Credit Information Market Study Interim Report

Annex 1: Data quality

November 2022

Contents

1 Introduction and our approach 2

2 Population coverage and information depth 4

3 Comparison of scores 11

4 Comparison of information on individuals 22

5 Impact of data differences on lending decisions 36

Technical Appendix 1: Sampling and matching methodology 49

1 Introduction and our approach

1. This annex describes our methodology and findings on the quality of credit information and its potential impact on lender decisioning processes.

Approach

2. To assess the quality of information held by the 3 large Credit Reference Agencies (CRAs), we requested the credit histories of a representative sample of UK individuals from each of Equifax, Experian and TransUnion. Individuals in the sample were chosen based on a particular date of birth. The information requested included personal identifiers of the individuals in the sample (eg names, postcodes and dates of birth), their credit history in the 6 years up to 1 August 2019, and their credit scores over the same period[1]. The Technical Appendix explains our sampling approach in more detail.

3. By data quality we mean population coverage and other aspects of the quality of credit information, such as accuracy. We use consistency of information between the CRAs as a proxy for quality.

4. Our methodology enabled us to receive a sample of individuals that could be matched across CRAs, allowing us to compare the data held by each CRA on the same individual. The matched sample therefore allowed us to identify differences in CRAs assessment of individual credit risk and the underlying data held by the CRAs.[2] We discuss alternative approaches and why we chose this approach over them in the Technical Appendix.

5. As we are unable to observe the 'perfect' credit file for each individual we have inferred the quality of data held by the 3 large CRAs by comparing the differences between them. If significant differences are found it can be inferred that at least 1 CRA either holds inaccurate information or is missing information. This may be due to a CRA not matching the data to the individual, wrongly matching data to an individual, or the CRA not receiving the data. If it is found that the data held by the CRAs is the same, it is possible that all 3 CRAs hold credit files that are missing information. However, this is quite unlikely, so where information matches across all 3 CRAs, this indicates that a credit file is more likely to be accurate and not be missing credit information. It is also likely that our methodology, which focuses mainly on those individuals known to all 3 large CRAs, means that our estimates provide a lower bound. Where we compare individual data between CRAs we limit our analysis only to where we can uniquely match IDs between CRAs. This means by construction we exclude cases where a CRA has the same details for multiple IDs, and where the same details map to multiple IDs at another CRA. This reduces the risk that the individuals we are comparing are not actually the same individuals.

6. In this annex we refer to the CRAs as CRA A, B and C. Note that these labels are not consistently used to refer to the same CRA throughout the annex to ensure anonymity is preserved.

[1] Much of the above analysis was conducted in 2019 and 2020, prior to the pandemic. Nonetheless, we have considered more recent sources (eg market publications, independent consumer research) and engaged with industry stakeholders (CISPs, consumer bodies, CIUs) throughout 2022 to validate our findings.

[2] For some metrics we only compare data between 2 CRAs, which vary in combination according to the metric. This is due to differing availability and comparability of metrics between CRAs.

7. Given the complexities of lending decisions it is difficult to assess harm arising from poor quality data. In this annex we show that our analysis has found material differences in data between the CRAs. For harm to arise that could be 'averted', lenders must be relying on CRA data that is sub-optimal (ie inaccurate, insufficient) and 'better' data must be available on that individual elsewhere which, had the lender relied on it, would have delivered a more efficient decision. We discuss this in Chapter 5 of this annex.

Scope of the analysis

8. In this annex we assess data quality in a number of dimensions:
   - In Chapter 2 we look at credit information coverage of the UK population. We compare depth of information by examining how many CRAs an individual is known to.
   - In Chapter 3 we compare a selection of credit scores offered by the 3 large CRAs. We also discuss individuals relative credit risk across CRAs.
   - In Chapter 4 we assess differences between the underlying data that CRAs hold on individuals. We also consider the possible causes of these differences.
   - Finally, we consider the impact of these differences in underlying data and scores on lending decisions in Chapter 5.

2 Population coverage and information depth

Introduction

9. In this chapter of the annex, we assess the quality of credit information in 2 aspects: by looking at the coverage of the population, and by examining differences in the depth of credit information held by different CRAs.

10. Our definition of coverage of the population means the proportion of the UK adult population for which CRAs hold credit information. If there is a large amount of the UK population who are not covered by credit information, there is a risk that many consumers will receive poor outcomes as lenders are less able to assess their credit risk.

11. Credit file depth is an important aspect of credit information. The more information a CRA holds on an individual, the more useful the credit file is in determining an individual's credit risk. Differences in the depth of credit information between CRAs can have a material impact on access to credit for individuals. If a CRA holds little information on an individual and another CRA holds more information, then the individual may receive different lending decisions based on the CRA chosen by the lender. Whilst lenders augment credit information that they receive from CRAs by using other sources, the depth of credit files still plays a significant role in lending decisions.

12. We identify individuals known to fewer than 3 large CRAs and discuss their characteristics. This enables us to examine whether certain groups of individuals, for example individuals in vulnerable circumstances, are more likely to be impacted by shallow credit files.

There are more IDs than the expected number of individuals at each of the 3 large CRAs

13. We wanted to examine the level of credit information coverage of the UK population. As discussed, poor levels of population coverage can result in poor outcomes for consumers.

14. We estimated the proportion of the population covered by the 3 large CRAs by comparing our sample sizes at each CRA with the number of individuals we would expect to see in the sample, shown in Table 1.

Table 1: Proportion of population covered by credit information at each CRA

CRA	Expected sample size	All IDs		IDs with thick files	
		Sample Size	***Ratio***	***Sample Size***	***Ratio***
CRA A	48,615	61,866	1.27	59,957	1.23
CRA B	48,615	75,521	1.55	52,916	1.09
CRA C	48,615	72,548	1.49	54,983	1.13

Source: FCA analysis on CRA and ONS data. Expected sample size is the 2019 Office for National Statistics (ONS) mid-year adult population estimate, adjusted down to reflect our methodology of sampling individuals born on one specific day

of the year in specific years. A small number of IDs have been excluded from this table based on geography, to ensure only individuals who would be included in the population estimate have been included.

15. Even if an individual is known to a CRA, a CRA might only have limited information on that individual ie a lack of depth of information or a 'thin file', as opposed to a 'thick file'. Thin file individuals have a shallow depth of credit file according to a CRA. Before discussing the coverage by thick files in Table 1, we first will discuss coverage of all files, thick and thin.

16. If all individuals were captured by all 3 large CRAs and there were no problems with matching records or measuring population, we should see a perfect match between individuals known to the 3 large CRAs and the expected sample size based on our estimates of the population.

17. A ratio equal to 1 would describe a perfect match. A ratio greater (smaller) than 1 means our sample is larger (smaller) than the expected sample based on ONS estimates.

18. There are opposing factors which can cause the number of individual IDs in CRA data to be larger or smaller than the number of actual people in the population. Firstly, the number of individual IDs could be smaller than the population because CRAs do not hold information on individuals who do exist. For example, the individual might not have credit information or other data which the CRA could use to create records. Alternatively, credit information may exist on an individual, but that information may not be shared with all 3 large CRAs.

19. Conversely, CRAs may have more individual IDs than the population because the CRA is not sufficiently sure that the records refer to the same person at a given point in their processes. As a result, an individual could be recorded multiple times. Other possible causes include the use of 'aliases' or alternative identities by consumers, underestimates of the true population by the ONS, and time lags on credit files for those gone-away or deceased (being captured in population estimates but not by CRAs).

There are more thick files known to each of the CRAs than the expected number of individuals in our sample

20. As CRAs vary in the precise definition of thin credit files we created a flag identifying whether an individual has a thin credit history. We consider that an individual has a thin file if the individual i) has 2 or fewer credit accounts (eg a credit card and a personal current account) and ii) those accounts were opened in the last 6 months. The results of this definition are included in Table 1.

21. Recent research by Experian revealed that over 5 million people in the UK are virtually invisible to the financial system.[3] It is not just those on the lowest incomes that are affected by the issue of financial exclusion. Thin file consumers come from a wide variety of backgrounds; ranging from young people who have no credit record (eg students), older people who may have paid off their debts or have limited use of credit, recent immigrants and expats.

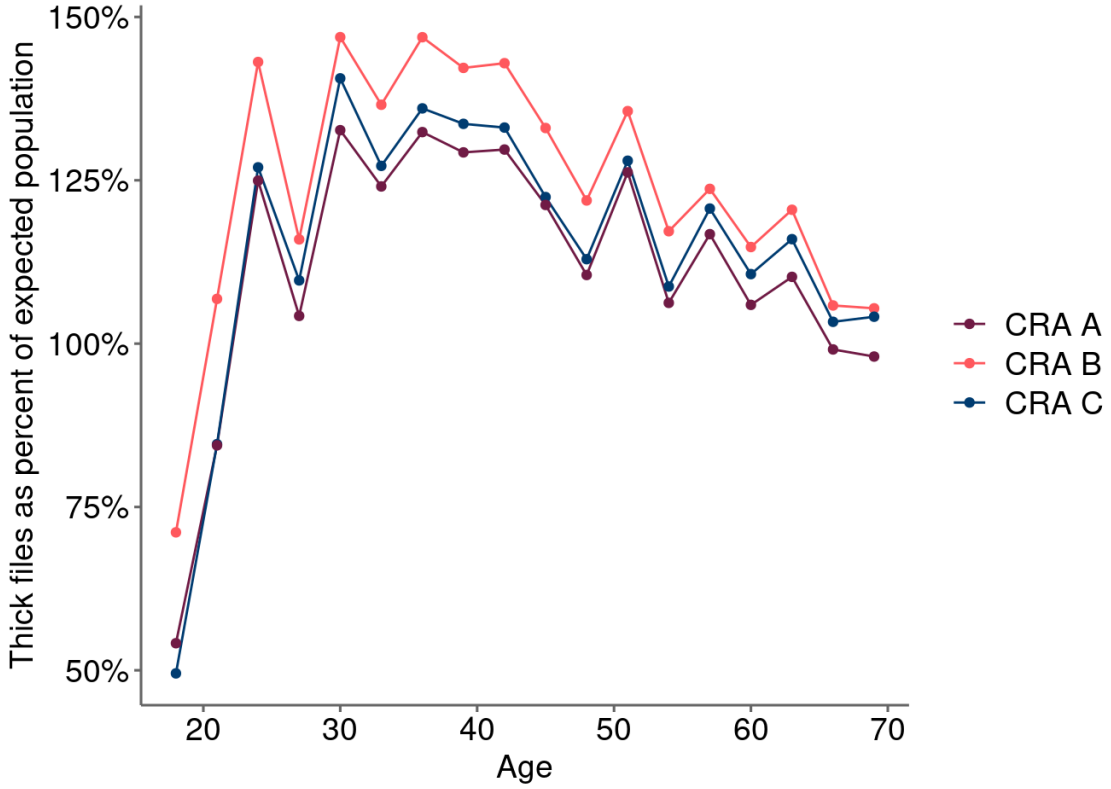
22. If a Credit Information User (CIU) receives only a thin file on an individual, then the CIU has the option of contacting another CRA to see if the other CRA has a thick file on the individual. This is called 'waterfalling' and the extent to which CIUs do this is

[3] www.experianplc.com/media/latest-news/2022/meet-the-5-million-credit-invisible-brits-still-at-risk-of-exclusion-from-the-financial-system/#:~:text=UK%2C%20March%2021%2C%202022%3A,about%20their%20financial%20track%20record

discussed in Chapter 5 of this annex on lending decisions. Thin files can still be problematic even if CIUs can get information from another CRA. For example, a CIU may not have arrangements with other CRAs, or the other CRA may also have a thin file on the individual (but the third CRA could have a thick file).

23. We also looked at how the ratio of thick files to population varies by age to examine whether age has an impact on the depth of credit files. This is shown in Figure 1. We see that the ratio varies by age. This is consistent with younger people being less likely to have accounts histories, and people opening fewer accounts as they reach retirement age.

Figure 1: The proportion of thick files as a percentage of expected population by age of individual



Source: FCA analysis on CRA and ONS data

We matched records across the 3 CRAs

24. We used information on names, date of birth and address to match IDs at CRAs to each other and create what we refer to as 'FCA IDs' for individuals. Specific discussion of the approaches we took can be found in the Technical Appendix to this annex.

25. Table 2 shows that the number of individuals with thick files known to all 3 large CRAs is around 96% of individuals in the UK who were eligible to be sampled. The number of individuals with thick files known to any 1 CRA is 30% greater than the number of individuals we would expect in our sample. This is consistent with the existence of 2 counteracting forces: that there are some individuals CRAs do not have information on, putting downwards pressure on estimated coverage ratios, and there are difficulties in matching individuals across records, causing multiple accounts to exist for some individuals, putting upwards pressure on estimated coverage ratios.

Table 2: Proportion of population covered by credit information (all FCA IDs vs. FCA IDs with thick credit history)

Individuals known to	Expected sample size	All IDs		Thick files	
		Sample Size	*Ratio*	*Sample Size*	*Ratio*
3 large CRAs	48,615	47,619	0.98	46,857	0.96
At least 2 CRAs	48,615	61,526	1.27	53,072	1.09
At least 1 CRA	48,615	99,459	2.05	63,382	1.30

Source: FCA analysis on CRA and ONS data. In the case where an individual is known to all 3 large CRAs, we record the individual as having a thick file if at least 2 of the 3 CRAs have the individual as having a thick file. A small number of IDs have been excluded from this table based on geography, to ensure only individuals who would be included in the population estimate have been included.

Analysis of relative credit information coverage suggests that some groups of people are more likely to have less coverage

26. Even though we can observe more accounts than individuals, there might still be many individuals who do not have credit information at all CRAs. To explore this issue in more detail we discuss relative coverage.

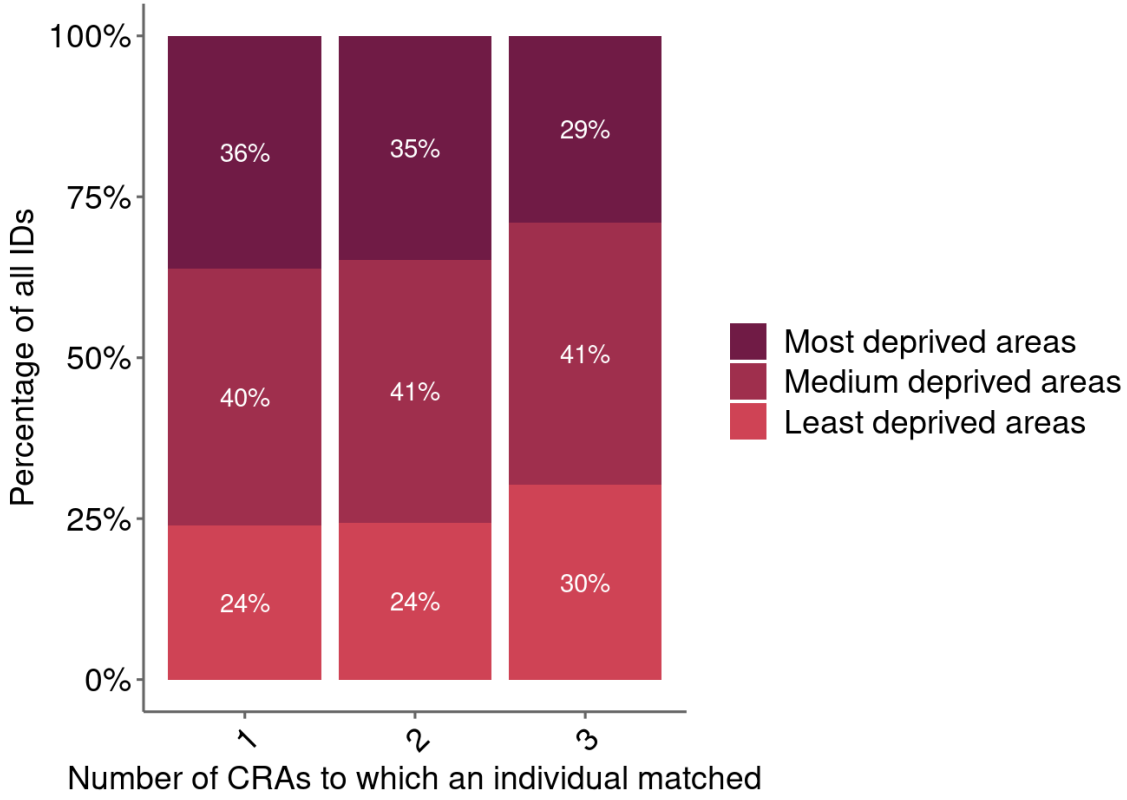
27. We wanted to check that certain groups of people were not more impacted than others by issues with information coverage if they were to apply for credit. The impact of poor credit information coverage on consumers who are in vulnerable circumstances can be greater than other consumer groups as they often have fewer credit options and denial of credit or worse credit terms impacts proportionately more on their standard of living. So, we wanted to check that coverage issues did not disproportionately affect those groups of consumers.

28. To do this we examined the proportion of individuals who were matched to 1, 2 or 3 CRAs by Index of Multiple Depravation (IMD) category.[4] The results can be seen in Figure 2.

29. The results show that individuals who matched to fewer CRAs tend to live in more deprived areas. Around 36% of individuals who matched to only 1 CRA live in the most deprived areas in England, compared to 29% with who matched to all 3 large CRAs.

[4] The IMD, which is produced by the Ministry of Housing, Communities and Local Government (MHCLG) and is available for England only. The IMD ranks every Lower Layer Super Output Area (LSOA) from 1 (most deprived area) to 32,844 (least deprived area). LSOAs are designed to be of similar population sizes (they contain an average of 1,500 residents) and are sometimes referred to as neighbourhoods in UK official statistical releases. The IMD combines information from 7 deprivation measures: income, employment, education and skills, health, crime, housing, and living environment.

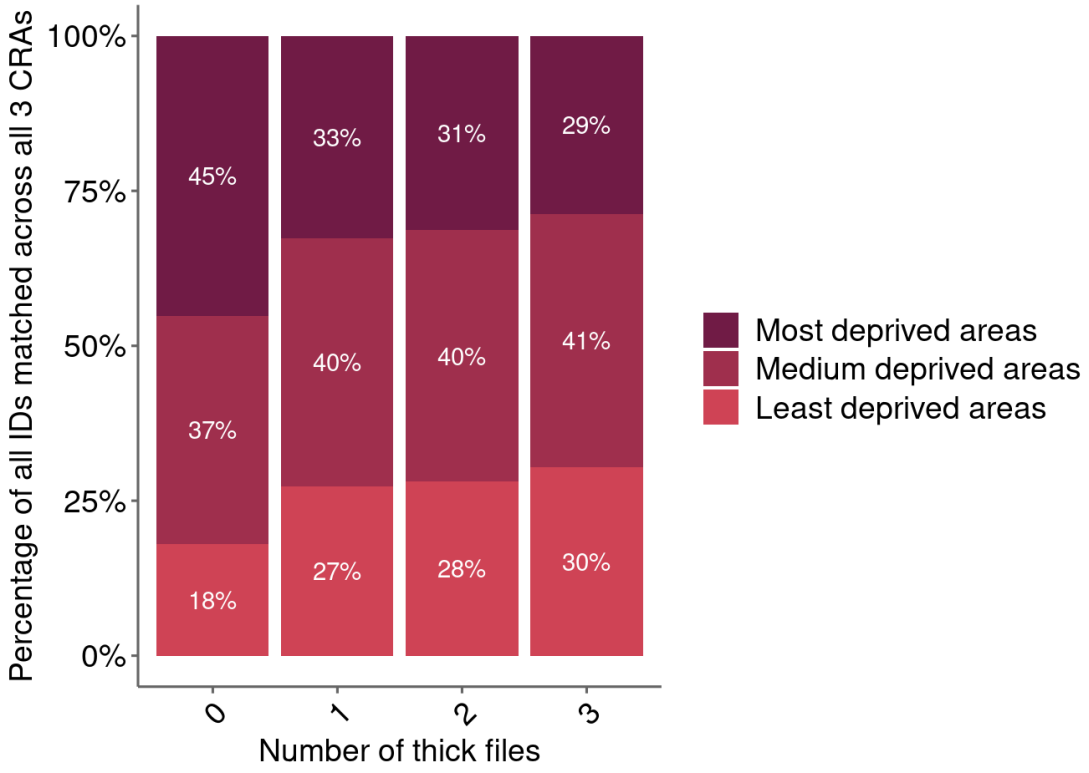
Figure 2: Proportion of FCA IDs living in deprived areas according to matching



Source: FCA analysis on CRA data

30. We also extended the analysis in Figure 2 to examine proportion of individuals that were covered by thick file credit information at 1, 2 or 3 CRAs by IMD category. This was to examine whether people living in more deprived areas were more likely to be thin filed. If a consumer has a thick credit history at 1 CRA and a thin credit history at the other 2 CRAs then they may receive different lending decisions based on the CRA used by the lender. This is also an indicator that the CRAs hold different information on the individuals. Figure 3 displays this analysis.

Figure 3: Proportion of FCA IDs living in deprived areas according to matching and thick files



Source: FCA analysis on CRA data

31. For individuals who matched to all 3 CRAs but where there was no thick file at any CRA, around 45% were in the most deprived areas, compared to a 29% with thick files at each CRA. This indicates that individuals more at risk of thin files coverage are more likely to live in more deprived areas than those with thick files with all the CRAs.

32. The data here, combined with the data on pre-matching coverage presents the following picture:

- As set out in Table 1, individual CRAs hold information on more IDs than there are people in our expected population. Much of this is driven by thin files which CRAs do not have sufficient data to match to other files, and when these are excluded the number of IDs is closer, but above, the number of people in the population.
- As set out in Table 2, when we match IDs across CRAs we have a similar picture to what we observe at individual CRAs – more IDs than individuals, but with the number of IDs with thick files at each CRA being similar to the expected population.
- This does not preclude lack of coverage being more likely for many individuals and can disproportionately affect those who are older, younger, and who live in more deprived areas.

33. The fact that CRAs have more accounts than individuals has implications for individuals whose underlying data is not contained in the same file account. An individual who at a given CRA has multiple accounts which reflect their information risks having that information not reflected in their credit information. Unmatched thin files do not include substantial information from CIUs and so the effect of this on individuals may be limited, so long as the CRA can match the individual to the ID with the most

information. If a CRA does not have sufficient information on an individual on a credit file, a credit search on that individual will either not return anything or return a thin file, which could affect their ability to get credit. The drivers of lack of coverage are a combination of lack of existing credit information at a given CRA (if for example a lender only reports to a single CRA), but another one is the absence of credit information.

3 Comparison of scores

Introduction

34. The credit scores provided by CRAs reflect consumers' credit risk and are among the factors that lenders might use in their decisioning processes.

35. In this chapter, we compare a selection of credit scores offered by the 3 large CRAs including those scores provided most commonly to lenders and those provided (sometimes via third parties) to consumers. We first describe what a credit score is. We then present the distribution of a selection of the latest generation of scores at the time we requested the information from the CRAs. We also show how the same individuals may have different relative credit risk according to different CRAs.

36. CRAs offer many credit scores, which can differ by factors including which market they target and whether they reflect information on financial associates (for example other people who opened joint accounts with the individual).

37. In order to compare scores between CRAs, we can take the rank of score of an individual as compared to other individuals.

38. Identifying differences, even large differences, in the ranks of similar scores between CRAs does not necessarily mean that these differences reflect competition working poorly. Material differences in similar scores between CRAs are a necessary but not sufficient condition for exploring whether some data practices lead to poor market outcomes. We explore in the following chapters of this annex whether differences in similar scores across CRAs are a result of differences in the credit information data used to generate these scores.

39. If 2 scores with the same definition, associated to the same individual from 2 different CRAs differ, this could be due to:
    - differences in the algorithm used to calculate the score
    - differences in the underlying data used to calculate the score

40. The algorithms used are CRAs' proprietary information and will differ across CRAs. Differences between CRA scores due to algorithmic differences would be consistent with a well-functioning credit information market as it would indicate that the CRAs are competing on the quality of their algorithms to better predict an individual's credit risk.

41. The underlying data may differ if CRAs obtain data from different data contributors, if data contributors provide different data to the CRAs, or if CRAs have different matching techniques. It would be concerning if the data differences are the result of differences in the data that CRAs receive from lenders. Data differences driven by differences in input contributions could indicate that some CRAs may not be able to obtain the inputs needed to produce effective scores.

CRAs offer a variety of credit scores to CIUs

42. The 3 large CRAs in the UK typically offer a range of credit scores to help lenders and other users assess individuals' credit risk. Other credit scores provided by CRAs are

used to inform other decisioning processes, such as affordability and account management.

43. As discussed in the main report, some lenders (usually smaller lenders) rely more heavily on the credit score provided to them by the CRAs to inform lending and other decisions. Lenders also use raw and summary data provided by the CRAs as well as credit scores to create their own scorecards. Whilst not all lenders directly use scores in their decisioning processes, scores are a proxy for the underlying credit information held by CRAs and are indicative of an individual's credit risk, so it is important to examine the extent to which scores differ across CRAs.

44. A credit score typically estimates the probability of the applied for account going 'bad' within a certain period. Some scores are predicated on whether the consumer will go 'bad' on any account they hold. The time period for going 'bad' will differ according to product type, for example mortgages will have a longer time period.

45. Credit scores may vary according to the definition of an account going 'bad' used, as discussed above, and several other factors. These include:
    - The time period considered. This is typically 12 or 15 months, but for some scores it can be shorter (eg 6 months for short term lending) or longer (eg 20 months for some markets such as mortgages).
    - Whether they include information on financial associates (eg someone who opened a join account with the individual).
    - Whether they focus on a specific market (eg mortgages, telecoms) or they are an all-markets version.
    - Whether they include other indices such as an affordability or indebtedness index.

Scores offered to CIUs vary materially by CRA

46. We wanted to assess the differences between the scores offered by CRAs to CIUs. As scores are a proxy for the information on an individual's credit file, large differences in relative score across CRAs may indicate large differences in underlying data. These differences can lead to different lending decisions dependent on the CRA used. To do so, we first focused our analysis on scores that have similar features across CRAs. Specifically, we considered those that:
    - are all-market versions
    - include information of financial associates
    - do not include either indebtedness or affordability indices where relevant

47. For this analysis our sample included only IDs matched by all 3 large CRAs. As the sample of individuals was the same across CRAs we would expect to see a similar ranking of individuals, even if the range of scores differ, if the CRAs are providing a similar picture of an individual's credit risk. We further limit the analysis to uniquely matched individuals (eg exclude where one ID at one CRA maps on to 2 IDs at another CRA).

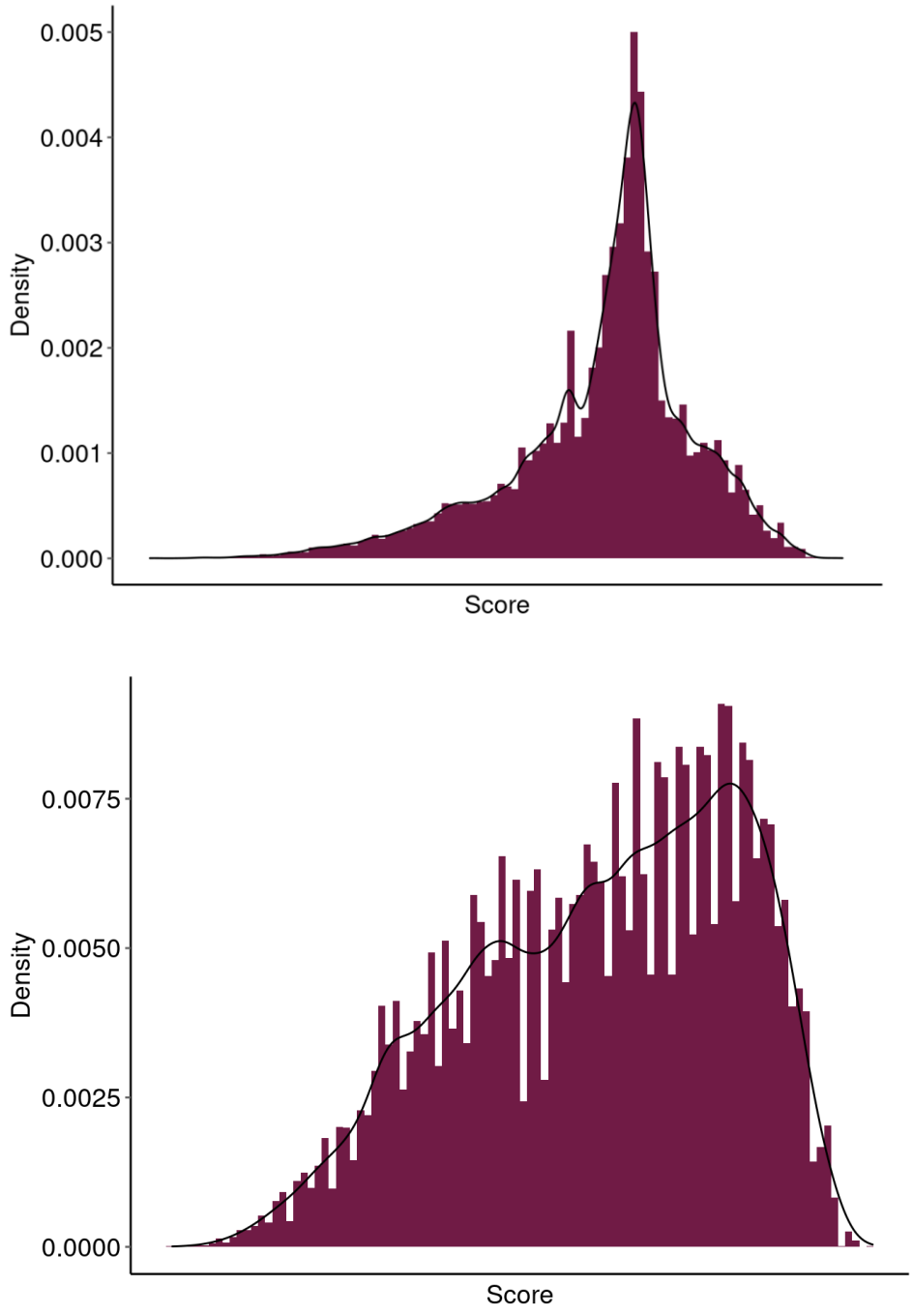
48. To enable comparison of scores across CRAs, we ranked individuals in our sample according to the given score from a given CRA. We compared scores provided to CIUs in 2 ways:
    - the distribution of scores
    - comparing where an individual may rank in 2 different CRA scores

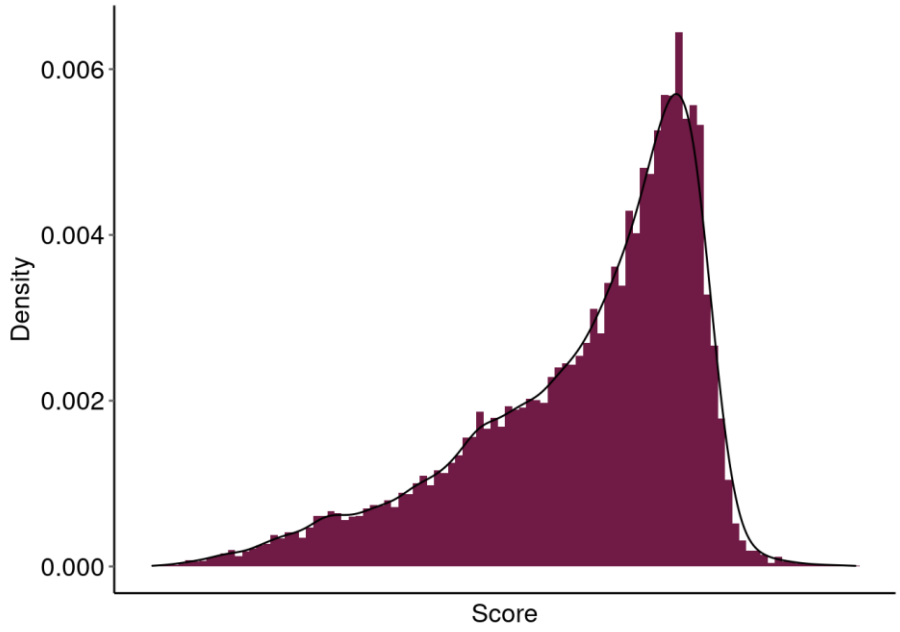
Score distributions vary by CRA

49. For a given type of score, different CRA can have different minimum and maximum possible scores. The scores also have different distributions of customers across those ranges. The figures in this section compare selected scores for the individuals in our sample that are uniquely matched across the 3 large CRAs.

50. We compared the dispersion of credit scores by CRA by examining the distribution of the scores. Figure 4 shows the distributions of 3 selected scores. Also, the distributions are skewed to the right. This indicates a concentration of high scores, and relatively more dispersion for lower ones.

Figure 4: Distributions of a selection of the latest generation of scores





Source: FCA analysis on CRA data

Individuals' relative indicative credit risk differs across CRAs

51. If individuals vary in credit risk according to CRA it implies that either underlying data or algorithms used to calculate the scores are different. Differences in credit risk perceptions across CRAs can lead to harm for consumers as detailed in Chapter 5 of this annex.

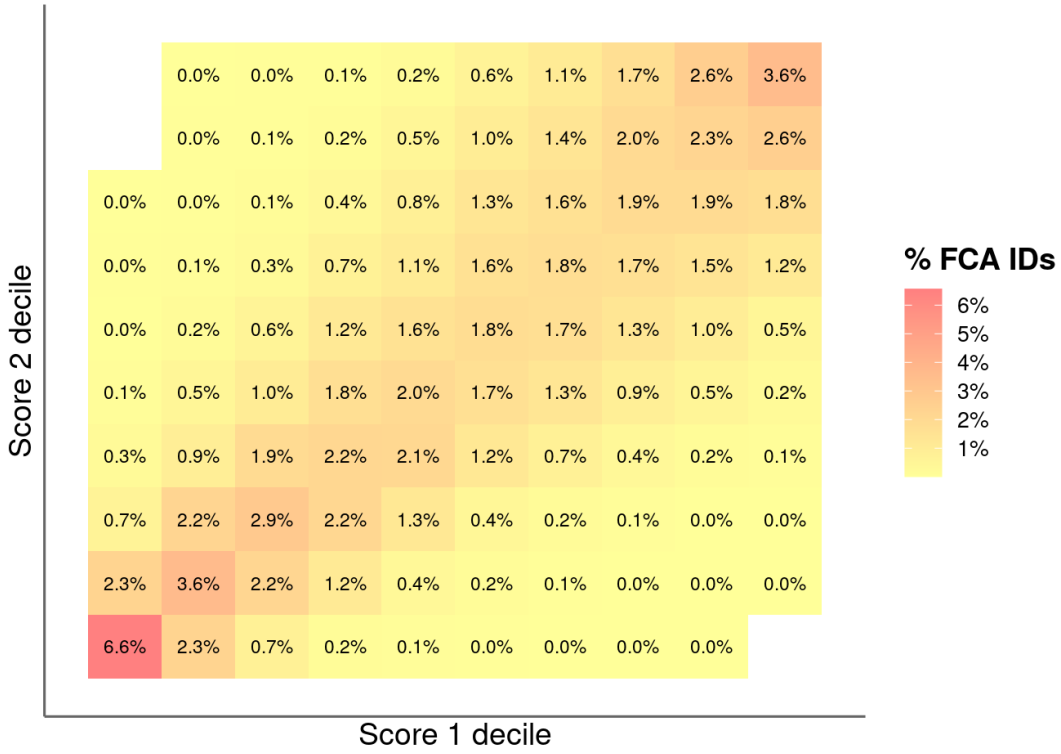
52. We examined whether the same individual is in a different ranking position in different CRAs according to the selected scores. As explained earlier in this chapter, we are able to undertake this by limiting our sample only to individuals who were uniquely matched across all 3 CRAs.

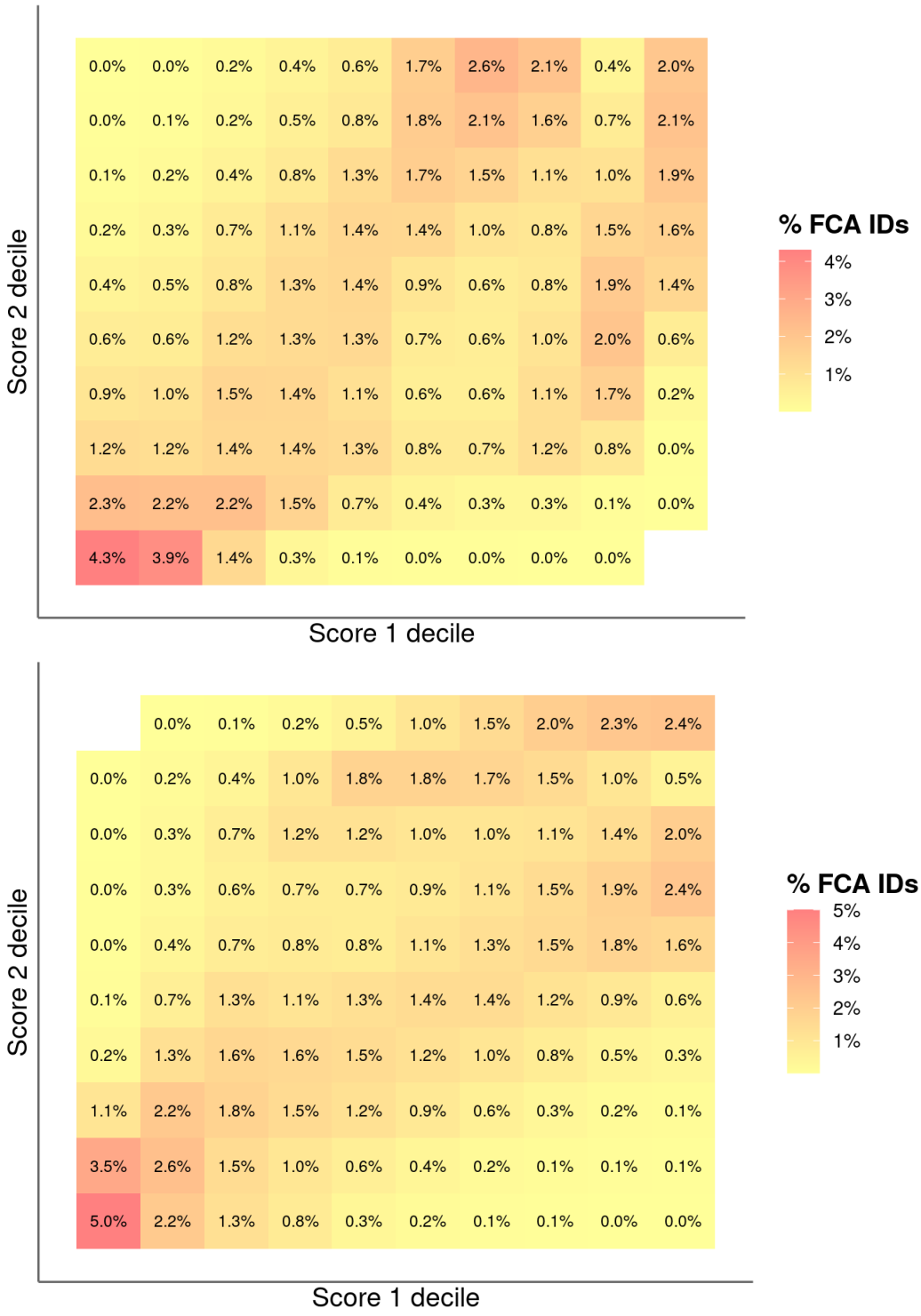
53. For example, an individual who is at the 4th decile of a score indicates that they have a better credit risk than 40% of the individuals in the sample. Our analysis allowed us to check if the same individual was in the 4th decile according to an alternative CRA. We then compared each possible pair of scores.

54. Figure 5 compares a score from each CRA in each possible combination of the 3 large CRAs. Each panel shows a comparison between 2 CRAs, and each cell shows the percentage of individuals in our sample for a particular combination of score deciles. The sum of all the cells in a panel is 100%. If a cell is missing, it means there are no individuals with that combination of scores for the 2 CRAs. If the individuals' scores were in the same decile across the 2 CRAs we would see dark red boxes diagonally from the bottom left to the top right, each showing 10% of the matched sample.

55. The scores used here are all-market scores which use information on financial associates. Scores which use information on associates are referred to as opt-in scores.

Figure 5: Pairwise comparison of opt-in scores

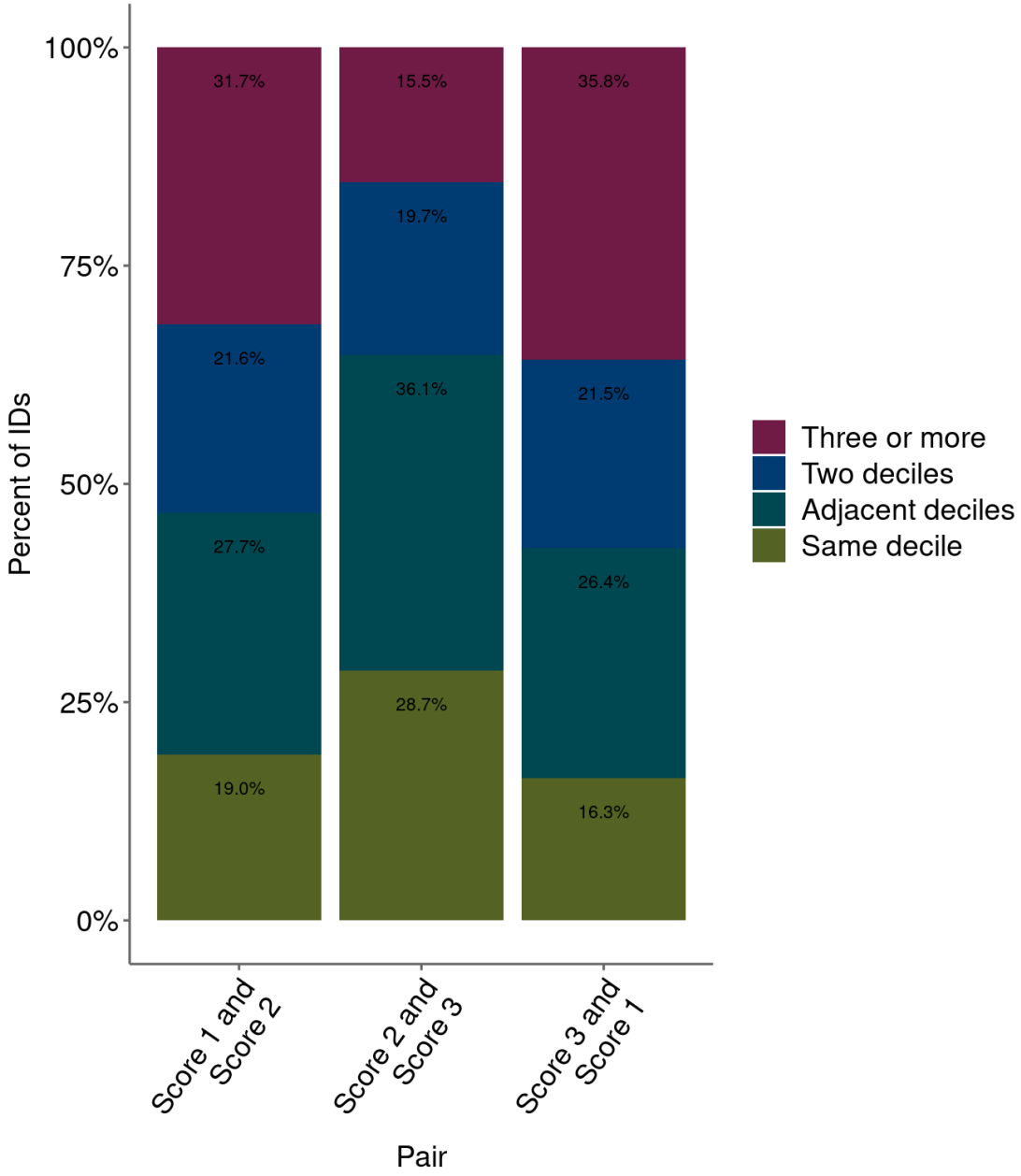




Source: FCA analysis on CRA data

56. Figure 6 illustrates the findings from Figure 5 and summarises the degree of correspondence between scores from different CRAs. Between 16.3% and 28.7% of individuals in our sample are in the same decile according to both scoring models, and between 15.5% and 35.8% have scores that were 3 or more deciles away from each other. This indicates that, on average, individuals are ranked differently in terms of credit risk across CRAs.

Figure 6: Pairwise comparisons of opt-in scores



Source: FCA analysis on CRA data

57. Figures 5 and 6 show that there is a large dispersion across the spectrum of credit risk between CRAs. There is greatest correspondence for individuals in the bottom decile of credit scores. This is unsurprising given a general agreement between CRAs of what makes an individual have an extremely high credit risk. We discuss differences in the data that drives these scores between CRAs in Chapter 4. For all other deciles, we observe significant dispersion of credit scores.

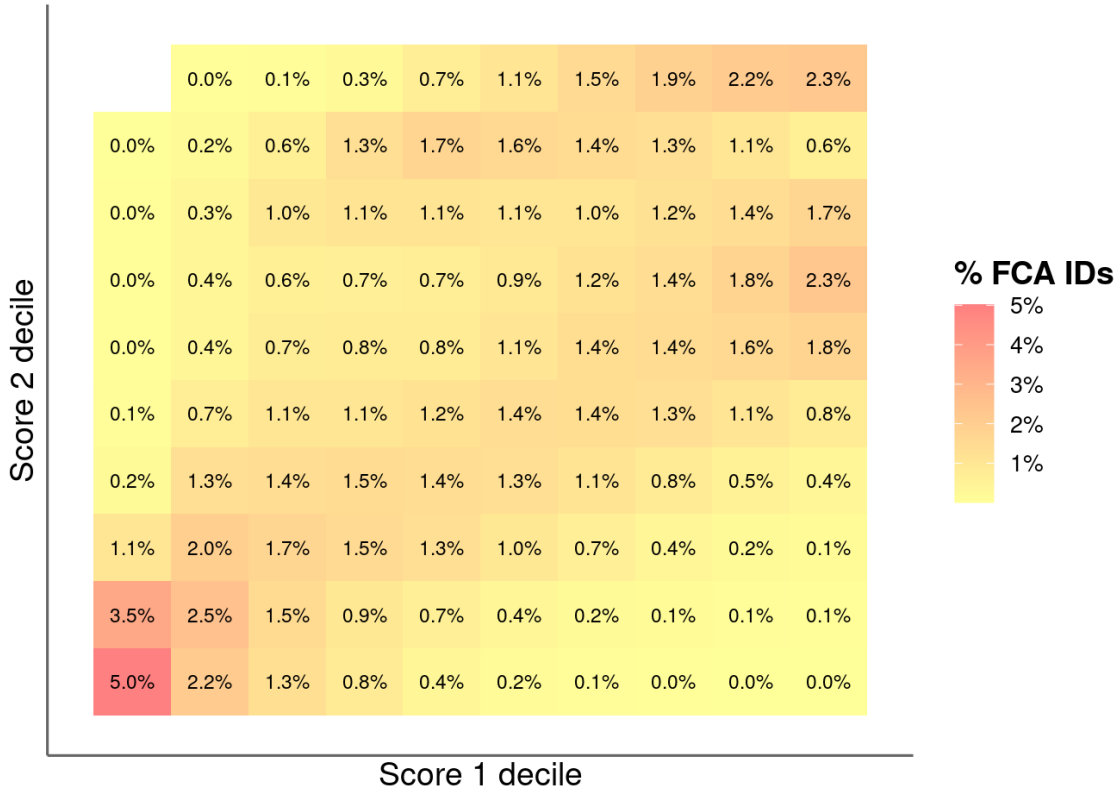
58. Overall, our findings show that the assessment of an individual's relative credit risk can differ significantly across CRAs. This means that a lender may have a different perception of a customer's credit risk depending on the CRA used.

Other types of credit scores provided also differ across CRAs

59. In the previous section we focused on the differences between opt-in scores across CRAs. Opt-in scores are just one type of score that CRAs produce and provide to CIUs. This section compares other types of scores that CRAs provide both within CRAs and across CRAs.

60. Scores which do not reflect information on associates are referred to as opt-out scores. Figure 7 below shows the pairwise comparisons between the main opt-out scores at 2 of the large CRAs where comparable opt-out scores were available.

Figure 7: Pairwise comparison of opt-out scores



Source: FCA analysis on CRA data

61. The comparisons of opt-out scores are similar to the comparisons of the opt-in scores, suggesting that different information on associates, or how the data affects scores, is not primarily driving differences in opt-in scores between CRAs.

62. We also looked at the consistency of opt-in and opt-out scores within CRAs. This was to examine the impact of associate information on an individual's credit score. It is important to examine how closely these scores align with opt-in scores to see the extent to which associate information alters an individual's credit score.

63. Figure 8 below shows pairwise comparisons of the opt-in and opt-out scores within 2 of the large CRAs where comparable opt-out scores were available. Unsurprisingly, these are more similar than the cross-CRA comparisons. We see that generally for an individual their opt-out and opt-in scores at a given CRA are very similar, as there is little dispersion around the diagonal.

Figure 8: Pairwise comparison within 2 CRAs of opt-in and opt-out scores

Source: FCA analysis on CRA data

64. CRAs also offer scores directly or through providers of Credit Information Services (CISPs) scores to consumers for educational purposes. These are known as educational scores.

65. Consumers can face harm if they have a different view of their credit risk to a lender. Consumers could either overestimate or underestimate a lender's assessment of their

credit risk and apply for unsuitable products or to lenders that offer less favourable terms than those they might qualify for. Furthermore, consumers may face harm if they believe that they are in a better credit position than they are. A result of this may be that they therefore do not take action to improve their credit score when it would be in their interest to do so.

66. We have data on educational scores for two of the CRAs. For one of the CRAs the educational score and opt-out score are the same while for another CRA, there is a very different scoring methodology which truncates scores above a certain level. This means that the scores the individual can be quite different in information from either the opt-in or opt-out scores provided by the CRA to CIUs.

67. The US Consumer Financial Protection Bureau (CFPB) looked at differences between consumer and creditor-purchased credit scores in the US.[5] The purpose of the CFPB's analysis was to quantify the harm related to consumers' inaccurate perception of credit risk. As explained above, a customer may apply for unsuitable products or accept unfavourable terms if their perception of credit risk differs from the lender's assessment. The CFPB found that for a substantial minority of customers different scoring models gave meaningfully different results, and our findings are in line with this.

Some CIUs do not face strong incentives to share data with multiple CRAs

68. If data contributors do not share data across all CRAs, a CRA may not observe the entire credit history of a given individual. For example, a CRA may not be aware that an individual has opened a high-cost short-term credit account and therefore may not be able to factor this in the calculation of the individual credit score.

69. We look at incentives for CIUs to share data with multiple CRAs as high levels of data sharing lead to more comprehensive credit information. Incentives to share data lead to more data being shared and better quality credit information for individuals which in turn leads to better lending decisions. CIUs have a strong incentive to share information with at least one CRA, as they are required to do so to get credit information from that CRA. Similarly, if a lender uses more than one CRA, for example because they fall back on another CRA if one CRA only has a thin file, or because they use data for multiple CRAs for an application, then they will also share information with multiple CRAs.

70. The requirements on lenders to share data reflect the Principles of Reciprocity.[6] These principles require CIUs using credit information to share data with at least those CRAs from which it purchases credit information but does not require CIUs to share data with all 3 large CRAs, although this is recommended.

71. CIUs still may have incentives to share their data with CRAs they do not receive information from. Sharing information can make switching CRAs easier or the threat of switching more credible, and CIUs, particularly larger ones, may view their sharing practices as part of general market integrity behaviour.

72. However, we know that many CIUs do not share data with all large CRAs, as discussed in Chapter 4 of this annex. This reflects the fact that there can be additional costs to share data with additional CRAs.

[5] https://files.consumerfinance.gov/f/201209_Analysis_Differences_Consumer_Credit.pdf

[6] https://www.scoronline.co.uk/principles/

73. Data contributors may incur additional costs when sharing data with additional CRAs, compared to the situation where they share data with only one CRA. Each CRA has its own format to receive data, however some (but not all) lenders we spoke to told us that CRAs accept data in other CRAs' format. A contributor sharing data with all 3 CRAs typically shares the same data with all 3 large CRAs in the same format.

74. Some respondents to the RFI also identified risks in sharing with more than one CRA because of increased likelihood of compromising customer data.

75. In addition to the incentives to share data with multiple CRAs, costs may also arise when dealing with data queries and disputes from CRAs and consumers. For example, if a contributor submits credit information to all additional CRAs, it may have to deal with requests of correction from consumers from 3 sources and will have to correct data with all 3 large CRAs.

4 Comparison of information on individuals

Introduction

76. CIUs can receive the information on individuals which CRAs use to generate credit scores from the CRAs. As part of our data request we also received this data from the 3 large CRAs. We received data on the data metrics that CRAs generate on individuals, and the underlying data that is used to generate these data metrics: CATO data, search data, electoral roll data and public data. We also received income data from 2 of the CRAs.

77. In this chapter, we discuss and assess potential sources of data differences. First, we analyse the extent of differences between the data on individuals we received from the 3 large CRAs. We then describe the possible reasons why data may be inconsistent between CRAs and discuss the extent to which differences are caused by data contributors sharing data with only 1 or 2 CRAs.

78. Data differences are a factor in credit score differences. As discussed previously, credit score differences amongst CRAs can lead to inefficient or inappropriate lending decisions and harm for consumers. It is therefore important to analyse differences between the credit information that CRAs hold on individuals.

Individuals credit information data metrics can vary across CRAs

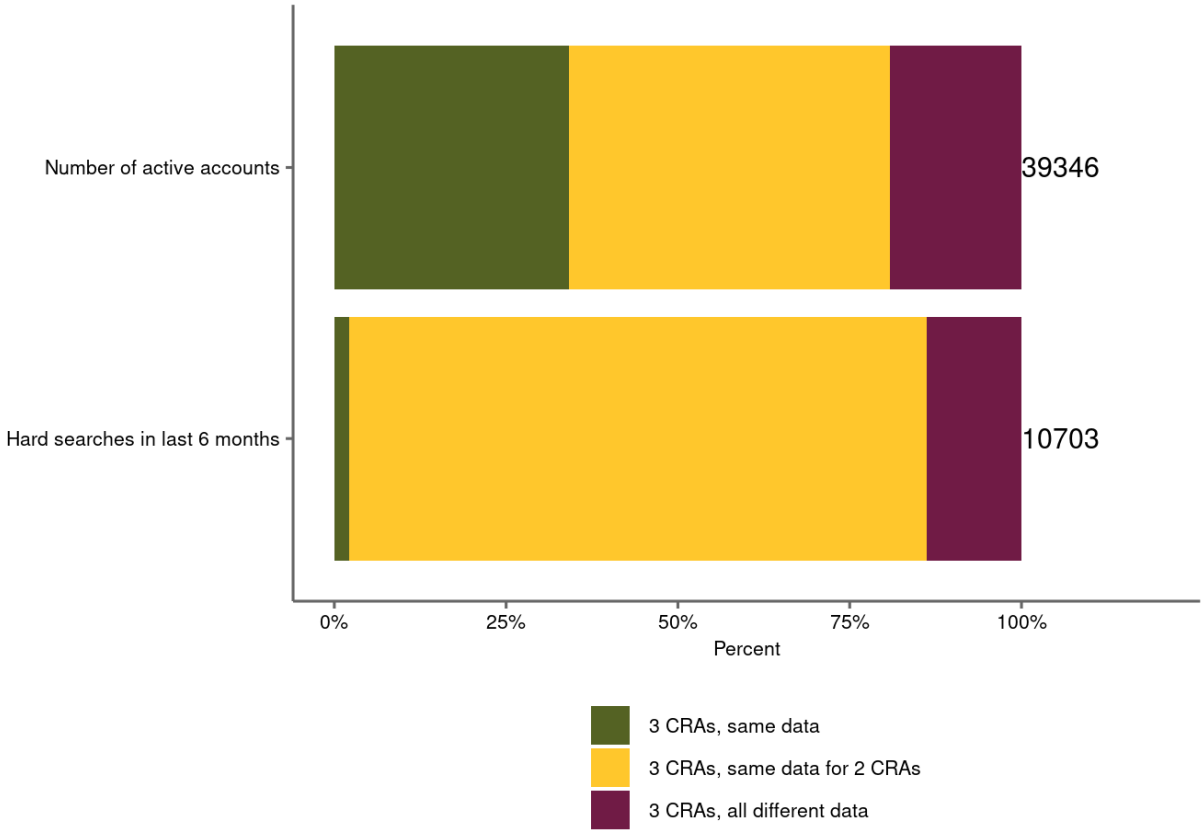
79. CRAs develop summary credit information data metrics about individuals from the raw credit information they collect. The data metrics that CRAs define and use are not necessarily the same between CRAs. For example, a CRA may summarise the raw data in a different way to other CRAs on the worst status on an active account in the last 3 months. CRAs may have different definitions of active accounts or worst status.

80. Our comparison of data across CRAs is limited to individuals we have uniquely matched across all 3 CRAs, giving a sample size of 39,809 individuals.

81. When comparing a given feature, we have excluded instances where no CRA has a positive observation from the analysis. For example, the comparison data on County Court Judgments (CCJs) is only for individuals where at least one CRA records a CCJ, as otherwise the data would be broadly identical, as most individuals do not have a CCJ. In addition, we allow some minor deviation of 5% for numerical data such as for income, before marking 2 CRAs as having different information.

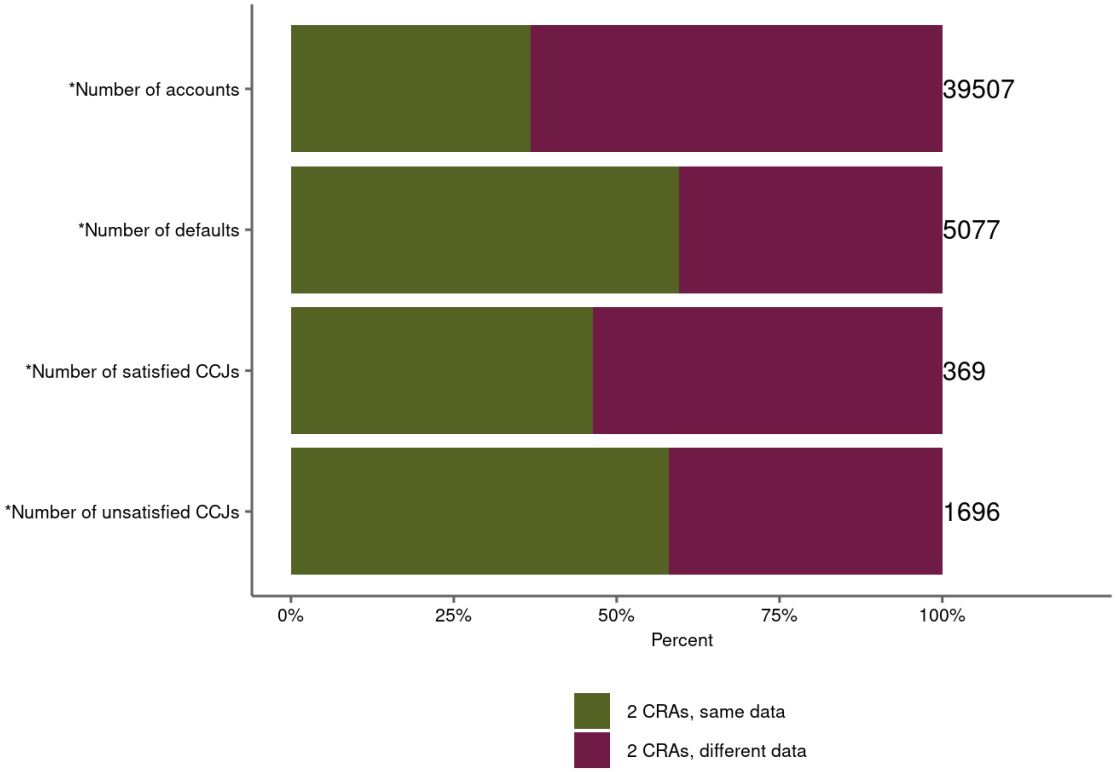
82. Figures 9 compares the data between CRAs for a selected set of data metrics that the CRAs generate on individuals. Figure 10 compares data between CRAs where only 2 CRAs have comparable information. The number to the right of each bar is the number of individuals who have been included in the calculation, according to the exclusion mechanism discussed in the previous paragraph.

Figure 9: Differences between the 3 large CRAs on a set of selected data metrics



Source: FCA analysis on CRA data

Figure 10: Differences between 2 of the large CRAs on a set of selected data metrics

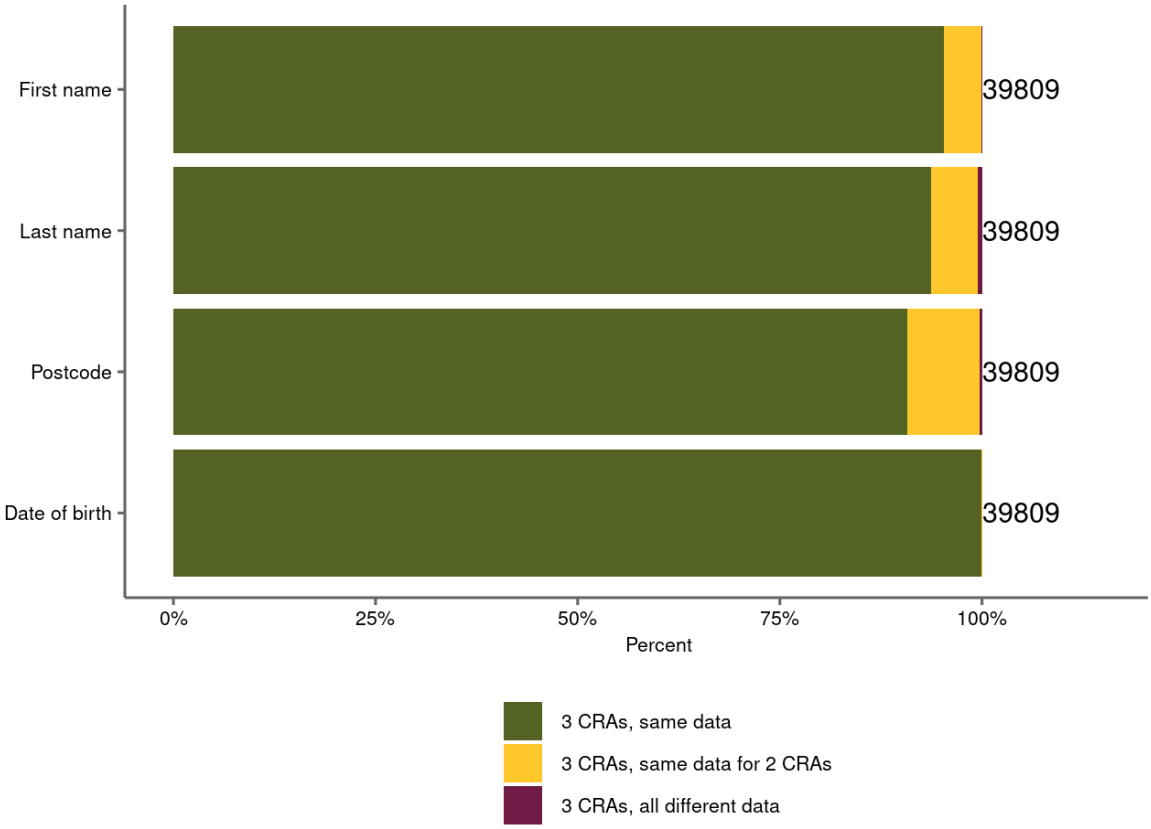


Source: FCA analysis on CRA data. *Comparison only includes information from 2 CRAs, however it is not necessarily the same two CRAs in all comparisons.

Individuals' primary information is broadly consistent across CRAs, however some individuals still see differences

83. We also compare the primary personal identifiers for each for each of the matched individuals in order to assess the level of similarity across CRAs. Differences in primary personal identifiers could lead to an individual not being found in the CRAs database when lenders ask for their information. This is shown in Figure 11. As individuals can have multiple identifiers associated with their account (such as historic postcodes) the primary details are not always the same for individuals, even across matched individuals.

Figure 11: Differences between the 3 large CRAs on a set of selected personal identifiers



Source: FCA analysis on CRA data

Individuals' raw credit information data can vary across CRAs

84. As part of our data request, we also received the data underlying the data metrics. This includes information on accounts, current account turnover (CATO), search data, electoral roll data, CIFAS data, and other public data.

85. Where CRAs have created different data metrics on an individual, this does not necessarily mean that the CRAs hold different underlying data on an individual. As discussed previously CRAs may summarise their raw data differently, leading to differences in the generated data metrics. Therefore, it is important to examine the underlying data.

86. Lenders use a wide range of data in their decisioning, therefore just examining the most similar data metrics across CRA is only of limited applicability. It is for this reason that we examine the raw, underlying credit file data.

87. Differences in data can also lead to significant harm for consumers. Table 4 below shows the percentage of lenders that told us in the BiFD request for information that they would automatically refuse a credit application for at least some of their products if the following factors were present. If a consumer has differing data between CRAs on any of these factors, it could lead to substantially different lending decisions and harm for the consumer. This includes data on Individual Voluntary Arrangements (IVAs).

Table 4: Percentage of lenders automatically refusing lending for the stated factors

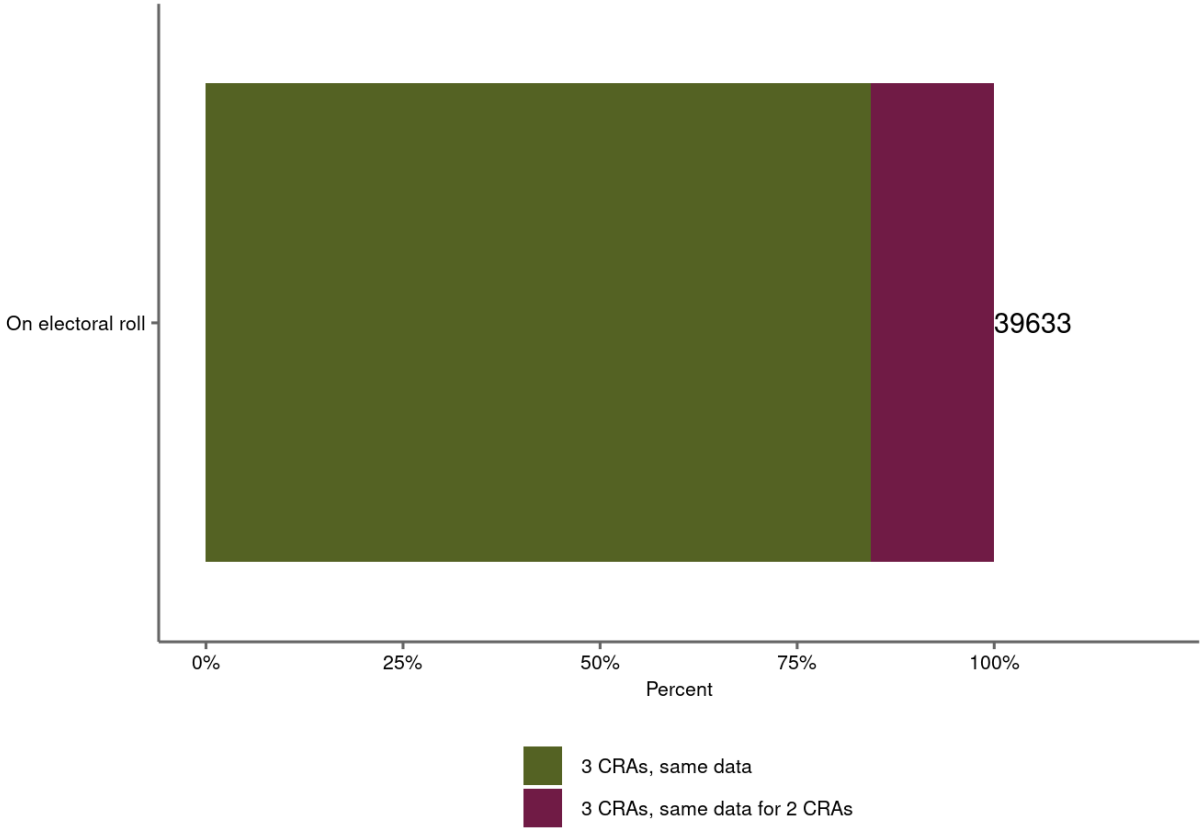
Factor	Percentage of Lenders automatically refusing a credit application
Individual currently in an IVA	94%
Unsatisfied CCJ in the last year	45%
Satisfied CCJ in the last year	39%
Default in the last year	27%
Unsatisfied CCJ in the last 6 years	15%
Satisfied CCJ in the last 6 years	12%
Bankruptcy in the last 6 years	67%
IVA in the last 6 years	70%

Source: FCA analysis on firm data gathered for BiFD RFI. Figure 15 below covers consistency of public data such as CCJs. Defaults are covered in Figure 13.

88. Furthermore, lenders also told us in our RFI that they take a variety of different factors into account, either separately or together, as well as those listed in Table 4 when assessing applications, such as presence on the electoral roll, the number of hard searches present on a credit file in certain time frame, the number of accounts an individual holds and levels of current debt. It is therefore important that we examine a variety of underlying credit information data held by the 3 large CRAs. Material differences in the data could lead to different credit application outcomes for individuals.

89. One data set we have received from the CRAs is electoral roll data. We examined the similarity of electoral roll appearance at the current address for individuals between CRAs. Figure 12 below demonstrates that for around 85% of individuals, CRAs are in agreement about whether or not the individual is on the electoral roll.

Figure 12: Differences between the 3 large CRAs on electoral roll data

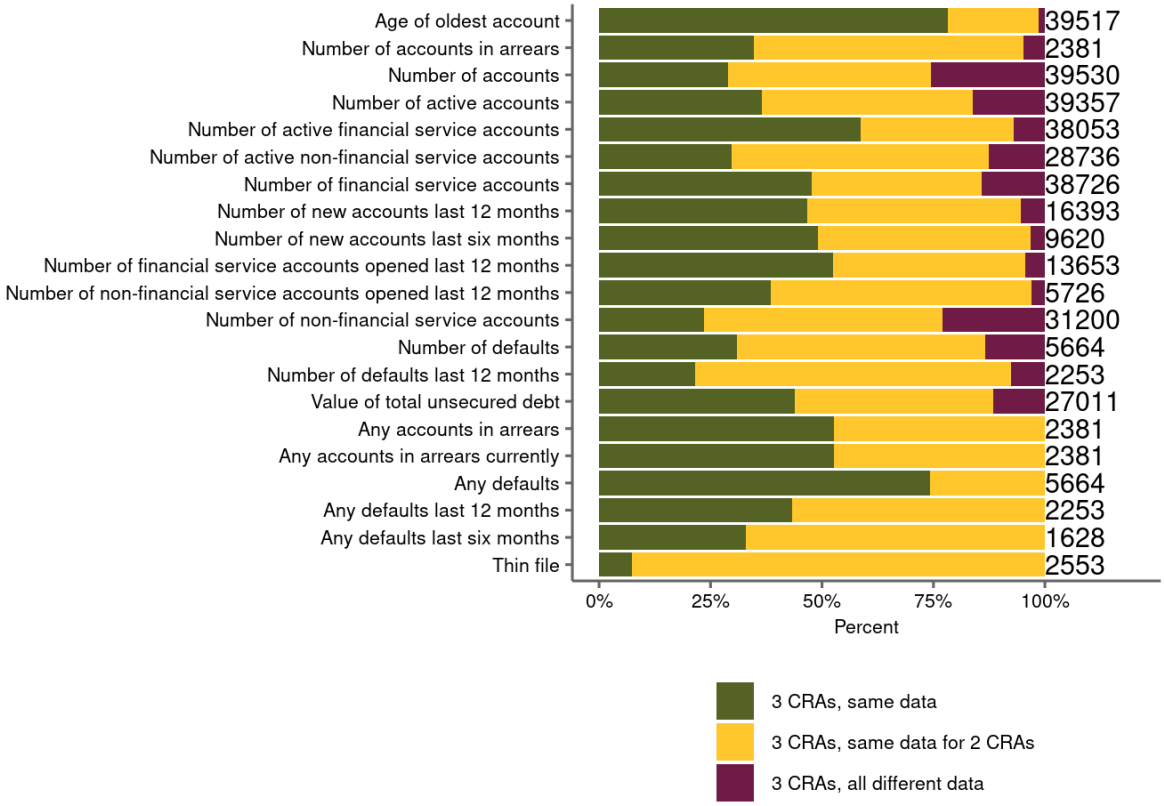


Source: FCA analysis on CRA data. For binary data (such as `On electoral roll') it not possible for 2 CRAs to agree while a third disagreeing, as seen in other comparisons.

90. We have also examined differences in account data across CRAs. This includes data on the number of active accounts, the age of accounts, number of defaults and total debt levels for a given individual. This dataset is made up of the main CAIS, SHARE and INSIGHT databases and it underlies many of the data metrics that CRAs create as many data metrics relate directly to accounts. Data differences between CRAs in this raw dataset could therefore lead to large differences in the data metrics generated (and shown in Figure 9 and 10).

91. Figure 13 below shows the differences between the CRAs on underlying accounts data. This figure shows that for some individuals, CRAs agree on things like the number of accounts, but for many individuals this is not the case. As this is driven by differences in underlying data, it is important to understand what type of accounts are driving this. This is useful for informing potential remedies for data gaps (ie missing information from an individual's credit file). We attempt to address this question in a later part of this chapter of the annex.

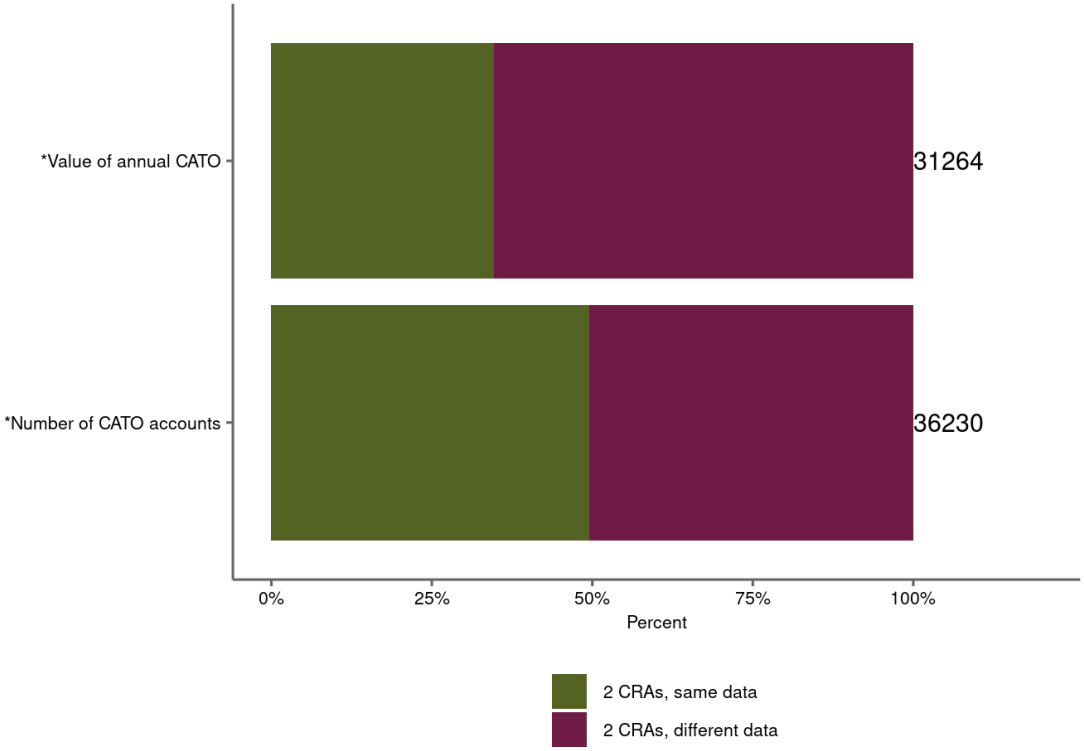
Figure 13: Differences between the 3 large CRAs on a set of selected account data



Source: FCA analysis on CRA data

92. In addition to accounts data, we also explored to what extent CATO data was the same between CRAs. We explored this for 2 CRAs. As Figure 14 below sets out, CATO data is frequently different for an individual depending on the CRA. This can potentially be explained by the differences seen in the number of accounts each CRA records, and differences in values for a given account.

Figure 14: Differences between 2 of the large CRAs on CATO data

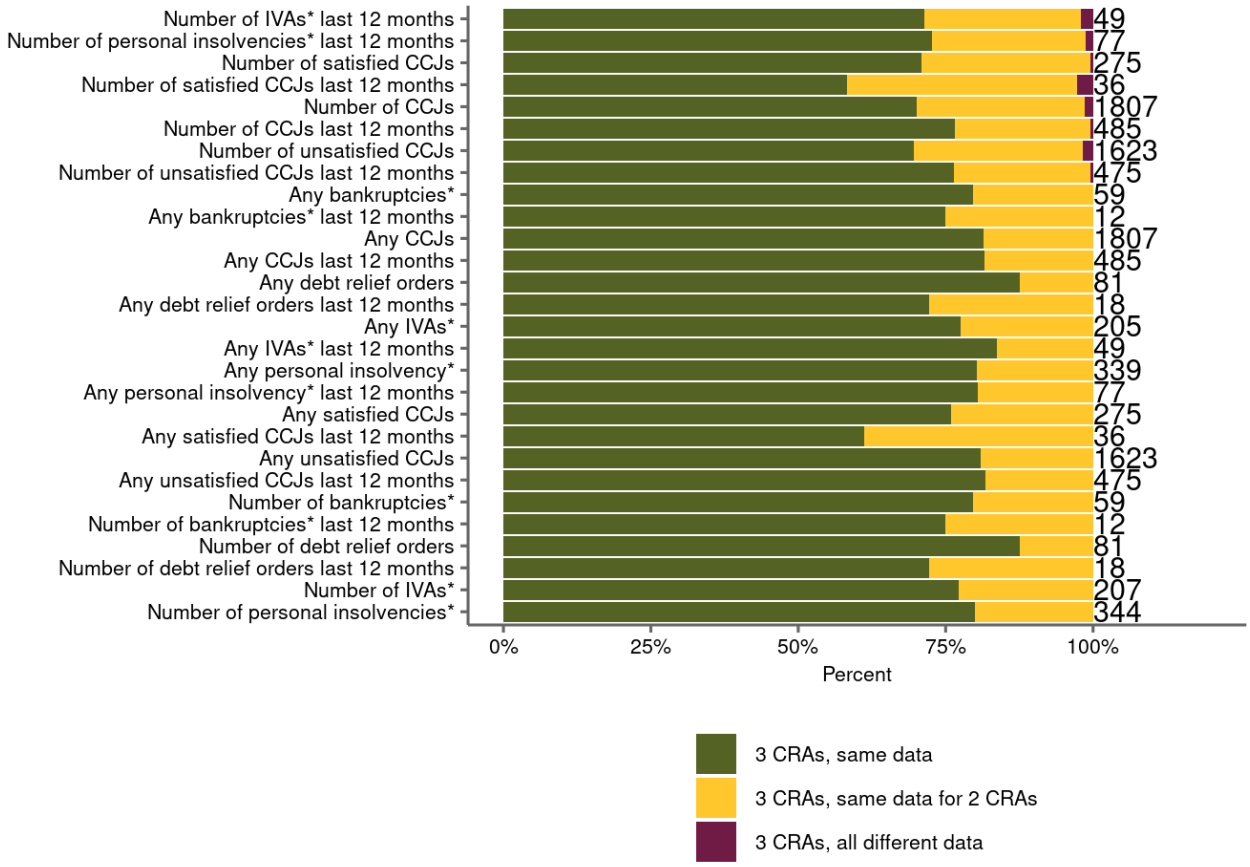


Source: FCA analysis on CRA data. *Comparison only includes information from 2 CRAs.

93. We also explored the extent to which public data differs for an individual by CRA. This is important as lenders told us that public information, such as CCJs and insolvencies, is one of the most important factors in lending decisions. Table 4 shows that public data is a significant consideration in lending decisions. Therefore, individuals could face harm if this data is not present at all CRAs.

94. Figure 15 below sets out that whilst for the majority of individuals CRAs are broadly in agreement on the number of CCJs an individual has, there are a significant number of instances of CRAs who record a different number of CCJs for an individual. Similarly, there are comparable levels of agreement between CRAs on bankruptcies, debt relief orders and IVAs.

Figure 15: Differences between the 3 large CRAs on a set of selected public data



Source: FCA analysis on CRA data. *Insolvencies include bankruptcies, sequestrations, protected trust deeds, debt relief orders and individual voluntary arrangements. IVAs include trust deeds and bankruptcies include sequestrations.

Data on utility and fixed-sum unsecured lending accounts is more different than other account types

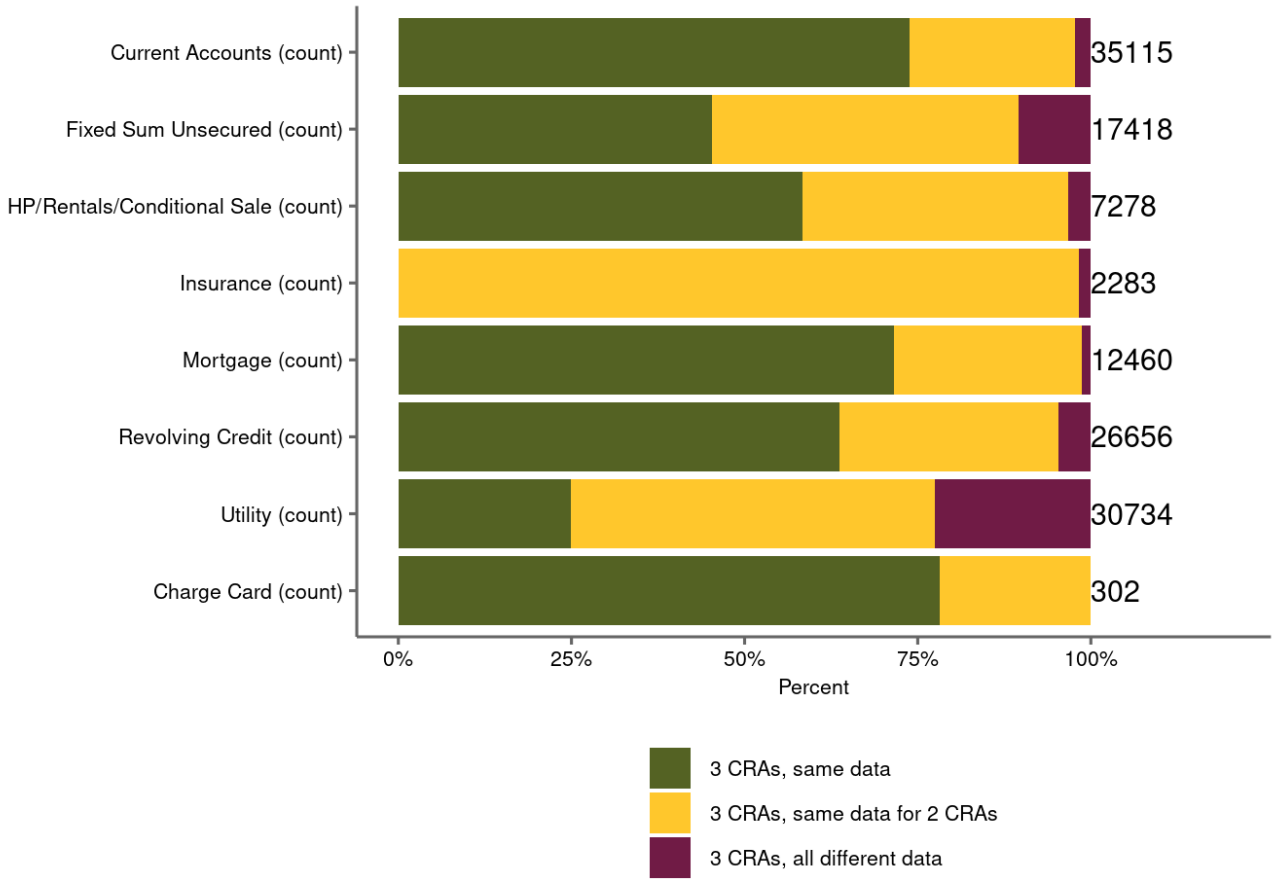
95. As we explored before, for many individuals, the number of accounts recorded varies by CRA. We wanted to explore to what extent this was driven by, for example, utilities firms reporting data to only one or 2 CRAs.

96. We undertook a mapping exercise to map product types as defined by each CRA to a unified product mapping. Such a mapping exercise is inherently problematic as product types may not map neatly to each other. As a result, we have undertaken 2 levels of mapping: layer 1, which is more granular, and layer 2, which is less granular.

97. As the layer 2 mapping is less granular, it is less likely to be affected by differences between CRAs of product types. For example, CRAs can have multiple different types of mortgages as product types, which may be difficult to map, but grouping these all together as 'mortgage' in layer 2 means that most products should be picked up.

98. Figure 16 below provide the results of the layer 2 mappings. Similarities in the number of accounts between CRAs is higher for current accounts and mortgages than utility products and fixed-sum unsecured lending. This is consistent with the latter types of products being less frequently shared to all 3 CRAs.

Figure 16: Differences between the 3 large CRAs on a set of selected financial products



Source: FCA analysis on CRA data

Income data varies greatly across the CRAs

99. Income data that is collected by the CRAs is an input into lenders' affordability assessments.

100. Affordability differs from credit risk as it is about how difficult it may be for a consumer to repay credit agreements in light of their wider financial situation. Affordability is therefore a 'borrower-focused' test whereas credit risk is seen as 'lender focused' test, however there is often overlap between the 2 measures. Discussion surrounding affordability is outside the scope of this work, but we have analysed differences in the income data held by the CRAs.

101. Differences between CRAs' income data can impact on lenders' affordability assessments of individuals. This can in turn impact the credit or terms of credit which a consumer is offered, and thus it is important to examine it.

102. For one CRA we received net income data and for the other we received gross income data. We also had data on national income statistics.

103. In order to make the income data comparable with the national statistics we therefore calculated individuals' gross annual income for the net income data using the marginal tax rates provided to us in the CRA's documentation.

Table 5: Gross annual income of individuals with CRA data and national statistics (all FCA IDs)

	CRA A		CRA B		National statistics
Metric	*Mean*	*Median*	*Mean*	*Median*	*Median*
Income	34,392	27,488	43,434	25,559	26,000

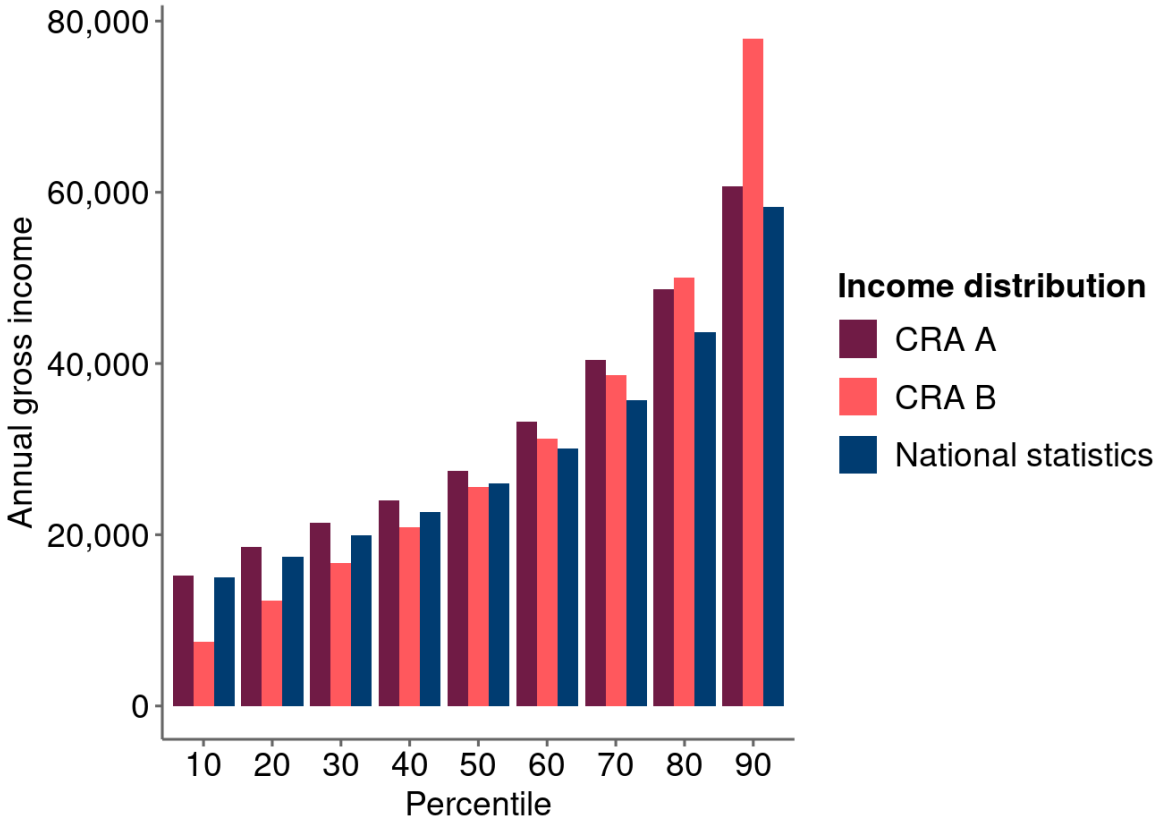
Source: FCA analysis on CRA and HMRC Survey of Personal Incomes data. Sample is limited to individuals uniquely matched across CRAs to allow comparisons.

104. When comparing median income estimates, we see a small but noticeable difference for individuals known to all 3 CRAs. This suggests that broadly, income data can be different across CRAs. We also see that the median income reported by each CRA is close but different to the national statistics.

105. CRA B's estimates for the mean income are much higher than the equivalent figures for CRA A. This is due to our calculation not considering tax implications of other income that higher earning individuals may receive.

106. Examining the differences in income data graphically can help explain this. Figure 17 below displays graphically, by percentile, income estimates for 2 CRAs and national estimates. Our results show that there are large income differences, especially for individuals at the extremes of the distributions.

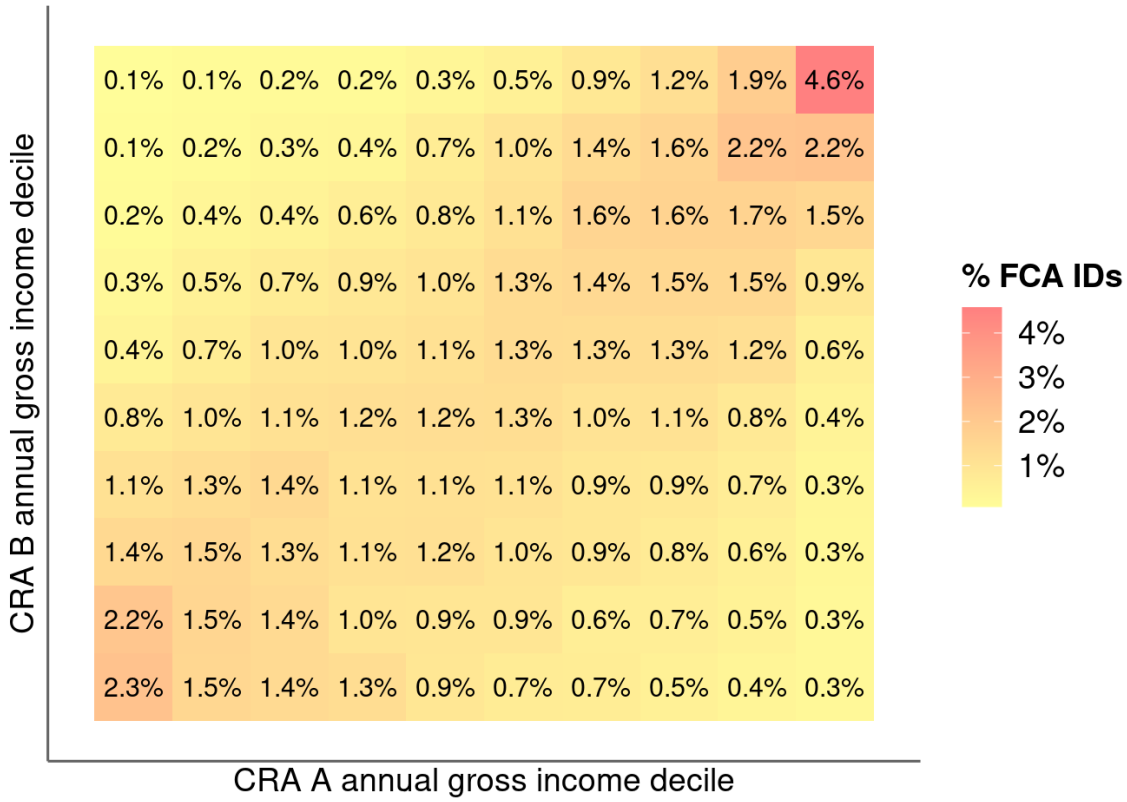
Figure 17: Inconsistencies in income estimates, comparisons with national statistics



Source: FCA analysis on CRA data

107. Figure 18 shows a heatmap which compares individuals' income data between CRAs. If we consider the same individual and compare their relative position within the distributions of the 2 CRAs, we find that over 54% of individuals have incomes that differ by 2 or more deciles.

Figure 18: Inconsistencies in income estimates



Source: FCA analysis on CRA data

108. The differences observed in income data between CRAs indicate that CRAs do not hold similar income data for individuals. This has impacts on the affordability assessments carried out by lenders who use this data. As discussed, this can have significant impacts on lending decisions and consumers can face harm as a result.

109. The potential for harm from these differences is somewhat constrained by the growth of Open Banking being used as an affordability insight by lenders. Open Banking can provide a more accurate and up to date reflection of a consumer's affordability. Although, this currently will only mitigate harm where lenders utilise Open Banking, the use of Open Banking may increase going forwards as the technology matures and uptake grows. The CRA Competition Annex discusses the impact of Open Banking on credit information, including take up and barrier to Open Banking adoption, in more detail.

A minority of data contributors do not share data with all 3 large CRAs

110. Differences in credit information may be due to some CIUs not sharing data with all 3 large CRAs. We therefore have examined the extent to which data contributors share data with each CRA.

111. We used information on which data contributors share data with each of the CRAs, and we estimated how many accounts held by individuals in our sample are provided by firms that share data with 1, 2 or 3 large CRAs.

112. Table 6 shows how many accounts are provided by the contributors that share data with 1, 2 or 3 large CRAs.

Table 6: Number of accounts provided by the contributors that share data with 1, 2 or 3 CRAs

CRA	Accounts provided by firms sharing data with all 3 large CRAs	Accounts provided by firms sharing data with any 2 large CRAs	Accounts provided by firms sharing data with 1 large CRA	Total
CRA A	*90%*	*7%*	*3%*	*100%*
CRA B	*92%*	*6%*	*2%*	*100%*
CRA C	*86%*	*8%*	*6%*	*100%*

Source: FCA analysis on CRA data

113. The table shows that between 86% and 92% of accounts held by individuals in our sample are provided by data contributors that share data with all 3 large CRAs. Between 6% and 8% of accounts are provided by firms that share data with any 2 large CRAs and between 2% and 6% are provided by firms sharing data with only 1 large CRA.

114. These results indicate that a minority of data contributors do not share data with all 3 CRAs, which is a source of data differences between CRAs. Moreover, this analysis does not cover the level to which data contributors share data with smaller CRAs, which can be assumed to be lower level than the above figures.

115. It is also worth noting that data contributors, albeit submitting to all 3 large CRAs, might share inaccurate information that is corrected only by one or 2 CRAs. This can also lead to inconsistencies between data held by each CRA.

116. Other sources of inconsistencies are due to difficulties in matching credit accounts to the correct individual file. Given the high number of different data contributors that report to the CRAs, some inaccuracies may inevitably occur. For example, a CRA may mistakenly add a new credit file belonging to an individual to the credit history of another individual, or fail to identify that the individual already exists in their data, and keep it separate. There are also likely to be inaccuracies arising from errors in the underlying data sent to the CRAs by lenders.

5 Impact of data differences on lending decisions

Introduction

117. A lender typically assesses a credit application using the information provided by CRAs. If the quality of information about the individual was poor, then the individuals might have different outcomes from credit applications or choose to make different credit applications where they have access to credit information pre-application.

118. This change in behaviour, from both individuals and lenders, could lead to many different types of negative outcomes. For example, if one CRA does not have comprehensive data on the applicant's credit history, then a lender relying on the information provided by that CRA may open a credit line while the same lender would have not chosen to do so if a different CRA was used. This results in harm if the consumer is unable to repay the debt. Alternatively, if a CIU has a lower score on an individual due to incorrect data then the individual would be at risk of losing access to credit. There can also be harms outside of specific individuals. For example, if a lender knows that CRA data may be inaccurate, the lender may choose to lend less, or to charge higher rates, as there is additional uncertainty in credit risk of individuals. Where data quality was perceived as unreliable, lenders may also face greater incentive to exit the market, or to choose not to enter.

119. Exploring and modelling all these possibilities is beyond the scope of this paper. This reflects the fact that using CRA data is not sufficient to calculate exactly how lenders make decisions. Instead, we use CRA data to consider the following.

120. Firstly, we explore to what extent lenders use multiple CRAs when assessing applications, which they may choose to do in the face of data differences between CRAs.

121. To explore how problems with data quality at CRAs can lead to negative outcomes, we examine how differences in credit information that we have identified in the previous chapters of the annex are linked to different lending outcomes. Because lending decisions can be highly complex and multi-dimensional, we explore how CRA scores are linked to lending outcomes. Exploring this allows us to examine what sort of individuals might be most likely to be affected by data problems.

122. We then undertake a more exploratory analysis to examine how lending decisions might have been different if a lender used an alternative CRA.

Lenders only use 1 CRA to assess most applications

123. We already know from qualitative information that each lender has different lending practices. As a consequence, the impact of data inconsistencies on lending decisions varies across lenders. Moreover, a lender may use 1 or multiple CRAs to determine whether to accept or reject an application. Thus, it is more likely that data inconsistencies have an impact on those decisions where only 1 CRA was used, while the impact is likely to be smaller when a lender combines data from more than 1 CRA.

Thus, we used search data to assess to what extent credit information users use multiple CRAs.

124. Search data is of 2 types: hard and soft. A lender may search an individual's credit report when he or she applies for credit, or when, for example, the individual is being chased for an outstanding debt. This search is called 'hard' and the CRAs keep these records on the individual's credit report for up to 2 years, so other credit information users can access them and base their lending decisions on the amount and type of applications the individual has made.

125. Soft searches, on the other hand, are not visible to other lenders, but are on an individual's credit information file held by CRAs. For example, an individual may shop around for credit and compare different deals before applying for a specific product. This search will not be visible to other lenders and consequently will not affect their decisions.

126. From our data request, we have search data from the 3 large CRAs. The datasets include all the queries done by credit information users about the individuals in our sample in the last 5 years. When we look at whether a lender used multiple CRAs, we limit our analysis to hard searches only. This is because any lending application will include a hard search, even ones which start with a soft search. Furthermore, lenders face different incentives in using multiple CRAs for a soft search, as the lender will not make a decision based on the soft search alone.

127. We mapped the different search categories at each of the 3 CRAs to either hard or soft search types. Moreover, in order to identify whether searches at one CRA were located at another, we mapped CRA identifiers for firms across each CRA for the most frequently used CIUs.

128. In calendar year 2018 around 55% of individuals in our sample made a credit application (26,479 out of 48,012, limiting the analysis to only individuals who have been matched across all 3 large CRAs, uniquely or otherwise).

129. It is useful to look at how many CRAs are typically used for a given application in order to gauge the extent to which lenders use multiple CRAs to assess applications. We therefore examined the number of CRAs used per application in a 2 year period.

130. Table 7 shows that for around 91% of applications, lenders used only 1 CRA. For 9% of applications CI users used 2 CRAs and for almost 0% of applications, credit information users used all 3 large CRAs. To note, we do not have data on the use of smaller, challenger CRAs for an application, however this is unlikely to materially impact our findings as the CRA market is heavily concentrated around the 3 large CRAs that we received data from.

Table 7: Number of CRAs used for an application undertaken between 31 Jul 2017 and 31 Jul 2019

Total # of applications	Applications sent to	Number of applications	% of applications
165,641	Any 1 CRA	150,038	91
	Any 2 CRAs	15,095	9
	All 3 large CRAs	508	~0

Source: FCA analysis on CRA data

131. Lenders may also correct for inconsistencies in other ways – such as seeking information from other sources (including directly from the consumer or by using more innovative new CRAs). Overall, our data suggests that the extent to which lenders mitigate for inconsistent information is limited, this is discussed briefly in the CRA Competition Annex of this report.

132. Lenders may also differ in the type of information requested. For example, a lender may obtain data on affordability from one CRA and data on credit risk from a different one.

133. The use of credit information varies also depending on the amount of information a CRA holds about an individual. For example, some lenders use a secondary CRA when the primary CRA does not hold data about a consumer or their credit file is thin. Others may ask applicants for additional information if information is missing. The Competition Annex discusses the reasons behind multi-homing in more detail.

Credit scores are strongly linked with lending outcomes

134. In this section we analyse the relationship between credit scores and success of application following a hard search. This is useful as it helps us to understand the potential impact of poor quality data on a consumer's access to credit.

135. The data that we received from the CRAs does not directly state what the outcome of a hard search was (ie did the applicant get an approval or rejection for the credit), however in some cases we can infer this. In particular, we can match data on searches to data on account openings. Where we observe a hard credit search made for, say, a credit card at a particular firm, we can then check to see if an account at this firm was opened by the individual in a reasonable window of time.

136. We cannot however observe whether the application was rejected and so this type of analysis is risky if done incorrectly. For example, if we considered soft searches, we may substantially underestimate the chance of a lender accepting a given application. Alternatively, if a customer makes an application then withdraws, for example if someone attempts to buy a house with a mortgage, but then this falls through, the there is a risk that the lender decision, which was not a rejection, could be considered as such. We have taken steps to address this by carefully considering search types, and product types. This can also be addressed by looking at products which we know from lenders automatically open accounts following a successful search. As a result we focus the analysis on products where an account is automatically opened after a successful search – credit cards, store cards and personal current accounts.

137. In this analysis we focused only on 1 CRA, for simplification. In this chapter we examine lending decisions, rather than cross-CRA comparisons, so using only one CRA in our analysis does not impact on the validity of the findings.

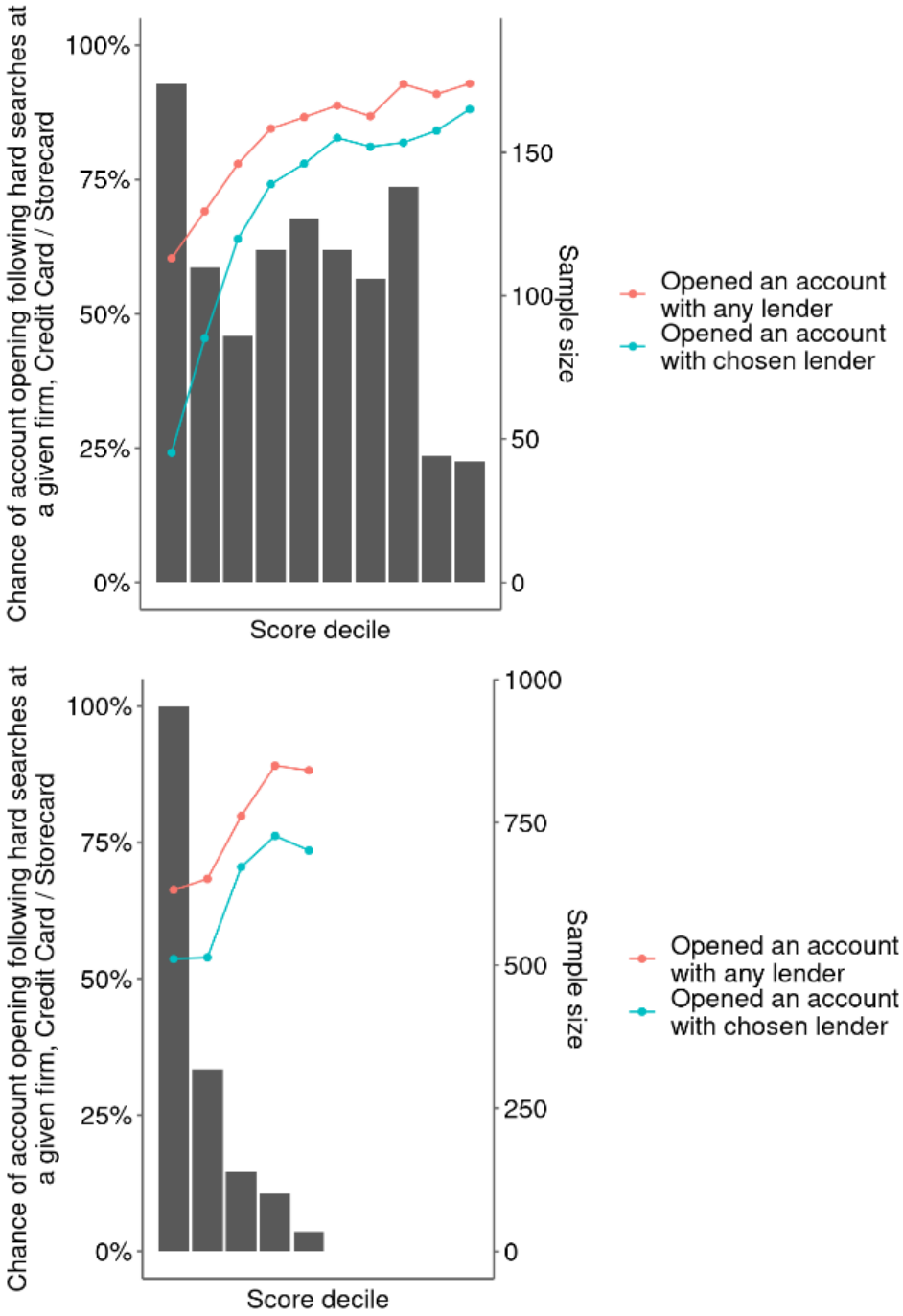
138. We can check our assumptions about whether account opening is a reasonable proxy for application acceptance by examining what happens to individuals with the highest credit scores. If many accounts were not opened for reasons other than rejections, then individuals with the highest credit scores would be likely to have lower levels of account opening than if accounts were only not opened following a rejection.

139. Figure 19 below provides an example for 2 providers, where acceptance for the highest scoring individuals is relatively high, supporting the idea that this methodology does reasonably identify the lending decision. We can see that for some individuals, those

reflected in the gap between the red and blue lines, being rejected from a product at one lender is followed with another lender accepting another application.

140. The distribution of scores is notably closer to the lower end for the second provider, indicating that most searches for the type of product at that provider are done by people with relatively low credit scores.

Figure 19: The probability of opening an account at 2 selected providers by a selected CRAs credit score



Source: FCA analysis on CRA data. Bars refer to the sample size, while lines refer to the chance of account opening. Results for the second graph for the top 5 deciles have been excluded because the sample sizes are very small. We considered all individuals in our sample who had a hard search for the selected products with the selected providers in a 5-year period between August 2014 and July 2019.

Even when a CRA only holds a thin file on an individual, many CIUs do not use additional CRAs for credit information

141. The proportion of applications from individuals with a thin credit history with at least one of the CRAs used in the application is between 2% and 3% for applications sent to 1, 2 or 3 CRAs.

142. Lenders can use a waterfall approach if the first CRA does not have enough information about the applicant, and use another CRA to get information. CRAs told us this is very common when the first CRA returns a thin or no file.

143. However, our data suggests that waterfalling is more limited: overall, out of the 4,185 applications done by individuals with thin credit histories with at least 1 of the CRAs in the application process, for 3,741 (89%) only 1 CRA was used and for 430 (10%) 2 CRAs were used.

CRAs have thick files on most applicants

144. In this section, we consider individuals in our sample who applied for a product and did not open an account. We also analyse how many of them were considered to have thin credit histories by the CRA used by the lender and how many would have had a thick credit history if the lender had used a different CRA. This is useful as it allows us to see the extent to which a different lending outcome may have occurred if a lender had used a different CRA to assess the application.

145. We considered the same products and lenders used for the analysis in the previous section. However, while the first analysis looked at those applications which led to an account opening, this analysis looks specifically at those applications that did not result in account opening.

Table 8: Impact of score inconsistencies on lending decisions: accounts not opened

Product	**Provider**	Total number of applications	Applications that did not lead to an account opening	Of these, applications with a thin file with the CRA used	Of these, applications with thick files with another CRA
Credit Card / Storecard	Firm A	1,222	413	1	1
Credit Card / Storecard	Firm B	2,194	1,016	8	4
Credit Card / Storecard	Firm C	675	294	0	0
Credit Card / Storecard	Firm D	1,393	587	9	2
Credit Card / Storecard	Firm E	5,866	2,768	6	4
Mail Order	Firm F	3,529	1,128	20	4
Credit Card / Storecard	Firm G	1,088	452	1	0
Current Accounts	Firm C	1,684	361	13	3
Current Accounts	Firm H	2,718	434	9	3

Source: FCA analysis on CRA data. In order to compare data to other CRAs, applications are limited to individuals uniquely matched across CRAs.

146. Table 8 shows that a small proportion of applications that did not lead to an account opening were submitted by individuals with thin credit history according to the CRA used by the lenders. However, often these also had thin credit history according to another CRA.

CRA data is not sufficient for understanding *how* a lender made a given lending decision

147. Lenders may make their lending decision using a CRA's credit score, non-score data from the CRA, or data outside of the CRA. Given the scale of the number of products and firms, and the likelihood that lending decision methods vary between each of these, we have not attempted to model how a given lender makes decisions.

148. If we knew exactly how a lender makes decisions, we could use this to consider the effect of data differences on outcomes. For example, if we knew that a lender accepted all applications where the individual is in the top 20% of the population for score, or alternatively if the individuals had no bankruptcies, then we could identify individuals for whom the lending decision would have been different if an alternative CRA was

used. However, we do not know these models in general, which may be highly complex, and the extent to which the lenders use data in their modelling that is outside of that supplied by CRAs.

149. In our conversations with lenders, some lenders told us that few lenders use the score calculated by the CRA, while others use the data and their own credit scoring model. Some lenders combine data from several CRAs and calculate the score using their own credit scoring model (ie 'multi-bureau' scorecards).

150. This implies that, when assessing the impact of data inconsistencies on lending decisions, we need to take into account the high degree of heterogeneity in how these decisions are formed. It is possible that data inconsistencies have a high impact if the same consumer makes a credit application to a specific lender, and little or no impact if they make the application to another lender, if that lender does not use data from CRAs as a main input in its lending processes. However, the likelihood of a lender not using CRA data as a least somewhat of an input to lending decisioning is extremely low. Thus, it is likely that data inconsistencies (if present) will have an impact, at least somewhat, on an individual lending decision.

Individuals who have very low scores at another CRA may be rejected if the lender had used that CRA

151. To circumvent the issue of identifying lending decision models, we assessed the number of individuals in our sample that opened a credit account whose application would have been likely to be rejected had the lender used a different CRA.

152. While we cannot reasonably estimate individual lending models, we can explore the implications of the following assumptions:

153. Individuals who were accepted by a lender using one CRA would have been unlikely to have been accepted if the lender used an alternative CRA where the individual has a very low score at that CRA. This is because the very low score likely reflects material negative information on the account, such as a default, or missing positive information, such as current account which is interpreted as positive by the CRA.

154. Conversely, individuals who were rejected by a lender at one CRA may have been accepted if the lender used an alternative CRA where the individual has a very high score at that CRA. This is because the very high score likely reflects either missing negative information, or the presence of an account which was not recorded at the other CRA.

155. That is, while we do not know how lenders make individual decisions, factors which cause a lender to take a very negative/very positive view of an individual, and factors which cause a CRA to take a very negative/very positive view of an individual likely overlap at the extreme ends of the spectrum.

156. In the next few figures, we examine the distribution of scores of individuals who applied for and opened a specific credit product and the distribution of scores of individuals who applied for but did not open a specific credit product. Examining these distributions allows us to identify score thresholds below which very few individuals were accepted.

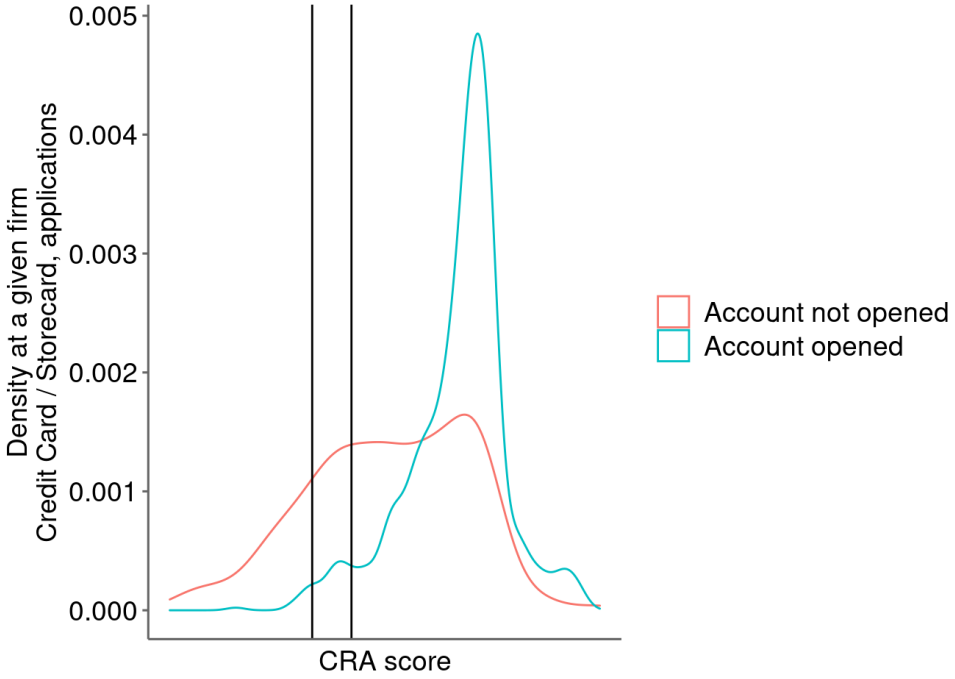
157. The blue curve in Figure 20 represents the distribution of scores of individuals who applied and opened a credit card account with a selected provider. The red curve

represents the distribution of scores of individuals who applied for the credit card but did not open it.

158. There may be several measurement errors that explain why we observe a low score for successful applications. First, we observe credit scores only every 6 months. As a result, the score of an applicant may decrease after an application and before we observe the score. Second, lending criteria may change over time.

159. As expected, the average score of successful applicants is above the average score of applicants who did not open an account. However, the 2 distributions overlap (eg a high score individual may still be rejected) because the score may be only one of the criteria used by this lender to accept an application. This is common and confirmed by qualitative information we obtained from credit information users. For example, other common criteria for accepting or rejecting an application may include the lender's appetite for certain consumer segments or availability of funds.

Figure 20: Distribution of scores of individuals who applied for a credit card with a selected provider



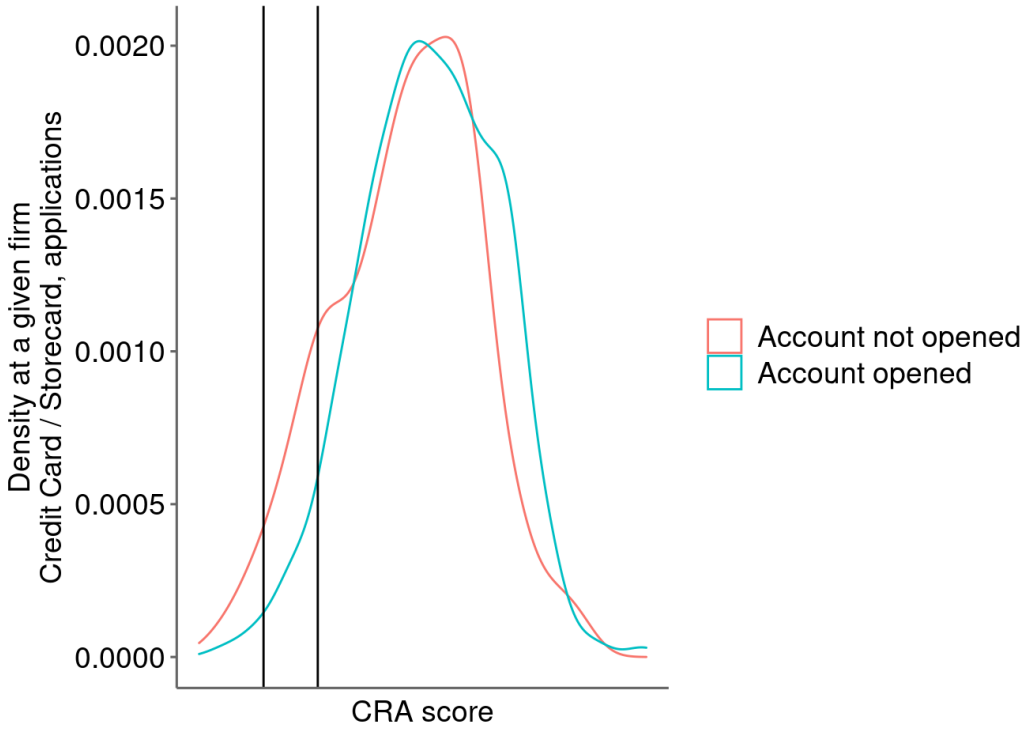
Source: FCA analysis on CRA data. The red curve shows the distribution of scores of individuals who opened a credit account with a selected provider and the blue curve represents the distribution of scores of individuals who applied for the credit card but did not open it. The black lines represent the 1st and 5th percentile of scores where we observe an account being opened following the application. In order to compare data to other CRAs, applications are limited to individuals uniquely matched across CRAs.

160. We identified the acceptance threshold by using both qualitative information provided by lenders and the data available. Figure 20 shows the 1st percentile and the 5th percentile thresholds.

161. We repeated the same exercise with a different lender in Figure 21. As before, the blue curve represents the distribution of scores of individuals who applied and opened a credit card account with a selected provider and the red curve represents the distribution of scores of individuals who applied for the credit card but did not open it. There is a significant overlap between the score of applicants who opened an account and other applicants. This analysis is less indicative of the conclusion above, and

instead suggests that the lending assessment is less similar to the scoring methodology of the CRA.

Figure 21: Distribution of scores of individuals who applied for a credit card with a selected provider



Source: FCA analysis on CRA data. The red curve shows the distribution of scores of individuals who opened a credit account with a selected provider and the blue curve represents the distribution of scores of individuals who applied for the credit card but did not open it. The black lines represent the 1st and 5th percentile of scores where we observe an account being opened following the application. In order to compare data to other CRAs, applications are limited to individuals uniquely matched across CRAs.

Modelling counterfactual lending decisions by looking at best and worst scores in lending pools

162. In this section we model counterfactual lending decisions by looking at the best and worst scores in the lending pools. This extends the analysis of score distributions and lending decisions undertaken in the previous section. The previous analysis would likely produce an underestimate of the number of affected individuals. As we can see from Figures 20 and 21 above many individuals with reasonably high scores appear to get rejected, for example because the lender cares about data different than to how the score is calculated, or because the specific details of the product were unaffordable.

163. As a result, we consider the case where a lender may reject a credit application if i) the applicant has a low score (perhaps because the CRA used considers them to have a thin credit history) or ii) the CRA used has no information about the applicant. These may not be the only reasons why an application is rejected, but we consider these plausible scenarios where data inconsistencies may potentially affect lending decisions.

164. For this analysis we have focused on credit cards, store cards and personal current accounts.

165. The providers included in our sample serve different consumer segments and are of different sizes.

166. We identified the score threshold as the 1st percentile of the score distribution of the applications who subsequently opened an account. We also did a sensitivity check using the 5th percentile. The thresholds (henceforth 'lending thresholds') are consistent with the qualitative information about lending criteria of certain lenders we have available.

167. It is important to note that this exercise is hypothetical because we cannot assess what would have happened if a lender had used a different CRA. In fact, different CRAs may estimate credit risk differently and lenders may adjust their lending decisions criteria if they change CRAs. We also acknowledge that scores are generally not solely used when lenders make an assessment and that lending decisions are highly complex. This exercise provides indicative results under the assumption that scores are a proxy for all the information that a CRA holds on a given individual.

168. If data and scores are inconsistent it is possible that lending decisions are affected. A lender may approve one credit application based on the data from one CRA while the same lender would reject the same application if it had used a different CRA.

169. Table 9 shows the proportion of individuals who opened a credit account and i) have a score below the lending threshold according to another CRA (and have a score above the lending threshold for the actual CRA) or ii) are not known by another CRA or iii) have a thin file according to another CRA.

Table 9: Impact of score inconsistencies on lending decisions: accounts opened

Product	Provider	have a score below the lending threshold according to another CRA		are thin files for another CRA
		Threshold = 1st percentile	Threshold = 5th percentile	
Credit Card / Storecard	Firm A	0.0%	0.1%	0.1%
Credit Card / Storecard	Firm B	0.7%	1.8%	0.0%
Credit Card / Storecard	Firm C	0.0%	0.3%	0.0%
Credit Card / Storecard	Firm D	0.0%	0.4%	0.2%
Credit Card / Storecard	Firm E	0.5%	1.6%	0.0%
Mail Order	Firm F	0.3%	2.1%	4.2%
Credit Card / Storecard	Firm G	0.2%	1.5%	0.0%
Current Account	Firm C	0.1%	1.2%	3.3%
Current Account	Firm H	0.1%	1.5%	0.8%

Source: FCA analysis on CRA data. In order to compare data to other CRAs, applications are limited to individuals uniquely matched across CRAs.

170. If we assume that a score below the threshold would automatically lead to a rejected decision, we can say that inconsistencies would have led to around 0% to 2% of accepted applications being rejected by each lender considering the 5th percentile threshold. Whilst this assumption is a simplification of how lending decisions are made in practice, it suggests the scope for harm for most consumers is limited as lending decisions do not appear to be significantly impacted by inconsistencies in credit information.

171. Overall, we found that the differences in scores, using this methodology, do not imply that differences in scores have a very large impact on lending decisions, however this is likely an underestimate, as we consider only individuals who had very bad credit scores (among those accepted) at other CRAs. As we show in Figures 20 and 21 there seem to be many individuals who are rejected despite having credit scores above the lowest percentiles of accepted individuals.

Alternative approach to modelling counterfactual lending decisions

172. The analysis that we set out above provides a reasonable lower bound on the number of individuals who may face different lending decisions if an alternative CRA was used.

Using a different approach to modelling counterfactual lending decisions can also help us to estimate the impact of poor quality data.

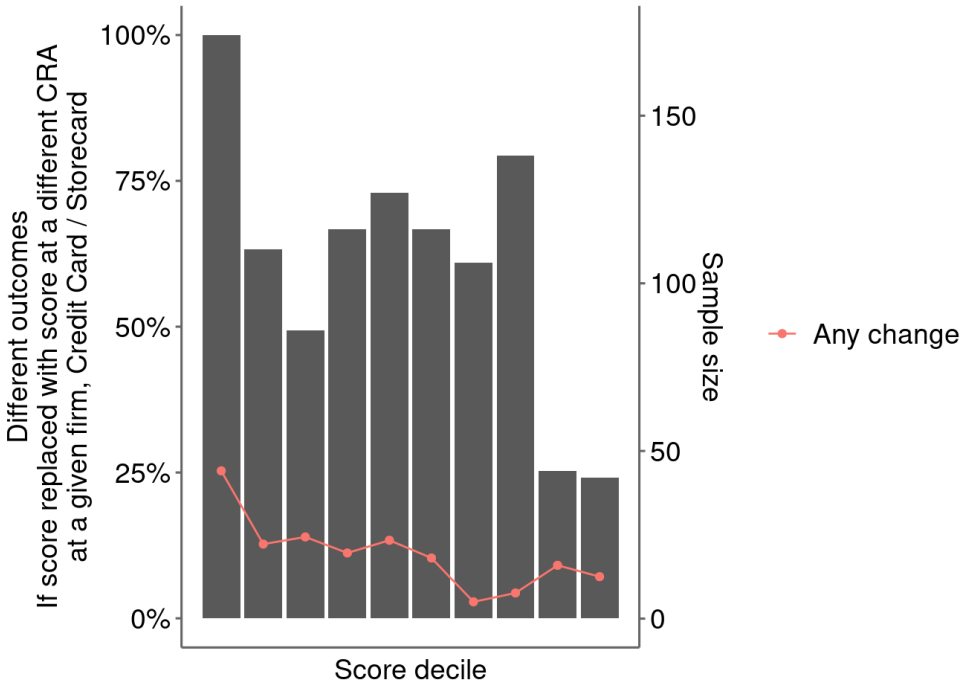
173. Using CRA score as a proxy for the information available to the lender, we can bucket individuals in groups and give them implicit chances of acceptance.

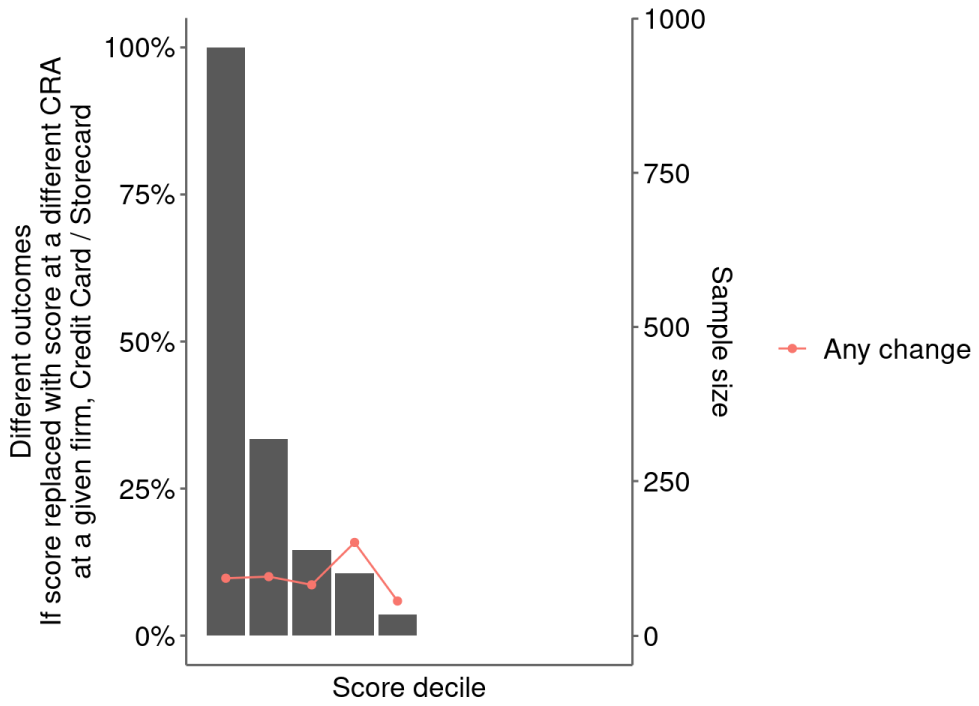
174. We can construct counterfactual probabilities with the scores from other CRAs, again using these as a proxy for the underlying data available to the CRA, rather than assuming the lender uses CRA scores directly. For example, an individual may be in a score decile with an average acceptance of 65% at the CRA the lender is using, but at an alternative CRA (at the corresponding score) be in a score decile with an average acceptance of 45%.

175. We can then undertake a simulation for individuals, assigning each applicant a random number, then marking them as accepted if that number is less than the average application percentage of the bucket.

176. We can then compare the 2 simulated outcomes. If we compared one CRA against itself, there would be no difference, but if we compare against another CRA, each individual who relatively moves has a chance of being accepted where they were rejected, and vice versa. Figure 22 shows the results of running this simulation for 2 lenders.

Figure 22: Simulated lending decisions for 2 selected providers





Source: FCA analysis on CRA data. Bars refer to the sample size. Results for the second graph for the top 5 deciles have been excluded because the sample sizes are very small.

177. The pattern of middle and lower scoring individuals being most likely to be affected reflects that in most cases, middle and lower scoring individuals are likely to be closer to the decision boundary for lenders.

178. This approach is an overestimate of the impact of using an alternative CRA. Relative scores between CRAs may vary because of the scoring methodologies they use as well as differences in data. However, the results here indicate some important points, in particular, as lenders can have complex models to assess applications, the impact of changes to CRAs do not just affect individuals on some cut-off by score.

Technical Appendix 1: Sampling and matching methodology

Introduction

179. In this technical appendix we present the methodology we used to sample, clean our data, and match individuals across the 3 large CRA. We also show the characteristics of our sample.

180. We first describe our methodology to sample the individuals from the CRAs' datasets. We then present methodology used to calculate the population coverage of the CRAs discussed in this annex. We also discuss how we cleaned the data we received. Following this, we describe how we matched individuals across CRAs. Finally, we show the characteristics of our sample.

Sampling methodology

181. To obtain a representative sample of the UK population, we requested the credit information from the 3 large CRAs of all individuals born on a particular date in each of the following years: 1920, 1923, 1926, 1929, 1932, 1935, 1938, 1941, 1944, 1947, 1950, 1953, 1956, 1959, 1962, 1965, 1968, 1971, 1974, 1977, 1980, 1983, 1986, 1989, 1992, 1995, 1998, 2001.

182. We also requested the credit information of financial associates of the individuals covered above (we distinguish hereafter between 'main subjects' and 'financial associates'). The sample was extracted on 1 August 2019 by all CRAs.

183. As we did not request information on individuals based on personal identifiers, but instead asked for information on all individuals born on a particular date, there are a number of challenges to our approach. In particular, the approach does not reflect how lenders and CRAs typically identify individuals and return relevant information, using key identifiers such as names and addresses. This means that it is possible that the data extracts from the CRAs do not include all relevant information on all individuals known to that CRA with that birth date.

184. In previous research we undertook (Occasional Paper No. 28: Preventing financial distress by predicting unaffordable consumer credit agreements: An applied framework[7]) we adopted a different approach. Here, we sent details of specific accounts to CRAs and asked for details on these individuals, which was used as the basis of comparing information on individuals across CRAs. A result of this approach was that there was no need to match identities between CRAs, as it was implicit in the sampling process.

185. The approach we have adopted in the current work (directly sampling and then matching individuals ourselves) is better suited for examining the issues we were concerned about. For example, our findings in coverage, that CRAs hold information on more implied individuals than there are people living here would be difficult to arrive at without doing the matching ourselves. In addition, we have also been able to identify and understand the challenge of 'multi-matching' where CRAs may reasonably not identify 2 records with the same personal information as the same person. Where 2 CRAs hold different records on an individual in our analysis, we are more easily able

[7] https://www.fca.org.uk/publications/occasional-papers/occasional-paper-no-28-preventing-financial-distress-predicting

to address the possibility that this is because the data held truly is different, rather than the CRA has the records, but did not return them because the individual was not matched.

Coverage

Estimating expected population

186. We estimated the proportion of the population covered by credit information by calculating the:

- Numerator as the number of individuals in our sample (excluding the few who are not in the regions covered by the ONS).
- Denominator as the number of individuals we might *expect* to have in our sample given ONS population estimates (expected sample size).

187. To calculate the denominator, we used the 2019 Office for National Statistics (ONS) mid-year population estimates, the closest date to when our sample was collected. This estimate incorporates common fluctuations in population attributed to migration, births and mortality and is the best, readily available, estimate of the UK adult population. We only considered the population born in the years we selected. We then calculated the probability of being born on the selected day using ONS data. The denominator is given by multiplying the UK population estimate in the selected years by the probability of being born in the selected calendar year. This is the number of individuals we expected to appear in our sample.

188. We used ONS population estimates to compare our sample with the UK population. ONS population estimates are broken down by geographical areas (Lower-layer Super Output Area (LSOA)) and age.

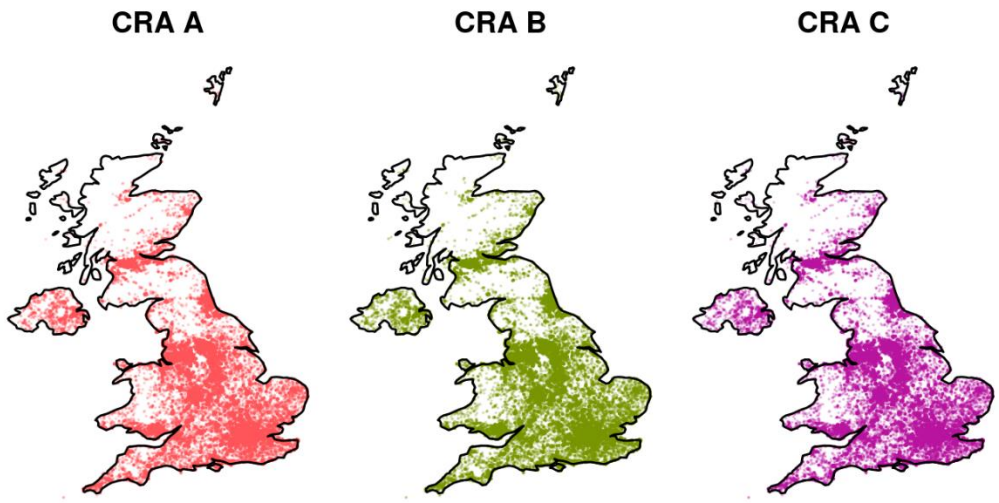
189. Given that our sample of individuals is drawn based on a date of birth, we do not have information about individuals for which CRAs do not hold a date of birth record. We consider that if CRAs do not hold date of birth information about an individual, it is also likely that this individual has a thin credit history (eg the CRA may be aware of this individual only based on public records, such as the electoral roll that does not include information about the date of birth). It is then possible that we are underestimating the proportion of thin file individuals in our sample. However, we do not consider this to have a significant impact on our result.

190. Our main findings on population coverage are detailed in Chapter 2 of this annex.

Checking geographic coverage

191. Figure 23 shows the postcode distribution of the IDs of the main subjects across the UK received by each CRA. The figure shows that individuals are concentrated around the most populated areas of the country, as expected.

Figure 23: Geographic distribution of main subjects from each CRA



Source: FCA analysis on CRA data

Addressing potential issues in the data

192. In some cases, the underlying data of an ID is not unique, and so a process needs to be undertaken to make data comparable between CRAs. For example, a CRA may have multiple versions of an account on record or record an account twice under slightly different details. Simply adding these accounts up could lead to a conclusion that underlying data was different, where in fact it was only the recording practices between CRAs that was different. As a result, we have undertaken steps to address this risk, and to only include what we interpret to be unique accounts, unique CATO data, unique searches, and unique public data, such as CCJs.

Matching methodology

193. To identify a consumer, CRAs use various combination of personal identifying information. However, when a data contributor sends data to a CRA, they might not send all of the identifying information (eg middle name or date of birth might be missing). To analyse differences in credit information between CRAs, we need to be able to compare the same individual across the 3 large CRAs using the available identifying information.

194. When matching IDs across CRAs we aim to identify the same individual held by each CRA. The 2 risks are that we can either have i) false positives when we match IDs that do not belong to the same individual or ii) false negatives when we do not match IDs that belong to the same individual. The more stringent the matching methodology, the

smaller the risk of having false positives and the larger the risk of having false negatives (and vice versa).

195. We used 2 matching methodologies to match main subjects across CRAs. First, we matched individuals based on first name, last name, postcode and date of birth (stringent matching). Second, we matched individuals based on postcode and date of birth only (loose matching) as a robustness check. We discuss the 2 matching methodologies in the next section.

196. For each ID of each CRA we created a set of strings that includes first name, last name, full date of birth, and full postcode. Using all variations of name, date of birth and postcode, we create all possible combinations of these variables.

197. For example, if for a given ID a CRA provided 2 versions of the first name, 3 versions of the last name, 1 date of birth and 4 different postcodes, we created 2x3x1x4 = 24 combinations for this ID.

198. Each ID has at least one combination of first name, last name, date of birth and postcode. This therefore means we may create some combinations that were not actually present in the original dataset. We then looked for the same strings across the 3 large CRAs.

Number of matched individuals and robustness check

199. The outcome of the matching described above was a set of matched IDs across the 3 large CRAs (these are either unique matches or multiple matches), a set of matched IDs across any 2 CRAs (either unique or multiple matches) and a set of unmatched IDs. We report this outcome for both the stringent and loose matching in tables below.

200. We assigned a unique FCA ID to the IDs that we matched across the CRAs using our matching criteria. As a result, a non-unique match has multiple IDs of a given CRA associated to the same FCA ID. For example, a CRA may fail to link a new credit file related to an individual to the ID belonging to him or her (eg because the CRA has a lag between data coming in and being matched). Therefore, the CRA has 2 different IDs which, using our methodology, are both matched with the same ID of the other 2 CRAs. Thus, we create a unique FCA ID which identifies the same individual, but this ID will be matched twice to the CRA that did not consolidate the new record.

201. We consider 2 IDs are uniquely matched if a string of an ID matches to a string belonging to one ID in another CRA's dataset. An ID is not uniquely matched if strings of a given ID match to either:

- strings of more than one ID in another CRA's dataset or
- strings of a different ID in the same CRA dataset

202. We flagged these instances and reported them separately.

203. Overall, using the more stringent criteria we created 48,012 FCA IDs which matched across the 3 large CRAs. Of these, 39,809 were unique matches. We created a further 14,143 FCA IDs across any 2 CRAs and, of these, 13,385 were unique matches. 38,866 IDs were not matched. Table 10 shows the results.

Table 10: Number of FCA IDs matched with the stringent matching

Stringent matching	CRA A IDs	CRA B IDs	CRA C IDs	FCA IDs
Unique matches across 3 large CRAs	39,809	39,809	39,809	39,809
Non-unique matches across 3 large CRAs	8,203	8,203	8,203	8,203
Unique matches across any 2 CRAs	9,704	11,021	6,045	13,385
Non-unique matches across any 2 CRAs	715	498	303	758
Unique matches within 1 CRA	17,889	13,336	7,470	38,695
Non-unique matches within 1 CRA	111	58	2	171
Total	76,431	72,925	61,832	101,021

Source: FCA analysis on CRA data

204. Each column shows the number of IDs matched in each CRA, with the last column displaying the number of FCA IDs that we have generated. The number of unique matches across the 3 large CRAs (in first row) is the same, which means we found 39,809 IDs uniquely identified by all 3 large CRAs.

205. As a robustness check we replicated the methodology described above using a restricted set of criteria (ie postcode and date of birth only). Using looser matching criteria should increase the likelihood to match IDs across CRAs. This may increase the number of false positive and the number of non-unique matched (eg by matching individuals born on the same day and living in the same postcode and having different names).

206. We assessed whether the lose matching i) increases the total number of matches across 3 large CRAs and ii) the number of unique matches. We found that the total number of matches across 3 large CRAs does not change materially, while the number of unique matches decreases. This suggests that the loose matching gives are a higher number of false positive.

207. Table 11 shows that, using the less stringent criteria, we created 47,856 FCA IDs which matched across the 3 large CRAs. Of these, 35,823 were unique matches. We created 12,343 FCA IDs across any 2 CRAs and, of these, 11,191 were unique matches. 31,940 IDs were not matched. The less stringent methodology increases the number of non-unique matches without increasing the number of matches.

Table 11: Number of FCA IDs matched with the loose matching

Loose matching	CRA A IDs	CRA B IDs	CRA C IDs	FCA IDs
Unique matches across 3 large CRAs	35,823	35,823	35,823	35,823
Non-unique matches across 3 large CRAs	12,033	12,033	12,033	12,033
Unique matches across any 2 CRAs	8,229	9,102	5,051	11,191

Loose matching	CRA A IDs	CRA B IDs	CRA C IDs	FCA IDs
Non-unique matches across any 2 CRAs	956	806	542	1,152
Unique matches within 1 CRA	15,362	10,610	5,414	31,386
Non-unique matches within 1 CRA	352	153	49	554
Total	72,755	68,527	58,912	92,139

Source: FCA analysis on CRA data

208. Given that the number of matches across 3 large CRAs is similar (48,012 vs. 47,856) but the more stringent matching allows to identify a larger number of unique matches, we used the stringent matching for the analysis.

Deceased and gone away individuals

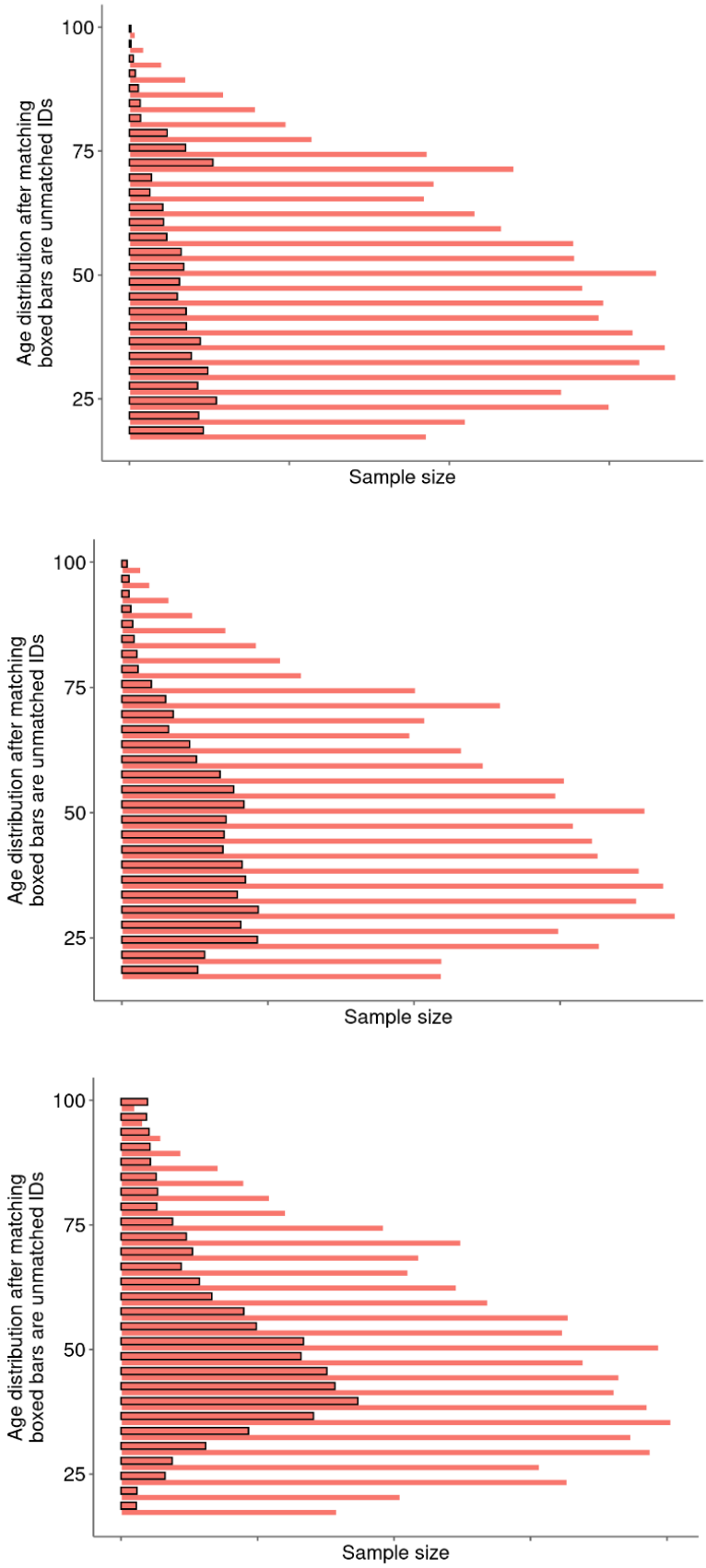
209. CRAs flag whether an individual is deceased or gone away. This is a based on a judgment made by each CRA, so CRAs may have differences in this information on a given individual between them.

210. To determine how to treat individuals who were flagged as deceased or gone away at a CRA we used a majority rule. If an individual is known to all 3 large CRAs, we removed individuals who at least 2 CRAs consider deceased or gone away. If an individual is known to only 2 CRAs, we removed the individual if 1 of the CRAs believes that they are deceased or if they are gone away. If an individual is known only to 1 large CRA, we rely on the only deceased and gone away flags available.

Unmatched individuals

211. Figure 24 compares the age distribution of the matched individuals (any type of matching) with the age distribution of the unmatched individuals. The number of unmatched individuals is represented by the dark-shaded bars. Naturally, given that we received a different number of IDs from each CRA, we also have a different number of unmatched in each CRAs dataset.

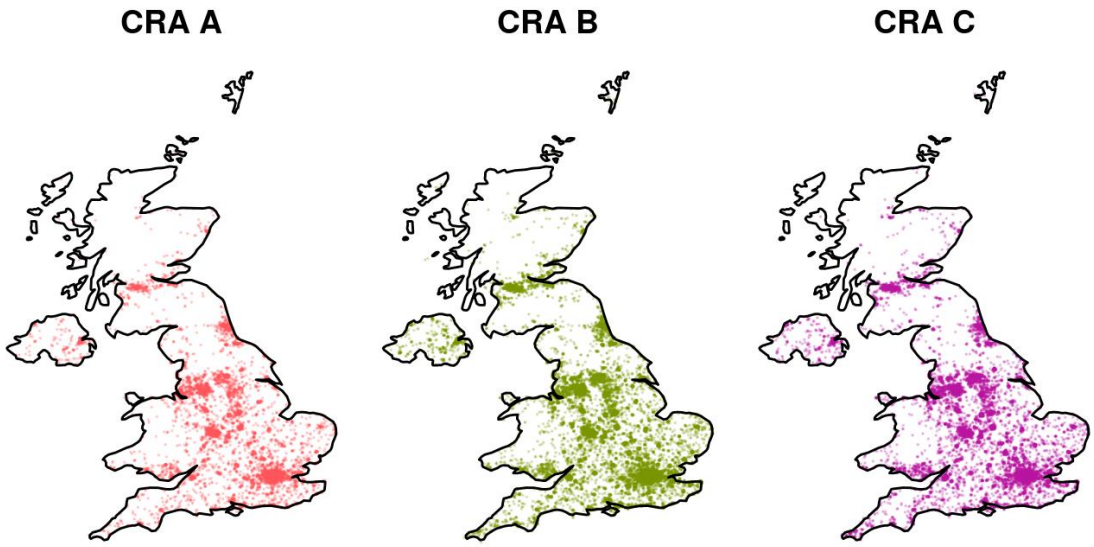
Figure 24: Age distribution of unmatched vs. matched IDs for each CRA



Source: FCA analysis on CRA data

212. Figure 25 shows the geographic distribution of the individuals that we did not match. Unmatched individuals appear concentrated around the most populated areas of the country as they do in Figure 23. This means that we are not systematically excluding individuals who live in certain locations.

Figure 25: Geographic distribution of unmatched vs. matched IDs for each CRA



Source: FCA analysis on CRA data

213. As the age and geographic distributions of unmatched IDs are similar to the distributions of the matched IDs, we can rule out the possibility that we are systematically excluding some categories of individuals because of matching errors.

Comparing information on individuals

214. To assess whether population coverage is correlated with borrower characteristics we used information about credit score, age, and age of accounts to assess whether individuals with lower income, who are younger or who have newer accounts are less likely to be captured in all 3 CRAs data.

215. We compared credit scores across the CRAs using the number of CRAs an individual is known to. This analysis explores whether individuals who are known by fewer CRAs have a lower credit score on average.

216. We found that individuals known to a subset of CRAs have lower average, and median credit score compared to individuals who are in all 3 datasets. Table 12 reflects individuals with both thick and thin files. Our findings do not change materially if we consider only individuals with thick credit histories.

Table 12: Credit scores, in percentiles, by CRA count

Individuals known to	CRA A	CRA B	CRA C
All 3 large CRAs	68	65	58
Any 2 CRAs	39	37	31
Any 1 CRA	25	26	30

Source: FCA analysis on CRA data. Deceased and Gone-away IDs removed according to our rule. For example, the average person at CRA A who is matched across all 3 large CRAs has a credit score in the 68th percentile for CRA A

217. Table 13 shows that individuals known to fewer CRAs seem to be slightly younger than individuals known to all 3 large CRAs. The findings do not materially change if we consider only individuals with thick credit histories.

Table 13: Age of IDs by CRA count

Individuals known to	All IDs	
	Mean	Median
All 3 large CRAs	49	48
Any 2 CRAs	46	42
Any 1 CRA	48	45

Source: FCA analysis on CRA data. Deceased and Gone-away IDs removed according to our rule

218. Table 14 shows that individuals known to a subset of CRAs have a lower average and median age of the oldest account. There may be several reasons. For example, individuals may be new to the credit market because they are young (eg they just turned 18) or they just arrived in the UK.

Table 14: Age (in years) of the oldest account by CRA count

Individuals known to	CRA A		CRA B		CRA C	
	Mean	Median	Mean	Median	Mean	Median
All 3 large CRAs	16	14	16	14	16	14
Any 2 CRAs	9	6	6	1	8	3
Any 1 CRA	1	0	1	0	8	3

Source: FCA analysis on CRA data. Deceased and Gone-away IDs removed according to our rule