentina's G20 Presidency. 2018 Priorities Paper' (www.gpfi.org/sites/gpfi/files/documents/GPFI%20Argentina%20Priorities%20Paper%202018.pdf).

Hausman, J. A. (1978). Specification tests in econometrics. Econometrica: Journal of the econometric society, 1251-1271.

International committee on credit reporting ICCR (2018). 'Use of Alternative Data to Enhance Credit Reporting to Enable Access to Digital Financial Services by Individuals and SMEs operating in the Informal Economy, June 2018.

Iyer, R., A. Khwaja, E. Luttmer, & K. Shue (2016). Screening Peers Softly: Inferring the Quality of Small Borrowers, *Management Science* 62(6), 1554-1577.

Jappelli, T., & Pagano, M. (2002). Information sharing, lending and defaults: Cross-country evidence. Journal of Banking & Finance, 26(10), 2017-2045.

Larrimore, L., Jiang, L., & Larrimore, J. (2011). Peer to peer lending: the relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research* 39, pp.19–37.

Liberati, C. & Camillo, F. (2018). Personal values and credit scoring: new insights in the financial prediction, Journal of the Operational Research Society, 69:12, 1994-2005, DOI: 10.1080/01605682.2017.1417684

Lin, M., Prabhala, N., & Viswanathan, S. (2013). Judging borrowers by the company they keep: friendship networks and information asymmetry in online peer-to peer lending. *Management Science* 59, pp.17–35.

Karapetyan, A., & Stacescu, B. (2013). Information sharing and information acquisition in credit markets. Review of Finance, 18(4), 1583-1615.

Kusi, B. A., & Ansah-Adu, K. (2015). Credit information sharing and its impact on access to bank credit across income bracket groupings. Int. J, 4(4).

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (1997). Legal determinants of external finance. The journal of finance, 52(3), 1131-1150.

Love, I., & Mylenko, N. (2003). Credit reporting and financing constraints. The World Bank.

Liu, D., Brass, D., Lu, Y., & Chen, D. (2015). Friendships in online peer-to-peer lending-pipes, prisms, and relational herding. *MIS Quarterly: Management Information Systems* 39(3), pp.729-742.

Mester, L., Nakamura, L. & M. Renault, M. (2004). Transaction accounts and loan monitoring. *Review of Financial Studies* 20, 529-556. FRB of Philadelphia Working Paper No. 04-20 (https://ssrn.com/abstract=606681 or http://dx.doi.org/10.2139/ssrn.606681).

Miller, M. (2000, June). Credit reporting systems around the globe: the state of the art in public and private credit registries. In World Bank. Presented at the Second Consumer Credit Reporting World Conference, held in San Francisco, California, October.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied linear statistical models (Vol. 4, p. 318). Chicago: Irwin.

Norden, L., & Weber, M. (2010). Credit Line Usage, Checking Account Activity, and Default Risk of Bank Borrowers. *Review of Financial Studies* 23, 3665-3699.

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, 74, 26-39.

Pagano, M., & Jappelli, T. (1993). Information sharing in credit markets. *The Journal of Finance*, *48*(5), 1693-1718.

Porta, R. L., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (1998). Law and finance. Journal of political economy, 106(6), 1113-1155.

Puri, M., Rocholl, J., & Steffen, S. (2017). What do a million observations have to say about loan defaults? Opening the black box of relationships, *Journal of Financial Intermediation* 31, 1-15.

Schonfeld E. (2011). Eric Schmidt's Gang Of Four: Google, Apple, Amazon, And Facebook – TechCrunch (techcrunch.com).

Thomas L.C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16,(2), 149-172.

Thomas L.C., Crook J.N., Edelman D. (2017). 'Credit scoring and its applications'. CIAM, 2nd ed.

About ECRI & CEPS

Founded in 1999 by a consortium of European banking and financial institutions, the European Credit Research Institute is an independent, non-profit research institute based in Brussels. ECRI provides in-depth analysis and insight into the structure, evolution and regulation of retail financial services markets in Europe. It derives its expertise from an interdisciplinary team of in-house researchers and a network of academic partners based throughout Europe. ECRI keeps its members and the wider public up-to-date on a wide range of topics related to retail financial services, credit reporting and consumer protection at the European level. Its operations and staff are managed by the Centre for European Policy Studies (CEPS).

The Centre for European Policy Studies (CEPS), founded in 1983, is among the most experienced and authoritative think tanks operating in the European Union today. CEPS serves as a leading forum for debate on EU affairs, drawing from its strong in-house research capacity, complemented by an extensive network of partner institutes throughout the world.

ISBN 978-94-6138-745-5

EUROPEAN CREDIT RESEARCH INSTITUTE (ECRI)
at the CENTRE FOR EUROPEAN POLICY STUDIES (CEPS)
1 Place du Congrès
1000 Brussels, Belgium
Tel: +32 2 229 39 11 • Fax: +32 2 219 41 51
info@ecri.eu • www.ecri.eu • www.ceps.eu