Chapter 1

Data Accuracy and Fair Lending Analyses

Jason Dietrich
March 2026

The views and opinions expressed in this paper belong solely to me and do not represent the views or opinions of any employer, institution, or organization with which I have been affiliated.

This paper is for informational and educational purposes only and is not intended to serve as professional or legal advice. I specifically disclaim all responsibility for any liability, loss, or risk, personal or otherwise, which is incurred as a direct or indirect consequence of the use or application of the contents of this paper. Every effort has been made to ensure that the information in this paper is correct. However, I assume no responsibility for errors, inaccuracies, or omissions. The use of this paper implies the reader's acceptance of this disclaimer.

I. Introduction

A data accuracy analysis is a key component of any fair lending review. The Interagency Fair Lending Examination Procedures (IFLEP)[1] conveys this point by listing verifying the accuracy of the data as the first step of the examination process. Data accuracy is so important because of the general view that data errors undermine the reliability of statistical analyses. The well-known phrase, “garbage-in / garbage-out" reflects this perspective. As with most aspects of data analyses, however, data accuracy issues, and their impacts on statistical analyses, are much more nuanced. For example, as we show in section II below, a dataset with significant data errors can still yield reliable statistical results that reflect the true underlying relationships of interest. Therefore, when conducting the data accuracy analysis, it is important to consider issues like, what specific variables have data issues, what are the types of data issues, how extensive are the data issues, are the data issues correlated with key variables for the given analysis, how might the given data issues impact the statistical results, and more. Incorporating these considerations into the data accuracy analysis will indicate how data errors impact the statistical analysis and result in more reliable and accurate statistical conclusions overall. In this report we explore these nuances of data accuracy analyses in a fair lending context.[2]

There is currently a large volume of available resources and guidance on the topic of data accuracy.[3] To a large extent, these resources focus on how to reduce the likelihood of errors occurring, how to identify data errors, and how to mitigate the effects of data errors during statistical analyses. Relatively few of these resources focus on assessing the impacts data errors

[1] See page 17 of Interagency Fair Lending Examination Procedures.
[2] Throughout this report we focus narrowly on just data accuracy and not the much broader concept of data integrity.
[3] Examples of available sources include, McGilvray (2021), Maydanchik (2012), Olson (2003), Moses et.al. (2022), Walsh (2021), Blokdyk (2021), Kumar (2023), “The DAMA Guide to the Data Management Body of Knowledge: (DAMA-DMBOK Guide),” Data Accuracy in 2025: Ensure Reliable, Quality Data, and Guide to Data Quality: Ensuring Accuracy and Consistency in Your Organization | WhereScape.

have on statistical analyses, specifically within a fair lending context. Because of this, we focus much of this report on the potential impacts that data errors can have on statistical analyses. Given the overall importance of data accuracy to statistical analyses in general, we do also include some discussions of strategies for reducing the likelihood of, and identifying, data errors that other researchers have covered extensively elsewhere. Overall, there are four key takeaways:

- Some types of data errors, even for a small number of applications, can have large impacts on disparity estimates.
- The impacts of data errors can vary significantly across the type of error, so data errors, even for large numbers of applications, do not necessarily invalidate statistical results.
- When developing strategies for dealing with data errors, it is important to differentiate between errors that are easy to identify at low cost, such as outliers, and errors that are impossible to identify, such as valid, but incorrect values.
- To maximize the value of statistical analyses, it is important to identify and correct data errors prior to the analysis, use analytical approaches that mitigate and estimate the impacts of data errors during the analysis, and incorporate the findings from these efforts into the overall assessment of the strength and reliability of the statistical analysis when drawing conclusions.

The remainder of the report is structured as follows. Section II presents empirical evidence of the types of impacts data errors can have on fair lending analyses. This section relies on simulated data to show how small numbers of a variety of types of data errors can impact interest rate disparities. One important distinction we make here and throughout the report is between data errors that are relatively easy to identify and correct prior to analysis, such as outlier values, and data errors that are nearly impossible to identify, such as valid, but incorrect values. Section III then discusses general considerations related to data accuracy and their impact on statistical results that are specific to fair lending analyses. Section IV discusses specific data error risks that arise during fair lending analyses, and strategies for mitigating these risks. Section V concludes the discussion.

II. Evidence of Potential Impacts of Data Errors on Fair Lending Analyses

In this section we generate empirical evidence showing the potential impacts data errors that commonly arise during fair lending analyses can have on estimates of pricing disparities for mortgages. For all analyses in this section, we use a simulation approach. Specifically, we simulate a set of data for a sample of mortgage loans, estimate a pricing disparity using this sample, and then inject a variety of types of data errors into the data to assess their impact on the estimated disparity.

The major benefit of using a simulation approach is that we know exactly what results the statistical analysis should generate, since we control the data generating process. This allows us to isolate just the impact of data errors, which is very important when comparing and contrasting how different types of data errors impact the statistical results. The major concern about using a simulation approach is the generalizability of the results. This is less of a concern here, since it is difficult to generalize the impact of data errors for any analysis, regardless of whether we use an actual real-world dataset. Generalizability is difficult because the impact of data errors will be analysis-specific and depend on several factors, such as the number and type of data errors, correlations between the errors and other variables in the analysis, and the objective of the analysis. Because of these complexities, our objectives here are only to show examples of the types of impacts that data errors could generate, and to highlight important analytical aspects to consider when assessing the potential impacts of data errors.

Pricing Analysis

For the pricing analysis, we inject a variety of types of data errors into a simulated dataset of mortgage loans to assess their impact on an interest rate disparity. Generating the simulated dataset is the first step in this analysis. We begin by specifying a simple decision-making process where a mortgage lender considers FICO score, property type, and income to identify an initial rate, and then allows loan officers to make discretionary adjustments to generate the final rate. To keep the analysis simple and focused, we deviate from typical mortgage pricing and do not allow applicants to buy up or buy down the rate and we assume the lender charges each applicant the same amount of fees. We then simulate data for a sample of 1,000 applicants and apply the pricing policies to generate the rate the lender charges to each applicant. Appendix A details the entire process we use to generate the simulated dataset.

Table A1 in Appendix A provides summary statistics for all 1,000 applicants in the simulated dataset. Tables A2 and A3 provide summary statistics by minority status. In general, the summary statistics reflect data from typical fair lending analyses of mortgages. FICO scores and incomes are generally high overall, and a small percentage of applicants apply for a mortgage secured by a manufactured home. Looking at the results by minority status, on average, minorities have lower FICO scores, lower incomes, and are more likely to apply for a loan secured by a manufactured home. The average rate is 6.7987% for minorities and 6.2887% for non-minorities, resulting in an unconditional disparity of 51.0 basis points (bps).

Given the differences in FICO score, property type, and income by minority status shown in Tables A2 and A3, as well as differences in the loan officer discretion by minority status discussed in Appendix A, it is not surprising to have a large unconditional disparity. To determine whether discrimination potentially drives any portion of the unconditional disparity, we use regression analysis to first control for the impacts of the legitimate policy factors and then re-estimate the disparity. Appendix B presents the details of this regression analysis. Table B1 provides the Ordinary Least Squares (OLS) regression results based on the 1,000-applicant sample. Consistent with the lender's decision-making process, applicants with lower FICO scores and lower incomes, and those applying for a mortgage secured by a manufactured home pay a higher rate on average. The estimated rate disparity is now 0.2160, which means that minorities pay a rate that is 21.60 bps higher on average than non-minorities after accounting for differences in legitimate policy factors. This disparity is statistically significant at the 99 percent confidence level (p-value <.0001). For the analyses that follow, this 21.60 bp disparity is the benchmark against which we compare disparity estimates based on data with errors.

To assess the impact that data errors can have on the estimated pricing disparity of 21.60 bps, we inject nine types of data errors into the simulated data and then re-estimate the pricing disparity. When injecting errors, we follow 5 guiding principles. First, we only inject the types of errors that commonly arise during fair lending analyses. Second, we assess the impacts of each type of data error independently and never inject more than one type of data error for a given analysis. Third, we inject errors into loans one at a time and sequentially. Fourth, we only inject errors into a small number of total loans (5).[4] Finally, we inject errors in such a way as to generate the largest impacts on the actual disparity estimate. Specifically, for each type of data error, we inject errors to elicit the largest increase in the disparity estimate, and then inject errors to elicit the largest reduction in the disparity estimate. This general approach provides one example of the possible range of impacts that a small number of each type of data error can have on a disparity estimate. Of course, different approaches, such as injecting errors into a larger number of loans, will lead to different estimated impacts.

[4] Our choice of 5 is solely meant to reflect a "small" number of errors and is not based on any formal statistical analysis. The results of the analysis would differ if we had chosen a different number of loans with errors.

The nine types of data errors we analyze are:

1. Dependent Variable (Rate): Mis-placed Decimal Points

   For this data error type, the decimal point on rate is incorrectly two places to the left. For example, if the actual rate for a loan is 6.25%, this error would change the rate in the data for analysis to 0.0625.

2. Dependent Variable (Rate): Outliers

   This data error type is an outlier value for rate. We define an outlier as four standard deviations above the mean. From Table A1 in Appendix A, the average rate is 6.4137% and the standard deviation is 0.5695. A rate four standard deviations above the mean is therefore 8.6917%. For example, if the actual rate for a loan is 6.25%, this error would change the rate in the data for analysis to 8.6917%.

3. Independent Variable (Income): Mis-placed Decimal Points

   For this data error type, the decimal point on income is incorrectly two places to the left. For example, if the actual income for a loan is $120 thousand, this error would change the income in the data for analysis to $1.20 thousand.

4. Independent Variable (Income): Outliers

   This data error type is an outlier value for income. We define an outlier as four standard deviations above the mean. From Table A1 in Appendix A, the average income is $74.4879 thousand and the standard deviation is 20.6265. An income four standard deviations above the mean is therefore $156.9939 thousand. For example, if the actual income for a loan is $120 thousand, this error would change the income in the data for analysis to $156.9939 thousand.

5. Independent Variable (FICO Score): Small Errors Near Policy Thresholds

   This data error type is a small error in the FICO score near a policy threshold, causing an incorrect mapping of FICO scores to policy categories. For our mortgage lender, the two policy thresholds for FICO score are 660 and 720, so this data error type changes a FICO score from just below (above) 660 or 720 to just above (below) 660 or 720. For example, if the actual FICO score for a loan is 659, this data error type would change FICO score in the data for analysis to 661. This error would cause the FICO_660 and FICO_660_720 0/1 flags used in the regression analysis to be incorrect.

6. Independent Variable (FICO Score): Incorrect Classification of Null/Missing Values When Generating 0/1 Flags for Regression Analysis

   For this data error type, the FICO_660 and FICO_720 0/1 flags used in the regression analysis do not accurately reflect loans with missing values. As an example, suppose the intention is to classify all loans with a missing FICO score into the under 660 category, i.e., the FICO_660 0/1 flag equals 1. A common coding approach to do this is, "if FICO score < 660 then FICO_660 = 1, else FICO_660 = 0". However, different statistical software treat missing values differently. Some software treat them as the smallest possible value, and some treat them as the largest possible value. If we were using a statistical software that treats missing values as the largest possible value, the previous example code would incorrectly classify loans with missing FICO scores as FICO_660 = 0.

7. Minority Variable: Incorrect Values

   For this data error type, the applicant's minority status is incorrect. For example, if the applicant for a loan is actually a minority (non-minority), this error would change the applicant's race in the data for analysis to non-minority (minority).

8. Duplicate Observations

   For this data error type, there are duplicate loans in the data.

9. Missing Values

   For this data error type, one or more variables have a missing value for a given loan, which causes that loan to be excluded from the regression analysis.

As constructed, data error types 5 and 7 are particularly challenging, because they are cases where a variable takes on incorrect, but valid values. It is relatively easy to identify these errors in small audit samples, but it is nearly impossible to identify and correct them across all loans in a dataset. Each of the other data error types are relatively easy to identify and correct across all loans in a dataset with a careful analysis of summary statistics and outlier plots. Although the data accuracy checks for actual analyses would typically identify and correct these errors, for the analyses in this section we assume that the data accuracy checks missed these errors.

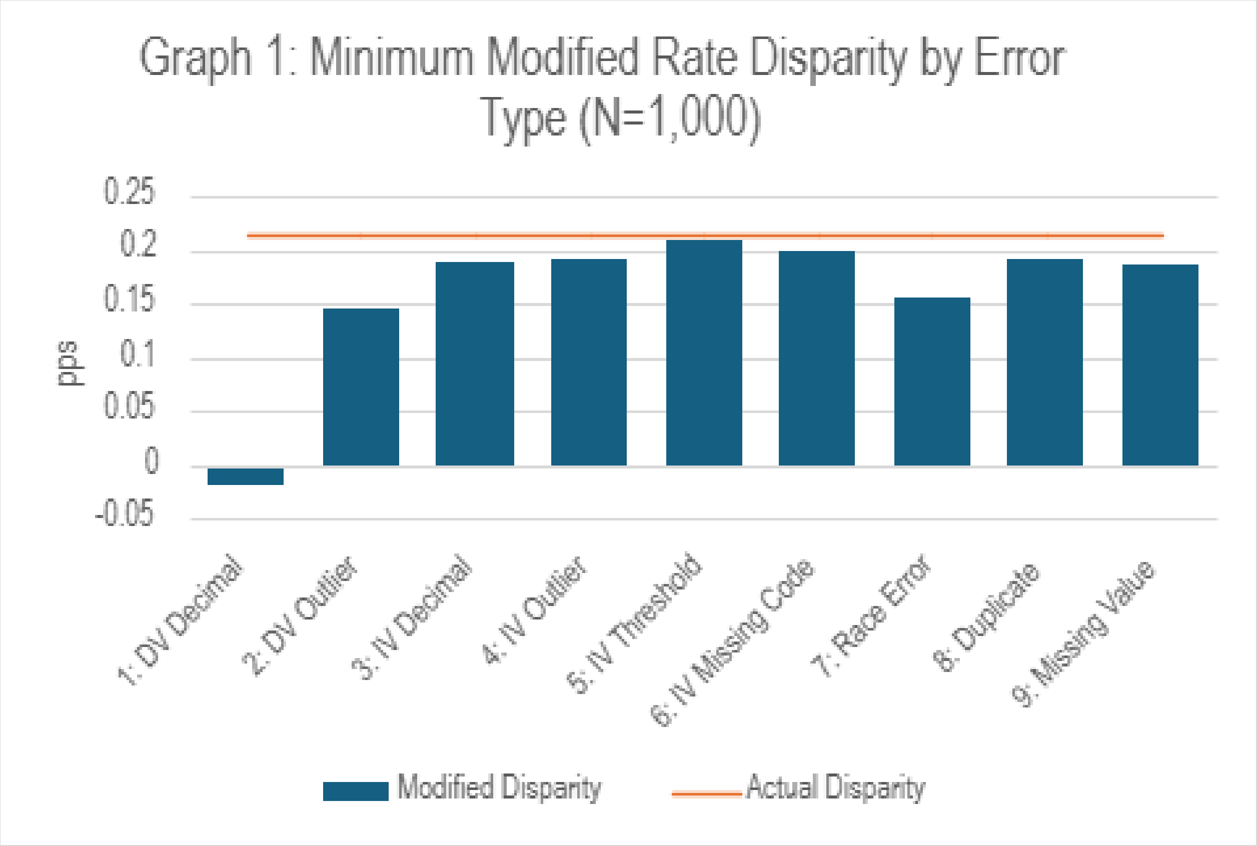
As an example of how we inject errors into the data and assess their impact on the disparity estimate, we walk through the process for identifying the largest reduction the first data error type discussed above (mis-placed decimal point in the rate) can have on the estimated rate disparity. We begin by injecting this error into the first loan, i.e., we divide the rate for loan 1 by 100. With this modified data, we re-run the OLS regression in Appendix B and save the disparity estimate. We then go back to the original dataset and inject the error into the second loan, i.e., we divide the rate for loan 2 by 100. With this modified data, we re-run the OLS regression in Appendix B and save the disparity. We continue this process for all 1,000 loans to generate 1,000 modified disparities. We then identify the one loan corresponding to the smallest modified disparity. Using the original data, we then inject the error into just that loan by dividing the rate for that loan by 100. This is now the dataset we use for the start of the second iteration. With this dataset, we again apply the same error injection process, one loan at a time, to generate 999 modified disparities.[5] We then identify the one loan corresponding to the smallest modified disparity. Using the dataset from the end of the first iteration, we then inject the error into just that loan by dividing the rate for that loan by 100. This is now the dataset we use for the start of the third iteration. We continue this process for 5 iterations, so at the end of the process, the dataset will have 5 loans with the rate error.[6] The difference between the original disparity estimate of 21.60 bps and the disparity estimate using this modified dataset at the end of 5 iterations is the estimate of the largest reduction this error type can have on the rate disparity. Following a very similar process, we also estimate the largest increase that this error type can

[5] For the second iteration there are only 999 modified disparities, because we do not inject the error into the one loan that already had an error injected from the first iteration.

[6] Note that this iterative approach may not identify the set of 5 applications for which injecting data errors would lead to the largest disparity estimate. We chose to use an iterative approach, since it was much more computationally feasible than injecting errors into all 8,250,291,250,200 possible combinations of 5 applications in a sample of 1,000 applications and then estimating regression models for each of these modified datasets.

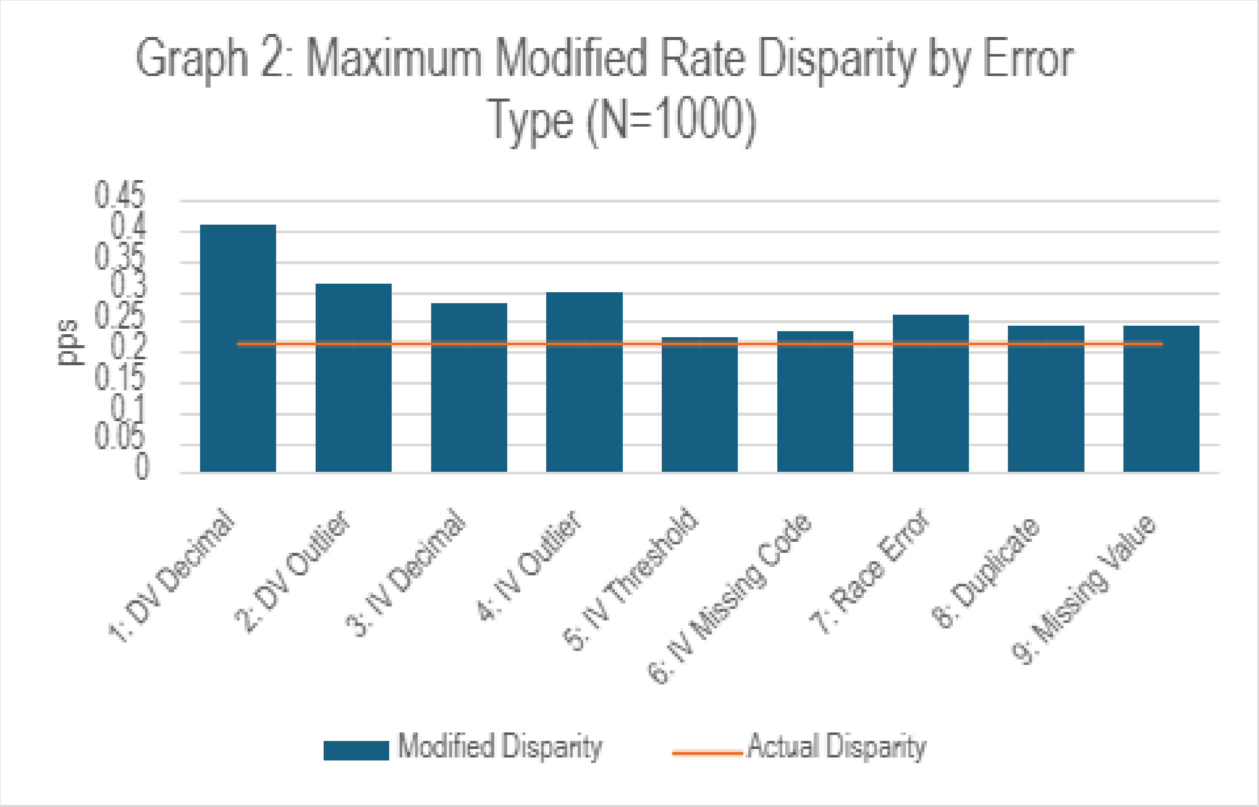
create on the rate disparity. We then use this overall approach to determine the largest reductions and increases that injecting each type of data error into 5 loans has on the disparity estimate.

Graph 1 provides a bar chart showing the minimum disparity estimates we could achieve by injecting each of the data error types into 5 loans. The horizontal line at 0.2160 denotes the actual disparity estimate using the dataset with no errors. As an example for how to read the bar chart, by injecting the first data error type (mis-placed decimal point in the rate) into just 5 loans we were able to reduce the disparity estimate from 21.60 bps to approximately –2 bps.



There are two main takeaways from Graph 1. First, data errors, even for a very small number of loans, can have a big impact on the disparity estimate, and lead to incorrect conclusions. This is consistent with the common belief that data accuracy is important because it can undermine the reliability of the statistical results. Second, although each type of data error impacted the disparity estimate, there was significant and meaningful variation in the magnitude of the impacts across error types. Mis-placed decimal points and outliers in rate clearly had the largest impacts on the disparity estimate. Although these impacts were large, they are not as concerning, because it should be relatively easy to identify and correct mis-placed decimal points and outliers across all loans in a dataset prior to the fair lending analysis. More concerning is the larger impacts caused by errors in the race variable. Since these errors are incorrect, but valid values, they are nearly impossible to identify across all loans in the dataset prior to the analysis. Each of the other data error types showed relatively small impacts. All of these results illustrate the more nuanced point that not all data accuracy issues undermine the reliability of the statistical results. It is important to re-emphasize here that these results are just one example of the type of impacts data errors could have on disparity estimates. Although it is somewhat intuitive that errors in the outcome and race variables showed the largest impact on the disparity estimate, these results might not be generalizable, since they are based on the specific assumptions we have made for these analyses.

Graph 2 provides a bar chart showing the maximum disparity estimates we could achieve by injecting each of the data error types into 5 loans. The horizontal line at 0.2160 again denotes the actual disparity estimate with no data errors. Not surprisingly the findings are somewhat similar to those in Graph 1 with errors in rate and race having relatively larger impacts on the disparity estimate. Unlike in Graph 1, outliers and decimal errors in income also show relatively larger impacts on the rate disparity here as well. Overall, the two main takeaways are again the same, data errors can matter, but they do not necessarily always matter.

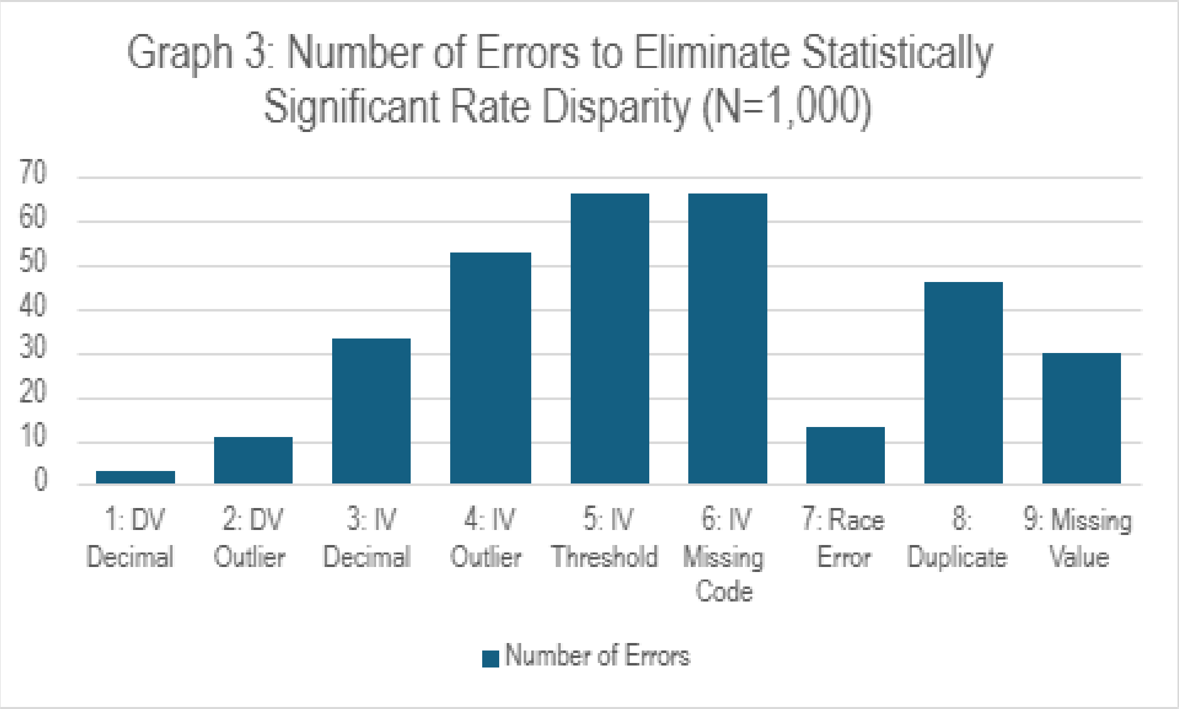


As an alternative way of viewing the impact of data errors on the rate disparity estimate, we also identify how many loans need to be in error before the statistically significant disparity in Table B1 is no longer statistically significant at the 95 percent confidence level. This representation of the results focuses directly on how data errors can change the qualitative conclusions of an analysis. For this analysis, we use the exact same analytical approach detailed above for Graph 1, but instead of stopping after 5 iterations, we continue until the estimated disparity is no longer statistically significant.

Graph 3 presents the results.[7] As an example of how to read the bar chart, we needed to inject error type 4 (outliers in income) into 53 loans before the estimated rate disparity was no

[7] Only 64 loans had FICO scores within the ranges (+/- 5 points) we used to define a threshold error. Injecting this error into all 64 of these loans did not eliminate the statistically significant pricing disparity. Therefore, for Graph 3 we just set the number of errors for this error type to 66, which is the maximum number of errors across all error types.

longer statistically significant. The results in Graph 3 are consistent with Graphs 1 and 2. Injecting just a small number of errors into the rate and race variables quickly eliminated the statistically significant pricing disparity, suggesting that errors in those two variables were the most impactful. In addition, there is wide variation in the number of loans that needed to be modified across data error types to eliminate the statistically significant disparity.



To put Graph 3 into context, for a lender with 1,000 loans, the error threshold for HMDA Reg C compliance reviews is 5.1%.[8] An error rate of 5.1% for our sample of 1,000 loans translates into 51 loans with errors. For both rate and race, as well as for error type 3 (decimal point error) for income, the numbers of loans with errors necessary to eliminate the statistically significant disparity were all well below this threshold. Therefore, the number of errors in these

[8] See the sample size table on page 4 of 201708_cfpb_ffiec-hmda-examiner-transaction-testing-guidelines.pdf

variables in Graph 3 would not raise concerns from a HMDA Reg C compliance perspective, but they would significantly undermine the reliability of the statistical results.

As a final exercise exploring the impact of data errors on rate disparity estimates, we show that garbage-in does not always result in garbage-out. Using the original simulated dataset from Appendix A, the rate disparity was 21.60 bps, which was statistically significant at the 99 percent confidence level. We now inject errors into the rate for every loan, so that every rate is incorrect. Specifically, for each loan, we take a random draw from a normal distribution with mean 0.25 and standard deviation 0.125 and add this amount to the actual rate to create a modified rate. This process altered the actual rates from between –27.8 bps and 67.8 bps across the 1,000 loans. Using this modified rate, we then re-ran the OLS regression in Appendix B. Table B2 presents the OLS results from this regression. The estimated rate disparity using this modified rate is 21.31 bps, which is statistically significant at the 99 percent confidence level. In addition, comparing the results in Tables B1 and B2, the OLS regression results are very similar to those using the correct data. In summary, compared to using the correct data, a data set where rate is incorrect for every loan yielded a disparity estimate that was less than one basis point different and a set of regression results that were very similar. Although a simple, and contrived example, it shows that data accuracy is more nuanced than simply garbage-in / garbage-out.

III. General Data Accuracy Considerations for Fair Lending Analyses

As discussed above, there are several aspects to data accuracy that analysts should consider when conducting data analyses. This section discusses 5 general aspects of data accuracy that are particularly relevant to fair lending analyses. These general points provide a foundation for the next section, which discusses specific data error risks that arise during fair lending analyses.

First, when conducting fair lending analyses, the expectation should be that the data will likely contain some level of data errors. When working with large volumes of data (both numbers of applications/loans and numbers of variables), it is difficult to ensure complete data accuracy, especially when there are valid, but incorrect values for variables. With enough time and resources, 100 percent accuracy is possible, but the costs of this strategy likely significantly outweigh the benefits. Consistent with this view, supervisory agencies explicitly acknowledge that 100 percent accuracy is not required. As one example, for mortgage data, the CFPB guidance on HMDA Compliance exams allows for tolerances for action date, application date, loan amount, and income.[9] As a result, it is important to use audit sampling to assess the prevalence and potential impact of data errors that are difficult to identify and correct, and to understand that some level of data errors does not necessarily invalidate the statistical results.

Second, as shown in the previous section, all data errors are not equal in their impact on the statistical analysis. Therefore, it is important to consider how likely it is that data errors that are difficult to identify and correct will substantively or meaningfully change the conclusions of the analysis. This is particularly important for fair lending analyses given the monetary and reputational consequences of a violation of the Equal Credit Opportunity Act (ECOA) or Fair Housing Act (FHA). Supervisory agencies acknowledge this point as well. For example, the IFLEP states, "If the LAR data are inconsistent with the information contained in the loan files, *depending on the nature of the errors* (author's emphasis), examiners may not be able to proceed with a fair lending analysis until the LAR data have been corrected by the institution."[10] When

[9] See page 40 of Home Mortgage Disclosure Act (HMDA)

[10] See 2021-june-17-fairlend.pdf, page 17.

conducting a data accuracy assessment of a sample of loans, it is therefore important to differentiate data errors that likely will have little impact on the statistical analysis from those that might have a larger impact on the statistical analysis.

Third, there are common and recurring patterns to data accuracy issues that are specific to fair lending analyses, especially those focused on mortgage data. One common pattern is that variables lenders are required to report under HMDA typically have better data accuracy than other variables because of the compliance risk, as well as the potential monetary penalties and reputational harm. A second common pattern is that data for denied applications are often less complete and accurate than for approved applications. For applications that lenders determine early in the process that they likely will not approve, there is less incentive to gather a complete set of accurate data. As one concrete example of this, many lenders consider eligibility criteria as a first step in the decision-making process to weed out clear denials early to avoid the cost of pulling a credit bureau report. Obviously, credit bureau data would not be available for these applications. Knowing that some applications and some variables are more likely to have data issues should help focus the data accuracy check. In addition, if it is possible to determine when credit decisions are based on only a subset of policy factors, analytical approaches such as estimating regression models for subsets of applications and using interaction variables to identify applications where a given variable was not considered can be utilized.

Fourth, for fair lending analyses, the data decision-makers rely on to make credit decisions may not accurately reflect the true characteristics of all applicants, and these differences may vary systematically across demographic groups. Income is a good example of this concern because it is complex and contains multiple components, and is therefore difficult for applicants to report accurately and consistently. Given these complexities, some demographic groups may systematically neglect to report some acceptable sources of income. In addition, loan officers might work harder with applicants from some demographic groups to ensure they report all acceptable sources of income. If either of these occur, when making credit decisions, decision-makers would not have data that accurately reflects the true characteristics of all applicants. During fair lending analyses, it is important to assess whether decision-makers treated all applicants fairly and consistently based on the data/information available to the decision-maker, even if the data do not accurately reflect the true characteristics of applicants. It is equally important to analyze variation across demographic groups in whether the data available to the decision-maker accurately reflects applicant's true characteristics. This second analysis is an important component of the data accuracy analysis of a sample of loans.

Finally, lenders' incentives are not always fully aligned with maintaining accurate data. Lenders have clear incentives to maintain extensive, clean data for internal analyses to accurately assess risk and identify evidence of discrimination. However, it is important to acknowledge that lenders also have incentives to not maintain accurate data. Supervisory agencies currently request large volumes of data from lenders during fair lending exams, which they use to conduct statistical analyses testing for discrimination. Statistically and economically meaningful disparities from these analyses often drive conclusions of potential violations of the ECOA and FHA. One potential strategy lenders can use in response is to argue that the statistical analysis is not reliable, because of data accuracy issues. This creates incentives to not maintain accurate data. We in no way are stating here that lenders are intentionally providing supervisory agencies with inaccurate data. We are just highlighting that lenders have some incentives to not maintain accurate data, and that this should be one consideration during fair lending analyses.

IV. Specific Data Accuracy Issues Relevant for Fair Lending Analyses

This section discusses specific data accuracy issues relevant for fair lending analyses. We focus specifically on data issues that arise during statistical analyses and do not cover the much broader concept of data integrity, which includes data accuracy as well as several other components such as equipment protection, access control, data encryption, backups, redundancy, training, maintenance, and more[11] Overall, there are three primary themes to the discussion:

- For data errors that are easy to identify for all applications, such as outliers, implement cost-effective strategies to quickly identify and correct these errors.
- For data errors that are impossible to identify for all applications, such as incorrect, but valid values, assess their prevalence and potential impacts on the statistical analysis.
- Include data accuracy analysis results into the overall assessment of the reliability of the statistical analysis when drawing conclusions.

We now present specific data accuracy issues that arise during fair lending analyses. To facilitate the discussion, we assume that an application is the relevant observation for analysis.

1. Data Entry is a Common Source of Data Errors

The credit application and corresponding financial reports are primary sources of data for fair lending analyses. For these documents, lenders typically need to use some data entry to transfer their contents into an electronic database for analysis. Data entry comes with heightened risks of transcription and transposition errors, which are often difficult to identify, as well as

[11] There is a large volume of available resources and guidance on the details of broader data integrity best practices. Some examples include, McGilvray (2021), Maydanchik (2012), Olson (2003), Moses et.al. (2022), Walsh (2021), Blokdyk (2021), Kumar (2023), "The DAMA Guide to the Data Management Body of Knowledge: (DAMA-DMBOK Guide)," Data Accuracy in 2025: Ensure Reliable, Quality Data, and Guide to Data Quality: Ensuring Accuracy and Consistency in Your Organization | WhereScape.

other errors, such as formatting, omission, and duplicate errors. Given these risks, safeguards and checks are necessary to minimize the likelihood of these errors.

A large volume of guidance currently exists on strategies for minimizing data errors during data entry,[12] so we highlight just two practices that are particularly important and relevant to the focus of this report. First, end data users should participate in developing and updating data entry policies, procedures, and processes to ensure the specific variables, data structure, formats, and accuracy meet their needs for fair lending analyses. Second, to the extent possible, lenders should use technology to minimize the use of manual data entry and to identify and flag potential data errors, especially errors that are difficult to identify, such as incorrect, but valid values.[13]

2. Obtain or Create a Data Dictionary for Each Source Dataset

Fair lending analyses often use data from multiple sources, such as the application, credit bureau reports, financial statements, tax documents, and more. It is important to obtain or create a data dictionary for each source dataset. For each variable in a source dataset, the data dictionary should contain:

- the variable name
- a brief description of the variable
- the source of the data

[12] See Transforming Data Entry: How to Eliminate Errors and Maximize Efficiency 2025 for one example of a thorough summary of data entry best practices.

[13] Some examples of this technology include Optical Character Recognition (OCR) which can convert electronic documents into editable electronic data (see Raj et.al. (2023) for a summary), Intelligent Character Recognition (ICR) which can convert hand-written documents into editable electronic data (see Winardi et.al. (2024) for a summary), Robotic Process Automation (RPA) which automates sets of repetitive data entry tasks typically conducted by humans and that can include OCR or ICR (see Axmann and Harmoko (2020) for a summary), and a variety of AI-based approaches, such as Intelligent Document Processing (IDP), AI-Powered Data Validation, Natural Language Processing (NLP), and Predictive Data Entry and Auto-Correction that learn from past data entry to flag for review values, formats, and structure during current data entry that deviate from expected patterns (see Kolandaisamy et.al. (2024) for a summary).

- the variable owner
- details on the creation of each variable generated as a combination of other variables
- the format of the values (character, string, integer, float, Boolean, date/time, other)
- a list of all valid values for categorical variables and valid ranges of values for continuous variables
- any special code values (such as 9999) and their meaning
- an explanation of every situation when the variable can contain missing values

Each dictionary should also include the current version of the dataset, when the version was first created, and when it was last updated.

3. Define Observations and Identify the Number of Observations in Each Source Dataset

It is important to clearly understand the definition of an observation in each source dataset, especially if it is necessary to combine multiple datasets for the analysis. For most datasets relevant to fair lending analyses, an observation will be an application or loan identified by a unique application or loan id. However, this is not always the case. For example, the data might contain a separate row of data for each applicant on the application. As a second example, for servicing data, there might be separate rows of data for each loss mitigation option the servicer considered for a given loan.

It is also important to create an inventory of all relevant observation counts, since these counts will be integral to assessing data accuracy throughout the statistical analysis. This inventory should include three general types of counts. First, it should include the total number of applications relevant to the analysis, which is typically defined as all applications with a final credit decision during a given time period. Lenders typically identify this count using the credit application source dataset and then validate it using independent sources. Second, the inventory should include the total number of observations in each source dataset. Given that different source datasets may have different definitions of an observation, the number of observations in a given source dataset may not be the same as the total number of applications relevant to the analysis. Finally, the inventory should include the total number of unique applications in each source dataset. When the total number of unique applications in a source dataset differs from the total number of applications relevant to the analysis, follow-up analyses are needed to understand the reasons for these differences.[14]

4. Dataset Format Transformations Can Create Data Errors

Source datasets are often plain text CSV files or EXCEL files, and Economists typically conduct fair lending analyses using statistical software packages such as STATA, SAS, R, or Python. Therefore, it is often necessary to transform source datasets into a format that is appropriate for the relevant statistical software. This transformation is relatively easy as most statistical software packages can read in and transform data from a variety of formats. In addition, there are several software packages such as STATTRANSFER designed specifically to transform data into a variety of different formats. However, these transformations can create data errors. Many of these errors stem from user error, specifically users not understanding or applying all of the encoding options for the given statistical package. However, errors can also arise because of a variety of formatting issues in the data.

[14] Either during data collection or during the statistical analysis, Economists often narrow the number of observations on several dimensions, such as business unit, delivery channel, product/program, and others. Given the importance of observation counts to monitoring data accuracy, it is important to generate each of the three general types of counts for each subset relevant for the statistical analysis.

When transforming data, it is important to always compare the data from the transformed dataset to the source dataset. Specifically, it is good to compare counts, summary statistics, and frequency distributions of each variable in the transformed dataset to the source dataset to make sure the transformation process generated no data errors. However, given the thorough data accuracy checks suggested in item 7 below, spot checks of key variables here often suffice.

5. Combining Source Datasets Can Create a Variety of Data Errors

When there are multiple source datasets, it is necessary to combine these datasets into one master dataset for analysis. Depending on the number of source datasets, variation in observation definitions across datasets, and a variety of formatting issues, this step of the analysis can create a significant risk of generating data errors.

There are two general approaches to combining datasets, merging, which combines datasets observation by observation, and appending, which stacks datasets. To help minimize the risk of data errors arising when merging source datasets, data analysts should consider the following:

- Ensure that there is a unique application identifier that is common to each source dataset.
- Ensure that each source dataset has one observation per application id by eliminating duplicate observations and collapsing applications with multiple observations into one observation for source datasets with multiple observations of data for some applications.
- Assess whether each set of potential duplicate applications are true duplicates across all variables in the dataset, or possibly very similar but slightly different applications at two different points in time.
- Ensure that each variable other than the unique application identifier is in only one source dataset; relatedly, if the same variable is in multiple datasets, verify they contain the same information, and if not, make sure to keep the correct variable.
- Ensure that each variable name other than the unique application id name appears in only one source dataset.

- Sort each source dataset by the unique application identifier and then merge the source datasets by that identifier.
- Once the merge is complete, verify that the total number of applications is correct.
- Once the merge is complete, analyze how many observations came from each of the source datasets, and verify that these counts meet expectations based on the inventory of counts from step 3 above.

To help minimize the risk of data errors arising when appending source datasets, data analysts should consider the following:

- Ensure that each dataset contains the exact same variables and variable names.
- Ensure that the data types (character, string, integer, float, Boolean, date/time, other) and value formats (percentage vs ratio for example) of each variable are exactly the same in all datasets.
- Once the append is complete, verify that the total number of applications is correct.

6. Data Formatting Can Create Data Errors

Generating summary statistics and running regression analyses require variables with a numeric format. Source datasets typically include some variables with non-numeric formats, such as character or string, so it is necessary to convert these variables to numeric format prior to the analysis. This conversion process is another source of potential data errors. For example, a character variable might include invalid numeric values, special characters, or delimiters that result in either missing values or incorrect values when converted to numeric format. As a second example, changing formats can alter the precession of values, such as changing “645.262” to 645.2 or maybe 600. As a final example, some string variables include leading zeroes that have meaning, and the transformation process that converts these variables into numeric variables can drop these leading zeroes.

When reformatting a variable, it is important to first review a frequency distribution of the variable to determine the precision of values and to identify any non-numeric characters. It is also important to develop a clear understanding of the coding commands that convert formats, all of their related options, and how they handle unusual characters. After reformatting a variable, review the frequency distribution of the newly created numeric variable to ensure all values are within expected ranges and that the reformatting did not generate any unexpected changes or missing values.

7. Search for Errors in Combined Data

A variable-by-variable search for data errors is needed after combining all source datasets into one master dataset for analysis. The data dictionaries for each source dataset and the inventory of counts discussed above will be particularly valuable resources during this search. Improving data accuracy of the master dataset will have immediate benefits by improving the reliability of the current statistical analysis. The variable-by-variable search will also provide a thorough understanding of the data, which will improve the current statistical analysis as well. In addition to these short-term benefits, this search will also have broader benefits as information on the numbers and types of data errors for each variable found in this data search can be used to improve overall data collection and processing, which in turn will improve data accuracy for other fair lending analyses.

The search for data errors should include univariate, bivariate, and multivariate analyses of each variable in the master dataset. The primary objective of these analyses is to locate data errors that can be identified at low cost for all applications. These analyses should also flag any valid, but incorrect values as well, but those errors should not be the focus here given the high cost of identifying these types of errors for all applications, especially for large datasets.

For categorical variables, the univariate analysis should focus on frequency distributions for each variable. Reviewing these distributions is a quick, effective, and low-cost approach to identify invalid values, special code values, and unexpected distributions of values. For continuous variables, the univariate analysis should focus on summary statistics (number of missing values, mean, standard deviation, minimum value, and maximum value) for each variable. The number of missing values can identify potential omission errors in the source data, as well as deletion errors created during the data collection, transformation, formatting, and combining steps of an analysis. The mean should match expectations for the characteristics of typical credit applicants for the given business model, channel, time period, and product. For example, a mean FICO score of 580 would be unexpected for a mortgage lender focused on prime applicants. Significant deviations from expectations would indicate potential outliers or systematic shifts, such as decimal point errors. Extreme values for the standard deviation (such as a value of 0) would suggest limited or no variation, or potential outliers. The minimum and maximum values quickly identify special code values, invalid values, and outliers. Summary statistics will identify most types of data errors in continuous variables. However, they will not catch all errors. For example, if two applications have invalid FICO scores of 1000 and 1005, the maximum value will identify the 1005, but none of the summary statistics will identify the 1000. Therefore, it is useful to augment summary statistics for continuous variables with histograms as an additional check to identify special codes, outliers, and unexpected distributions.

The bivariate and multivariate analyses check for consistencies among groups of variables. The number and extent of these checks will depend on the relationships between the variables in the given dataset. As one example of these checks, combined loan-to-value (CLTV) should always be greater than or equal to loan-to-value (LTV). As a second example, if a proprietary credit score was not generated for underwriting of applications for product A during the last quarter of the year, the multivariate analysis should verify that the proprietary credit score contains missing values for applications for product A with a final underwriting decision in the last quarter of the year. As a final example, it is important to check the relationships of date variables, since there often is a logical sequential order to dates, such as application date, decision date, and then note rate date.

For many of the data accuracy checks suggested above, it is important to utilize variable-specific filters prior to the checks to filter the data to just applications that should have valid values for the given variable. As one example, when generating summary statistics of a continuous pricing variable, such as APR, only include originated loans, since APR is typically missing for non-originated loans. As a second example, lenders often use internal codes, such as 2222 or 9999 to maintain confidentiality, such as for income of employees, or when actual values are not relevant, such as when the decision-maker does not consider a variable for a given product or program. In these instances, generate frequency distributions, summary statistics, and histograms using all applications to identify these special codes, and then generate a second set of results excluding these special codes to create more accurate representations of the relevant distributions of the variables.

These data accuracy analyses can be time and resource intensive, so it is worth exploring automating these checks using a batch approach instead of conducting them interactively, variable-by-variable. Information from the data dictionaries on valid values and skip patterns, the numbers of applications from the inventory of counts, and business-specific rules can all be incorporated into computer code that generates the data accuracy checks, summary statistics, frequency distributions, and graphics in batch and then generates a report identifying just the data issues requiring additional review. Although automation would be significantly faster and reduce costs, especially if conducting data accuracy assessments for several datasets, we recommend the interactive, variable-by-variable approach if possible since it has the significant added benefit of providing a comprehensive understanding of the data for analysis.

8. Address Errors in Combined Data

If the univariate, bivariate, and multivariate analyses identify potential data errors, the next step is to address these errors. Although replacing all incorrect values with correct values is ideal, this may not be the most cost-effective approach in all instances. Depending on the volume and types of errors, as well as the difficulty in identifying the correct values, there are a variety of approaches and tradeoffs to consider when deciding how to address data errors.

If the number of errors for a given variable is small, the ideal approach is to replace all incorrect values with correct values. The original source datasets should be the first place to look to resolve errors. For any remaining unresolved errors, check the code, intermediate data, and processes used for data input, transformation, formatting, and combining. Exclude from the analysis any applications with data errors that remain unresolved after all of these checks. It is important to include the number of applications excluded and the reasons they were excluded in the overall summary of the statistical analysis and results.

If the number of errors for a given variable is large, replacing all incorrect values with correct values may not be cost effective. In addition, simply excluding the applications with data errors is not a feasible option either, since excluding a large volume of applications would likely significantly impact the statistical analysis. In this instance, there are several possible analytical approaches for addressing these errors. We present three standard approaches here that are easy to apply and interpret. First, for a variable with a large number of data errors, generate a 0/1 variable flagging all applications with a data error, as well as a set of 0/1 flags classifying all remaining applications into mutually exclusive categories. Then, create an interaction variable by multiplying the 0/1 data error flag and the 0/1 minority flag.[15] Including the 0/1 data error flag and the interaction variable in the regression analysis will provide estimates of how applications with data errors impact outcomes in general and how data errors potentially impact disparity estimates specifically. As an example, suppose a large number of applications had errors in the DTI variable. The first step is to generate a 0/1 variable flagging all applications with a data error. For the remaining applications, transform the continuous DTI variable into a set of mutually exclusive 0/1 flags, such as for DTI values of 0-36, 36-43, and 43-75. The interaction variable would just be the data error flag times the 0/1 minority flag. In this example the regression model would include the 0/1 data error flag, the minority flag, the interaction variable, and any two of the three remaining DTI flags. The coefficient estimate on the 0/1 data error flag would indicate how applications with DTI errors impact the outcome in general and the coefficient estimate on the interaction variable would indicate how applications with data errors potentially impact the disparity estimate specifically.

The second approach again starts by generating a 0/1 variable flagging all applications with a data error. The next step is to replace all incorrect values with either the mean, median, or mode value of the given variable. The final step is to create an interaction variable by multiplying the 0/1 data error flag and the 0/1 minority flag. Including the 0/1 data error flag and

[15] This assumes the analysis focuses on only one demographic group. If the analysis includes multiple demographic groups, it is necessary to generate separate interaction variables for each demographic group.

the interaction variable in the regression analysis will provide estimates of how applications with data errors impact outcomes in general and how data errors potentially impact disparity estimates specifically. As an example, suppose a large number of applications had errors in the DTI variable, and that the median DTI value using only applications with correct DTI values is 30. The first step is to generate a 0/1 variable flagging all applications with a data error. Then, in the DTI variable, replace all incorrect values with the median value of 30. The interaction variable would just be the data error flag times the 0/1 minority flag. In this example the regression model would include the 0/1 data error flag, the minority flag, the interaction variable, and the continuous DTI variable. The coefficient estimate on the 0/1 data error flag indicates how applications with DTI errors impact the outcome in general and the coefficient estimate on the interaction variable indicates how applications with data errors potentially impact the disparity estimate specifically.

A third approach is to pull a random and representative sample of applications, correct all of the data errors in this sample, conduct the statistical analysis using this sample, and then use the statistical results to infer results about the overall set of applications. The significant advantage of this approach is that the volume of data errors needing correction would be manageable. The disadvantage is the uncertainty that comes with using a sample to infer conclusions about the entire set of applications.

9. Check Constructed Variables and Corresponding Code

A key component of statistical fair lending analyses is to construct all variables exactly how the lender considered each variable when making credit decisions. The process of constructing variables for analysis is another common source of data errors. As one example, lenders often apply thresholds when considering variables during underwriting and pricing. For example, a lender might categorize all applicants with a FICO score below 620 as higher risk. When constructing this FICO variable for analysis, it is easy to mistakenly use the code "if FICO <= 620" instead of "if FICO < 620," which would mis-code applicants with a FICO score of 620 into the incorrect FICO category for analysis. As a second example, using the code "if FICO < 620" to create a 0/1 flag for applicants with low FICO scores instead of explicitly including a minimum value using the code "if FICO >= 350 and FICO < 620" would classify all applicants with a missing FICO score into the under 620 category if the statistical software being used treats missing values as the smallest possible value, and into the 620 and above category if the statistical software being used treats missing values as the largest possible value.

After constructing variables for the analysis, use cross-tabulations to ensure that the code is correctly classifying all applications as expected. In addition, when possible, independently validate code to ensure accuracy, especially code used to generate the final statistical results. Finally, conduct a periodic quality assurance check of all code used for the fair lending analysis to ensure there are no coding errors.

10. Check for Valid, but Incorrect Values

As noted above, one of the biggest challenges with data accuracy is identifying valid, but incorrect values, especially for large datasets. These challenges are somewhat less of a concern for credit transaction data since it is common for lenders to conduct verifications of key data such as income, employment, assets, debts, expenses, and residences. In addition to verifying the specific information applicants provided, these verifications can enhance data accuracy even further if they also verify whether applicants accurately provided all relevant information.

Although these verifications clearly improve data accuracy, some valid, but incorrect values will likely still exist, since errors might occur during verification calculations and when entering these data into electronic databases. In addition, some loan officers might fail to update initial values or the system may prohibit updates to initial values. Finally, lenders do not conduct these verifications for every variable relevant to fair lending analyses. Therefore, it is important to consider additional steps to address valid, but incorrect values.

One of the most cost-effective strategies for addressing valid, but incorrect data is to use an audit approach focused on estimating error rates and the potential impacts of errors. Since audit reviews typically use sampling, they avoid the costs of checking the entire dataset for errors, which can be significant especially for larger datasets. Conducting the audit review at the end of the statistical analysis to narrow the review to only the variables used in the analysis can further reduce costs.

The first step of these audit reviews is to generate a random and representative sample. As a rule of thumb, this sample should contain at least 30 applications. However, a larger sample of 100 or more is better, especially if the overall number of applications for the analysis is large. Using this sample, identify every data error for every variable used in the statistical analysis. Classify each error as either impactful to the statistical analysis or not. Although this classification will be difficult at times, there should be some clear-cut cases. For example, a FICO error that results in an application being classified into an incorrect policy category should be flagged as impactful, while a FICO error that does not change the FICO category would not. The final step is to analyze applications with any errors and applications with impactful errors. Possible components of these analyses include,

- Generating the overall number and percentage of applications with at least one error

- Generating the number and percentage of applications with errors separately for each variable
- Separately for each demographic group, generating the number and percentage of applications with errors, both overall and for each variable
- Generating unconditional disparity estimates using the audit sample after first correcting any data errors and excluding applications with uncorrectable data errors, and comparing these disparity estimates to the disparity estimates based on the audit sample as originally drawn
- Using the final model specification from the statistical analysis, generating conditional disparity estimates using the audit sample after first correcting any data errors and excluding applications where data errors could not be corrected, and comparing these disparity estimates to the conditional disparity estimates based on the audit sample as originally drawn

The overall objective here is to assess the extent of valid, but incorrect values and their potential impacts on the statistical analysis to help inform the assessment of the reliability of the overall statistical analysis.

An alternative to an audit review based on sampling is to check for valid, but incorrect errors for all applications. There are a wide variety of available AI-enhanced techniques that can help identify valid, but incorrect errors, especially in large datasets (see Kapoor (2025)). Very generally, the AI-enhanced approaches search for complex data patterns, and small deviations from these data patterns, within variables, across groups of variables, across datasets, and over time. These patterns are much more difficult to identify using standard, rules-based validation checks. Given that the number of errors identified with these approaches will likely be large, consider using the analytical strategies for addressing data errors discussed in item 8 (“Address Errors in Combined Data”) above. In addition, similar to the audit approach, a set of summary statistics characterizing the errors that the AI-enhanced techniques identified can help inform the assessment of the overall reliability of the statistical analysis.

11. Incorporate Results from Data Accuracy Checks into the Overall Statistical Results

As a general matter, statistical disparities should not be presented or discussed in isolation. Instead, when assessing the statistical results and drawing conclusions, it is important to discuss disparities together with the objectives of the analysis, the quality and extent of the data, the quality of the statistical analysis, and the results of robustness tests. The data quality component should focus primarily on the accuracy of the variables used in the statistical analysis and the potential for errors in those variables to impact the statistical conclusions. The following lists possible types of information to consider,

- An assessment by the data analyst who conducted the data accuracy analysis of overall data accuracy and the potential impacts of data errors on the statistical results
- Separately for each variable used in the statistical analysis, the number and percentage of applications with an error, as well as a summary of the types of errors
- Separately for each variable used in the statistical analysis, the number and percentage of applications with an error that was likely to impact the statistical results, as well as a summary of the types of errors
- A narrative summarizing the approaches used to identify, correct, and address incorrect values, as well as details about the results, such as the number and percentage of incorrect values that were corrected and regression results estimating the impacts of applications with data errors
- A narrative summarizing the approaches used to estimate the prevalence of valid, but incorrect values and their potential impacts on the statistical results, as well as details about the results of these efforts

V. Conclusion

A data accuracy analysis is a key component of every fair lending review. As with most statistical analyses, there are many details and nuances to consider when conducting data accuracy analyses. In this report we discussed a variety of these details and nuances including different types of data errors common to fair lending analyses; how and when data errors can occur; the potential impacts errors can have on statistical results and how these impacts can vary across different types of errors; strategies for identifying and correcting errors prior to the statistical analysis; and approaches to addressing data errors during the statistical analysis.

We provided empirical evidence showing, not surprisingly, that data accuracy can matter as some types of data errors, even for a small number of applications, can have large impacts on disparity estimates. However, we also provided empirical evidence showing that the impacts of data errors can vary significantly across the type of error, so data errors, even for large numbers of applications, do not necessarily invalidate statistical results. Overall, it is therefore important to conduct a thorough and comprehensive data accuracy analysis that includes identifying and correcting data errors, using appropriate analytical techniques to mitigate the effects of data errors, and assessing the potential impacts of data errors on the statistical results.

References

Axmann, Bernhard and Harmoko Harmoko. “Robotic Process Automation: An Overview and Comparison to Other Technology in Industry 4.0,” presented at the 10th International Conference on Advanced Computer Information Technologies, September 2020.

Blokdyk, Gerradus. “Validity and Reliability in Statistics: A Complete Guide - 2020 Edition,” April 16, 2021.

Kapoor, Anjali. “AI-Driven Data Cleaning: Intelligent Detection and Correction of Data Errors,” *International Journal of Computer Technology and Electronics Communication*, Volume 8, No. 1, 2025.

Kolandaisamy, Raenu, Heshalini Rajagopal, Indraah Kolandaisamy, and Glaret Shirley Sinnappan. “The Smart Document Processing with Artificial Intelligence,” presented at the International Conference on Artificial Life and Robotics, 2024.

Kumar, B. Santhosh. Data Integrity and Data Governance, IntechOpen, 2023.

Maydanchik, Arkady. Data Quality Assessment, Technics Publications, 2012.

McGilvray, Danette. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information, 2nd Edition, 2021.

Moses, Barr, Lior Gavish, and Molly Vorwerck. Data Quality Fundamentals: A Practitioner's Guide to Building Trustworthy Data Pipelines, O’Reilly Media, 2022.

Olson, Jack E. Data Quality: The Accuracy Dimension, 1st Edition, Morgan Kaufmann, 2003.

Raj, Aaryan, Sakshi Sharma, Janvee Singh, and Aastha Singh. “Revolutionizing Data Entry: An In-Depth Study of Optical Character Recognition Technology and Its Future Potential,” *International Journal for Research in Applied Science and Engineering Technology*, Vol. 11, Issue 2, February 2023.

The DAMA Guide to the Data Management Body of Knowledge: (DAMA-DMBOK Guide). Bradley Beach, New Jersey: Technics Publications, 2010.

Walsh, Susan. Between the Spreadsheets: Classifying and Fixing Dirty Data, 1st Edition, Facet Publishing, 2021.

Winardi, Sunaryo, Gunawan, Fans Mikael Sinaga, Farrel Rio Fa, and Frederic Davidsen. “Literature Study: Intelligent Character Recognition (ICR) Technology for Learning Innovation,” presented at the 2nd International Conference on Technology Innovation and Its Applications (ICTIIA), September 2024.

Appendix A: Generating the Simulated Dataset for the Pricing Analysis

As a first step in generating the simulated dataset for the pricing analysis, we use the specifications below to generate data for minority status, FICO score, property type, and income for 1,000 applicants. These variable specifications reflect the type of data generally found in typical mortgage applications. Specifically, on average, minority applicants have lower FICO scores and incomes, and are more likely to apply for a loan secured by a manufactured home.

Minority = 1 if a random draw from a uniform distribution > .75; 0 otherwise

FICO Score
Random draw from a N(660, 40) distribution if Minority = 1
Random draw from a N(700, 50) distribution if Minority = 0

Manufactured Home (MH)
If Minority = 1 then MH = 1 if a random draw from a uniform distribution < .2; 0 otherwise
If Minority = 0 then MH = 1 if a random draw from a uniform distribution < .1; 0 otherwise

Income (in thousands)
Random draw from a N(60, 15) distribution if Minority = 1
Random draw from a N(80, 20) distribution if Minority = 0

We then use a set of mortgage pricing policies that we created for this report to determine the interest rate for each of the 1,000 applicants.[16] There are three parts to these pricing policies. First, the lender formally considers three variables when setting interest rate: FICO score, property type, and income. Specifically, the lender charges different rates to applicants with FICO scores below 660, 660 to below 720, and 720 or above; charges a higher rate for loans secured by manufactured housing; and considers income in continuous form, charging a lower

[16] Mortgage pricing is very complex, consisting of multiple components, such as rate, points, and fees, as well as tradeoffs among these components. All of these complexities detract from the objective of this analysis, which is to generate examples of the types of impacts data errors can have on statistical analyses. Therefore, we use a simplified set of pricing policies, which depend on only three factors (FICO score, property type, and income), do not allow applicants to buy-down or –up the interest rate, and assume that each applicant pays the same amount of fees.

rate to applicants with higher incomes. Second, the lender allows loan officers discretion to deviate from formal pricing policies. We incorporate this discretion with the variable DT, which equals a random draw from a uniform distribution multiplied by 0.4 for each minority applicant, and 0 for every non-minority applicant. Constructed in this way, minorities pay 20 basis points (bps) more on average due to loan officer discretion than non-minority applicants. Third, we allow for the possibility that random errors occur when the lender applies pricing policies, which is not uncommon, especially when there is a large volume of applications and loan officers. We incorporate these errors with a random draw from a normal distribution with mean 0 and standard deviation of 0.5 for each applicant. Equation A1 presents this entire pricing policy in equation form.[17] [18]

$$Rate = 7 + (.2 * FICO_{660}) + (.1 * FICO_{660-720}) + (.1 * MH) - (.01 * Income) + (1 * DT) + \big(1 * N(0,.5)\big) \tag{A1}$$

Applying Equation A1 generates interest rates for each of the 1,000 applicants in the sample. As an example of how equation A1 works, suppose applicant 1 has a FICO score of 680, an income of $100,000, and applies for a loan secured by a manufactured home; the lender applies a discretionary upcharge of 25 bps; and an error of 5 bps was made when determining the rate. Based on equation A1, the lender would charge this applicant a rate of 6.5% (= 7 + (.2 * 0) + (.1 * 1) + (.1 * 1) - (.01 * 100) + (1 * .25) + .05).

[17] Equation A1 does not explicitly include a flag for FICO scores above 720, because when a regression equation includes an intercept term and a continuous variable, such as FICO score, which is transformed into mutually exclusive categorical variables that cover the entire range of the variable (i.e. < 660, 660-720, and >= 720), one of the categorical variables must be excluded from the equation.

[18] We have chosen each of the parameters in Equation A1, i.e., the .2, .1, .1, -.01, 1, and 1 to generally reflect the types of relationships seen in mortgage pricing. Different parameter values will generate different rate values and different simulation results.

Table A1 provides summary statistics for all 1,000 applicants in the simulated dataset, and Tables A2 and A3 provide summary statistics by minority status.

Table A1: Summary Statistics for 1,000 Application Pricing Sample

```
Variable          N            Mean        Std Dev       Minimum        Maximum
------------------------------------------------------------------------------------
rate            1000          6.4137        0.5695         4.7428         8.2940
fico            1000        692.1616       49.5524       533.8335       876.5885
fico_660        1000          0.2630        0.4405         0              1
fico_660_720    1000          0.4520        0.4979         0              1
mh              1000          0.1360        0.3430         0              1
income          1000         74.4879       20.6265        18.4386       128.5331
minority        1000          0.2450        0.4303         0              1
------------------------------------------------------------------------------------
```

Table A2: Summary Statistics for Minority Applicants in the 1,000 Application Pricing Sample

```
Variable          N            Mean        Std Dev       Minimum        Maximum
------------------------------------------------------------------------------------
rate             245          6.7987        0.5494         5.5302         8.2940
fico             245        659.6281       37.3430       533.8335       767.9843
fico_660         245          0.5020        0.5010         0              1
fico_660_720     245          0.4531        0.4988         0              1
mh               245          0.2245        0.4181         0              1
income           245         58.2693       14.8714        18.7127       101.1408
minority         245          1             0              1              1
------------------------------------------------------------------------------------
```

Table A3: Summary Statistics for Non-Minority Applicants in the 1,000 Application Pricing Sample

```
Variable          N           Mean        Std Dev       Minimum        Maximum
------------------------------------------------------------------------------------
rate             755         6.2887        0.5179         4.7428         7.8773
fico             755       702.7189       48.4407       561.6101       876.5885
fico_660         755         0.1854        0.3889         0              1
fico_660_720     755         0.4517        0.4980         0              1
mh               755         0.1073        0.3097         0              1
income           755        79.7509       19.4658        18.4386       128.5331
minority         755         0             0              0              0
------------------------------------------------------------------------------------
```

Appendix B: Regression Model for the Pricing Analysis

The typical approach to fair lending statistical analyses is to first build a regression model that accurately reflects the lender's stated decision-making process, and then to add a minority flag to check whether being a minority explains any remaining variation in the outcome variable. Based on this approach, and equation A1, we include the following variables in the regression model:

- 0/1 flag for FICO scores < 660
- 0/1 flag for FICO scores from 660 to below 720[19]
- 0/1 flag for manufactured housing
- Income as a continuous variable
- 0/1 minority flag

There are two important caveats here. First, the lender's decision-making process (equation A1) includes a random noise component, which we do not explicitly include here. The random noise component is equivalent to the error term of a regression model, which includes all unobserved factors that impact the dependent variable (interest rate in our example).[20] Since it is unobserved, it is never included when estimating a regression model.

Second, the lender's decision-making process (equation A1) includes DT, but we do not include it here. We constructed the DT variable such that the discretionary adjustment was 20 bps higher for minorities on average. This disparity in discretionary adjustments could be justifiable. For example, minorities may have been less likely to initiate negotiations on rate. Alternatively, this disparity in discretionary adjustments could be a result of discrimination. If we include the DT variable, the regression model would accurately estimate that the rate was higher for

[19] See footnote 16 for an explanation for why we did not include a 0/1 flag for FICO scores above 720.

[20] The error term and its properties are the foundation of regression analysis. The details of error terms for regression analyses are extensive and complex, and therefore beyond the scope of this report. We encourage interested readers to explore any introductory text on multivariate regression analysis, such as, "Introduction to Econometrics" by G.S. Maddala and Kajal Lahiri, for more details.

applicants who were charged a higher discretionary adjustment. In addition, the estimated coefficient on the minority flag would be close to 0, since the lender did not explicitly consider minority status when setting rate (see equation A1). If the discretionary adjustment is justifiable, then it is perfectly acceptable to include DT as a control variable in the regression model. However, if the discretionary adjustment is discriminatory, including DT would mask that discrimination. Since it is typically unclear initially what is driving disparities in these discretionary adjustments, the appropriate approach is to exclude DT to start. By excluding the DT variable, the coefficient estimate on the minority flag will capture any differences in loan officer discretion across minority status, which is 20 bps on average here. A follow-up review would then assess whether there is justification for the differences in how the lender applied discretion when setting rate.[21]

Table B1 presents the OLS regression results for the pricing analysis. Since we are using simulated data, per equation A1, the estimated coefficients in the regression should be 7 for the intercept, 0.2 for fico_660, 0.1 for fico_660_720, 0.1 for MH, and -0.01 for income. Similarly, given the way we incorporated loan officer discretion, the coefficient for the minority flag should be 0.20. Each of the estimated coefficients in Table B1 are close to these values, but not exactly the same. The small differences occur because the decision-making process includes a random component and the sample size is relatively small. If in the decision-making process the random component was eliminated, the variance of the random component was set to a really small value, or the sample size was really large, the coefficient estimates from the OLS regression would be approximately the same as in equation A1.

[21] It should be noted that the exact same arguments made here for the DT variable also apply to the FICO score, income, and property type variables as well. The reason we treat those three variables differently is because the lender’s choices are more likely to influence DT, and therefore the risk of potential discrimination is higher.

Table B1: Regression Results for the 1,000 Application Pricing Sample

```
Dependent Variable: Rate

                               Sum of           Mean
Source              DF        Squares         Square        F Value       Pr > F

Model                5         89.6055        17.9211         75.99       <.0001
Error              994        234.4316         0.2359
Corrected Total    999        324.0370

Root MSE            0.4856       R-Square      0.2765
Dependent Mean      6.4137       Adj R-Sq      0.2729
Coef Var            7.5720

                               Parameter     Standard
      Variable        DF       Estimate       Error       t Value    Pr > |t|

      Intercept        1        6.9833        0.0731       95.55      <.0001
      fico_660         1        0.2104        0.0447        4.70      <.0001
      fico_660_720     1        0.0704        0.0376        1.87      0.0615
      mh               1        0.1469        0.0453        3.24      0.0012
      income           1       -0.0098        0.0008      -11.74      <.0001
      minority         1        0.2160        0.0425        5.08      <.0001
```

Table B2: Regression Results for the 1,000 Application Pricing Sample with Errors in Each Rate

```
Dependent Variable: Rate

                               Sum of           Mean
Source              DF        Squares         Square        F Value       Pr > F

Model                5         88.7961        17.9592         71.63       <.0001
Error              994        246.4357         0.2479
Corrected Total    999        335.2317

Root MSE            0.4979       R-Square      0.2649
Dependent Mean      6.6676       Adj R-Sq      0.2612
Coef Var            7.4678

                               Parameter     Standard
      Variable        DF       Estimate       Error       t Value    Pr > |t|

      Intercept        1        7.2395        0.0749       96.61      <.0001
      fico_660         1        0.2093        0.0459        4.56      <.0001
      fico_660_720     1        0.0690        0.0385        1.79      0.0739
      mh               1        0.1426        0.0465        3.07      0.0022
      income           1       -0.0098        0.0009      -11.46      <.0001
      minority         1        0.2131        0.0436        4.89      <.0001
```