1-28-2011

Financial Crisis Inquiry Commission Hedge Fund Market Risk Survey: Methodology Report

Timothy M. Mulcahy

Johannes Fernandes-Huessy

Kenneth Kuk

Financial Crisis Inquiry Commission Hedge Fund Market Risk Survey: *Methodology Report*

TIMOTHY M. MULCAHY, JOHANNES FERNANDES-HUESSY, KENNETH KUK

PRESENTED TO:
Financial Crisis Inquiry Commission

PRESENTED BY:
NORC at the
University of Chicago
4350 East West Highway, Suite 800
Bethesda, MD 20814
(301) 634-9300

JANUARY 28, 2011

TABLE OF CONTENTS

Background

The Financial Crisis Inquiry Commission (FCIC) was established by the U.S. Congress on July 15, 2009 to study the causes of the financial crisis and investigate fraud and abuse in the financial sector (a full list of the Commission's charge is available on FCIC's website).[1] The Commission reported its findings January 26, 2011 to the Congress the President, and the American people.[2]

One major area of inquiry for the Commission focused on the role that hedge funds may have played in the crisis. To gauge this, the FCIC sought information on hedge funds' market risks before, during, and after the financial crisis (e.g. investment portfolio profiles, investment strategies, risk exposure, involvement in derivative investments, etc.).

The Commission contracted with NORC at the University of Chicago (NORC)[3] to collect, prepare, aggregate, and de-identify survey microdata from participating hedge funds to protect respondent confidentiality and create a public use data file.[4] All microdata were processed and analyzed by a team of disclosure analysis specialists[5] and maintained in the NORC Data Enclave, a secure data warehouse that houses sensitive microdata from various federal statistical agencies and other confidential data foundations and/or academic data producers.[6]

This methodology report discusses the processes employed in aggregating, cleaning, recoding, analyzing, harmonizing, and de-identifying the survey data, and provides detail across each stage of the process. The final product of this effort will be a completely de-identified, aggregate public use dataset that has been diagnosed and treated for consistency and quality.

Survey Development & Implementation

FCIC staff designed and distributed the market risk survey, and responded to all survey specific inquiries made by respondents. FCIC investigators also were responsible for encouraging firms

[1] Financial Crisis Inquiry Commission. 2009. http://www.fcic.gov/about/. Retrieved on December 21, 2010.

[2] http://fcic.gov/report.

[3] Founded in 1941, NORC at the University of Chicago is a 501c(3) nonprofit organization that conducts public interest research. Although NORC's national studies are its most well-known, our projects range from local to regional and international. NORC has headquarters on the University of Chicago's campus, and offices in Chicago's downtown Loop, Washington, DC, Bethesda, Maryland, and Berkeley, California, as well as a field staff that operates nationwide. NORC's clients include government agencies, educational institutions, foundations, other nonprofit organizations, and private corporations.

[4] Before working with the microdata Identifiable information such as establishment names was excluded from the surveys by NORC's principal confidentiality officer to ensure that no such information would be known by NORC and/or FCIC analysts.

[5] The authors of this report want to a acknowledge senior NORC statisticians Fritz Scheuren and Michael Yang for their expert guidance and feedback throughout this project.

[6] NORC at the University of Chicago. 2010. http://www.norc.org/dataenclave/ Retrieved on December 21, 2010.

to participate.[7] The Commission's master sample list frame included 243 firms; however, we do not provide a definitive response rate due to a number of issues, first and foremost to protect respondent confidentiality. In addition, some firms changed ownership or dissolved during the reference period. Further complicating matters, survey respondents were not sampled consistently. Respondents included privately owned sole proprietorships, funds managed by a larger financial institution, and firms that managed multiple "children" funds. Given these challenges, we compiled the dataset at the respondent level, meaning that each observation is equal to one survey respondent. Although we cannot provide a precise response rate, we are confident that it is greater than 70 percent.

NORC implemented the data collection process and instructed respondents on encrypting their data and submitting completed surveys via a secure file transfer protocol. NORC analysts Survey extracted the response data to form a master dataset using an algorithm custom developed by NORC analysts using a set of custom-developed Visual Basic for Applications (VBA) algorithms.

Respondents provided data that formed two different types of data tables: (1) counterparty-, collateral- and product-related tables (henceforth referred to as "descriptive tables")[8] and (2) position and value tables (hereinafter "numeric tables")[9]. For descriptive tables, respondents were asked to complete descriptive information and the corresponding dollar amount (e.g. value of contract, collateral value, etc). For numeric tables, respondents were asked to complete net or notional values of a particular financial product or position. Thus, whereas data in descriptive tables contained descriptive information, those in numeric tables were strictly numeric. The strategies we employed to handle these two tables will be discussed in subsequent sections of this report.

The descriptive tables appear on the Repo, Commercial Paper, and Derivative Exposure sheets. In each case, respondents were asked to identify counterparties to particular transactions, a dollar outstanding or notional amount, the type of collateral used in the case of repurchase agreements, and the type of product for derivative transactions. Repurchase agreements and commercial paper tables also included a "haircut" (margin) value. In order to present these data in the most analytically useful manner, we created separate tables that show the mean dollar outstanding/notional amount (and "haircut" where applicable) by counterparty within each quartile. For repurchase agreements, we also show the mean dollar outstanding and "haircut" by type of collateral; and for derivative transactions, we show the mean notional value by product type.

[7] The FCIC provided hedge funds three options: (1) submit their raw microdata directly to the Commission; submit their microdata to NORC where the microdata would be de-identified and aggregated; or face a possible subpoena for noncompliance.

[8] These tables contain information on the counterparties involved in REPO contracts and commercial papers. Information includes value, haircut, discount, and collateral description. These tables also include information on net notional derivative exposure by counterparty with the type of financial product described.

[9] All tables not included in category (1) fall under category (2).

In so doing, we categorized all collateral and product types and reported each category that did not appear frequently enough to be reported in an "Other" category. Arranging the tables in this manner assists the analyst in seeing changes over time among similar funds in the same way we do in the master data set. For instance, we can see how a particular size class of a fund increased or decreased their transactions with a particular institution over time, or how that group of funds increased or decreased their demand for a particular type of derivative.

Challenges Related to Data Mining, Recoding, Cleaning, & Harmonization

Lack of an Item Validation Component

After compiling the dataset, NORC analysts ran multiple diagnostic tests to check for data consistency and identified a number of potential problems with the raw microdata. First, the survey lacked an item validation component to prevent respondents from entering irrelevant, invalid, or out of scope values. Respondents were able to delete columns and rows, lock down workbooks and worksheets with passwords, and rename worksheets, which led to a number of invalid values in both the descriptive and numeric tables. These values included different combinations of symbols and characters indicating missing or "not applicable" values, as well as adding units (i.e., "$", "*", etc.) to numeric values. We developed an algorithm to detect and correct the variable type. In the numeric variables, we coded all "not applicable" values[10] as missing. In some instances, respondents reported dollar amounts in percentages (with "%" sign) or vice versa (with "$" sign). Rather than attempting to interpret these inconsistent inputs, we coded them as missing values. In other instances, respondents reported percentage figures (e.g., "haircut", discount rate, etc.) with the symbol "%". When these numbers were extracted from the surveys, MS Excel (the format selected by the Commission for the survey) automatically divided the numbers by 100. To maintain consistency, we reviewed all of the surveys, and multiplied all affected values by 100.

Recoding String Variables

It is important to note that string variables that contained information on counterparty and brokerage names, as well as product specifications in the descriptive tables, required significant recoding efforts. Since survey respondents could not select from a drop-down list, at times respondents (referring to the same counterparty or brokerage firm) provided slightly different names or spellings. Consequently, we recoded and standardized all counterparty names. We also recoded the collaterals and products into categories. These recoding processes were essential for micro-aggregation and de-identification, and will be discussed more fully in subsequent sections of this report.

Numeric Values in Descriptive & Numeric Tables

[10] "Not applicable" appeared in various forms, such as "Not Applicable", "N/A", "na", "/", "-", and "*".

Some numeric values in both the descriptive and numeric tables presented inconsistencies that required logical editing. First, we made logical edits to some firms' responses to assets under management (AUM) – not only because it was the only measure of firm size, but also because it was critical to determining whether other variables were valid. While survey respondents were asked to provide their high, average, low, and current AUM within the period of reference in billions, about 20 percent reported values in erroneous orders of magnitude. Since AUM information was extremely important to our analysis (especially in later stages), we concluded that logical data imputation was the most appropriate solution.

As of January 2009, the largest U.S. hedge fund had an AUM of about $38.6 billion.[11] No respondent identified their high AUM as being over $60 billion and under $1 trillion. Thus we can conservatively assume that no U.S. hedge fund had an AUM of over $60 billion at any time during the reference period.[12] Therefore we considered all values outside of this range as "misreported" and adjusted them by dividing them by 1000. We repeated the process until there were no inappropriately large AUM figures. We did not see these types of problems in the lower tail of the distribution and thus left those values as they were.

Another issue stemmed from the fact that the surveys sent to respondents were pre-populated with dummy values. Dummy values, pre-filled in the surveys to assist respondents in reporting appropriate values in each field, were whole numbers ranging from "0" to "5", and differed across tables. What's more, a number of the dummy values were interspersed with valid values, at times making it difficult to interpret missing values and distinguish valid from dummy values.

Process for Dealing with Dummy Values

For comparison purposes, we extracted dummy values from the original survey and appended them to the compiled dataset. Next, we developed an algorithm that flagged all values that were equal to the dummy value in each variable. While dummy values that appeared in blocks (e.g., an entire table filled with dummies) were easily identified, treating isolated values was comparatively more complicated because the dummy value itself could sometimes be valid. Consider the following example:[13]

Period 1	Period 2	Period 3	Period 4	Period 5	Period 6
0.79	2	1	2	2	1.8

[11] Bloomberg. March 4, 2009. http://www.bloomberg.com/apps/news?pid=newsarchive&sid=aQ1Lote4Iwfo. Retrieved on December 22, 2010.

[12] January 1st 2007 to the point at which they reported

[13] Not real data.

Assume these numbers represent the values of a firm's outstanding debt from period 1 to period 6 in millions. In this example, "2" is the dummy value for the entire table. This means that if the entire table is filled with "2," it is likely that the respondent has not reported any value. However, "2" could also be interpreted as a valid value. While having an outstanding debt of $2 million is plausible, we cannot definitively distinguish valid from dummy values, i.e., "quasi" dummy values.

While "quasi"-dummy variables that appear in blocks can be difficult to interpret at times, there are also instances where a clear pattern can be observed. Consider the following example:[14]

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
16	1	1	1	12	1	1	1	7.8	1	1	1	14

Assume these numbers represent the values of a firm's outstanding debt from period 1 to period 13 in millions, collected quarterly. In this example, "1" is the dummy value for the entire table. In contrast to the previous example, we can observe a clear pattern, i.e., the respondent only reports values in the first quarter of a given year. While it is still plausible to have an outstanding debt of $1 million, having $1 million of debt only in Quarters 2, 3, and 4 is less likely. A more plausible explanation would be that the firm only has annual data for outstanding debt, and hence reports only these values in Quarter 1. In these instances, we are more likely to interpret "quasi"-dummy variables as missing values. To deal with this issue, we developed a general decision rule that block quasi-dummy variables are to be interpreted as missing values only if they appear as three or more in a row (as shown in the above example). In other words, "quasi"-dummy values are retained if they demonstrate a singleton or a doubleton pattern.[15]

Interpreting Missing Values and "0"'s

Interpreting missing values correctly is crucial in that it affects the number of observation by which the quartile aggregates are divided when quartile averages (produced as part of the de-identification process) are produced. If respondents removed the dummy value from a field, and did not replace it with another value; the value would be reported as missing. On the other hand, if respondents replaced a non-"0" dummy value with a "0", the value would be reported as "0". The practical differences between the two are particularly important in the context of this survey. Consider a scenario wherein "Respondent A" failed to provide information on the outstanding value of all REPO contracts in period 1. That is, "Respondent A" did not enter anything in the field after removing a non-"0" dummy value. This could be interpreted as a missing value, but it also could be interpreted as a"0". For example, if a fund was not involved in REPO contracts in period 1, the outstanding value would essentially be "0".

[14] Not real data.

[15] Note that "singleton" and "doubleton" are general descriptions of isolated valid values in a table, and should not be confused with their mathematical definition in set theory.

While equating imputed dummy values to "0" is a convenient solution, there are important implications. In the scenario described above wherein a respondent only reported first quarter figures but retained the dummy values in fields for other quarters, it is crucial to distinguish clearly between missing values and "0"'s. Essentially, while one could interpret missing values in Q2, Q3, and Q4 for example because a respondent is only able to retrieve annual data (and for data that had followed similar patterns, e.g. reporting monthly data in weekly data tables), this assumption would not hold true if the imputed dummy values were translated into "0"'s. And since missing values and "0"'s were interchangeable for most variables in most observations, all "0"'s, invalid values, a decision rule was codified that imputed dummy values were to be excluded when quartile averages were computed.

Correcting Negative Values and Erroneous Orders of Magnitude

Some respondents filled certain strictly positive variables with negative values, such as the fund's external debt. A number of respondents reported values from an accounting perspective, resulting in positive values for asset inflow and negative values for outflow. For consistency purposes, we imputed all negative numbers in variables related to debt, redemption payments[16], and outstanding value of reverse repo contracts[17] by taking the absolute value of these numbers. Moreover, some funds reported long positions in positives and short positions in negatives; whereas others reported both in their absolute values. Again for consistency purposes, data were imputed such that long and short positions were reported in absolute values in the final dataset. Negative values found in other fields were considered legitimate and were therefore retained.

At various points in the survey, respondents were asked to provide information in different orders of magnitude, leading to some misreported numbers in erroneous orders of magnitude. For instance, while respondents were asked to identify their AUM in billions, they were instructed to enter other responses in millions. Some respondents, however, reported numbers in the actual number or in thousands. This issue was fairly easy to uncover as some firms reported 10-digit numbers in the fields rather than in millions, meaning that a particular asset class was worth millions of times more than their AUM which is impossible. NORC analysts developed a process to handle extreme outliers, which we describe more fully below.

Creating Outlier Detection Models to Deal with Measurement Error

[16] Redemptions are payback to investors by request, which usually takes a certain time to process given hedge funds' low liquidity. Some firms reported them in negatives.

[17] Repurchase contracts, also known as "Repo," are contracts between two parties, one of which agrees to sell securities to the other party with the premise of buying them back within a designated period. Essentially, the original buyer of Repo is the lender who borrows money from the original seller using the securities as collateral. In a reverse repurchase contract, the two parties swap. The initiator agrees to buy securities from another party with the premise of selling them back at a later date. From the original buyer's perspective, it is like lending money to another party. In the context of data reporting, reverse repurchase contracts might be reported in negative because they are the exact opposite of repurchase contracts.

Measurement error is not new to the field of survey methodology. In fact, over the years statisticians have developed different models to identify outliers. Unfortunately, most of these models (e.g., Grubbs' test[18] and the Pierce's test[19]) are based on distribution assumptions that assume normality. For the hedge fund market risk survey, outlier detection was more challenging because the number of observations for most variables was very low, which made it very difficult to model the probability density functions. And since financial data are often significantly skewed, applying outlier detection models that require a normal distribution would only exacerbate the problem. We developed two different models to address this issue.

The first outlier detection model we created was a ratio sensitivity analysis. Essentially, we compared each number to a set of other numbers within the same observation and determined a reasonable ratio threshold beyond which a value should not practically exceed. AUM was considered the universal comparison variable because any net position or contract value should be a certain percentage of the firms' total assets. Even derivative exposure and cash flow should not exceed AUM by a certain multiple. In other words, we assumed a correlation between AUM and other variables.

Next we determined the correlation coefficient for the ratio analysis. After computing the AUM-variable ratio of the entire data matrix, we developed an algorithm that flagged data points if the ratio violated a certain set threshold. The algorithm then tested a number of thresholds and examined the number of flagged data at different thresholds, eventually indentifying the optimal ratio that minimized false flags, while also balancing the effects of false positives and negatives.

When working with heavily skewed financial data, such as the Market Risk Hedge Fund Survey, outliers may be valid data points and thus should not be discarded. On the other hand, when an outlier is determined to be invalid, it should be excluded. However, as noted previously, due to the small sample size of this effort excluding outliers could cause significant analytic distortion to the quartile averages. More importantly, decisions to exclude may necessitate data suppression in the de-identification process because if the number of respondents is too small, reporting quartile averages can significantly increase disclosure risk. Therefore we codified a decision rule to handle outliers by imputation, which essentially involved changing outlying values using donor microdata.

Sensitivity Analysis and Imputation

Before imputing the data, we developed a plan to re-shape extreme outliers into reasonable values without adding excessive arbitrariness and noise to the data. Critical to this process, we assumed that respondents unintentionally misreported numbers at incorrect magnitudes, i.e.,

[18] Grubbs, Frank E. 1969. Procedures for Detecting Outlying Observations in Samples. *Technometrics*, 11(1): 1-21.

[19] Pierce, Benjamin. 1852. Criterion for the Rejection of Doubtful Observations. *The Astronomical Journal*, 2(21): 161-163.

that all misreported numbers were merely the true values reported at different orders of magnitude.[20]

In summary, our first model assumed that extreme outliers resulted from respondents reporting values at incorrect orders of magnitude. Therefore, we repeatedly imputed the values by dividing the flagged value by a constant factor until the AUM-variable ratio fell below a reasonable threshold. If respondents reported numbers at incorrect orders of magnitude, they could only be 1,000 or 1,000,000 times larger than the true values. Therefore, after identifying the optimal AUM-variable ratio, we divided the flagged values by 1000. Then we re-ran the algorithm twice to identify and re-flag the values that still remained outside of the reasonable zone.

To recap, the optimal AUM-variable ratio resulted from a sensitivity analysis in which the algorithm computed the data matrix at different ratios to identify the one that would be conservative enough to create the least statistical noise. In the end, we employed an extremely conservative threshold because erroneous data imputation would not only create noise, but also would create data inconsistencies.

The sensitivity analysis yielded a ratio of 1:80 for the notional value of derivative exposure (of certain product) and a ratio of 1:10 for all other variables would be, on the one hand, conservative enough to minimize statistical noise caused by data imputation; while on the other hand, sufficient enough to deal with the maximum number of invalid outlying data points. An AUM-variable ratio of 1:10 meant that only when the value of the particular contract or a position was 10 times of the observations' AUM would we consider the value to be invalid. Given the hundreds of different types of financial products that each fund issued, we are confident that this definition is extremely conservative. While it would be more difficult to affirm whether a 1:80 AUM-exposure ratio is conservative enough, with an average debt-equity ratio of less than 1:3, we assumed that respondent most likely could not be exposed to a particular derivative class that was worth 80 times its total assets.

Developing a Secondary Outlier Detection Model

The first outlier detection model, using 1:10 and 1:80 ratios, yielded 2,409 flagged values in the first run, which was about 0.5 percent of the total data points (including missing values). After the first round of division, 943 flags remained. After the second round, only 25 remained. Lacking a theoretical backup for detecting outliers for a third round, we were confident with a residual rate of about 1 percent and thus proceeded to the second outlier detection model. The need for a second outlier detection model was evident in that some irregularities still remained in the dataset even after computing the quartile averages. Surprisingly, in a number of instances, we observed significantly larger numbers in lower quartiles.

[20] We are confident in these assumptions, and it is important to note that these assumptions are much weaker than assuming any certain form of probability density function with such a small sample size.

Although only 25 flags remained after processing the dataset with the sensitivity analysis model, we developed a complementary algorithm to detect outliers in a slightly different manner as a way to both cross check and validate the model. The second model was a conditional summary statistics analysis. The algorithm was developed such that it loops through every observation in each variable, removing one observation at a time, and comparing the conditional mean and standard deviation[21] to the original summary statistics in each quartile. Essentially, we created two separate sets of derivative data matrices (at the individual firm level) that enumerate the means and standard deviations of each variable to determine whether each non-zero observation should be removed from the sample. Next, we compared these numbers to the original means and standard deviations of each variable to detect whether some observations were creating significant distortions.

For instance, if the mean without "Observation 1" for "Variable A" is only 10 percent of the original quartile average, then "Observation 1" alone contributes to 90 percent of the quartile aggregate, a fairly clear indication that "Observation 1" appears to an outlier. In other words, the smaller the percentage, the less likely the observation is to be valid. Of course, 10 percent is only one example. Thus we included a sensitivity analysis component that flagged values on the derivative data matrices at different cutoff points in increments of 5 percentage points and then compared the flagged values yielded from these different cutoff points. The algorithm eventually showed that 25 percent was the optimal threshold. The number of flagged values declined significantly as the cutoff decreased from 30 percent to 25 percent, and remained flat thereafter.

Methodologically, the conditional standard deviation analysis procedure was very similar. The algorithm created data matrices filled with conditional standard deviations of each variable, as each observation was temporarily removed from the original dataset. Then, it compared these derivative data matrices to the original dataset to measure the effect caused by removing one observation on the overall distribution. If the standard deviation without "Observation 1" for variable A is only 25 percent of the standard deviation from the original dataset, the inclusion of "Observation 1" created a significant increase in variance. Then, the algorithm would flag values using different percentage thresholds. The sensitivity analysis demonstrated that a 35 percent threshold was the most reasonable cutoff for the standard deviation.

Because the second outlier detection model defined outliers comparatively more leniently than the first model, we chose not to impute the flagged data points. Instead, we reported the variables that were flagged by the program and included the standard deviation of each variable in the final dataset, thereby allowing data analysts to determine whether a variable

[21] Conditional mean and conditional standard deviation here referred to the mean and standard deviation of a variable when one observation is taken away from the sample. Consider the conditional mean for variable A with the absence of observation 1 to be $\hat{u}(A1) = E(u(A1) \mid x(1) \notin \{A\})$ where x(1) is the value of observation 1 and u(A1) is the mean of A without observation 1 (similarly for standard deviation). The conditional analysis algorithm eventually creates a new data matrix that can be compared to the original mean and standard deviation to measure the stake each observation contributes to the summary statistics of a variable.

should be used or removed to answer researcher specific questions (see Appendix 1 for a full list of variables that were flagged by the second model).[22]

After we conducted the first round of outlier treatment, only 25 flagged values remained. However, in the second round outlier detection, we identified 419 potentially invalid data points across 29 observations. Much of the difference is due to the fact that the second outlier detection model employed a more lenient definition of outliers. For the same reason, we only treated outliers by data imputation in the first model, but retained all flagged values in the second. In other words, we were able to minimize data loss and maximize data utility by using a conservative detect-and-impute model before introducing a more aggressive detection-only model. In so doing, we are confident that most of the outliers were treated without adding too much noise to the individual level dataset.

Micro-aggregation and De-identification

The microdata collected as part of the Hedge Fund Market Risk Survey are extremely sensitive because individual level data may disclose confidential financial information about funds, or even individuals. In general terms, disclosure can be classified into two types. While data providers are obligated to prevent disclosure of sensitive information from the dataset, they also have to consider the scenario of an intrusion. Essentially, disclosure risk is concerned with the possibility that the intruder will be able to determine a correct link between a microdata record and a known unit.[23] With the amount of information available to the public today, disclosure control is increasingly difficult.

In the main, data can be treated or de-identified to reduce disclosure risk. The obvious tradeoff is that there is a clear tension between increasing data confidentiality and data analytic utility. Once a dataset is treated or de-identified, its analytic value inevitably decreases due to loss of information.[24] The challenge is to identify an optimal point that maximizes data analytic utility and confidentiality. The goal of de-identification treatment is to restrict the amount of information an individual can glean from a dataset, while still providing analytic utility. While disclosure risk cannot be eliminated without limiting data analytic utility, it is often difficult to determine the most appropriate level that balances disclosure control and analytic utility. Indeed it varies depending on a number of factors, including the number of people who have access to the information the agreement between the data provider and users, the data producers' goals and objectives, intended audience, and risk tolerance.

[22] Additional detail regarding the components of the de-identified aggregate dataset (i.e. standard deviation of variable) will be discussed in depth below in the de-identification section.

[23] Skinner, Chris. 2009. Statistical Disclosure Control for Survey Data. S3RI Methodology *Working Paper M09/03, Southhampton Statistical Sciences Research Institute*, 21pp.

[24] Singh, Avinash. 2009. Maintaining Analytic Utility while Protecting Confidentiality of Survey and Non-survey Data. *Journal of Privacy and Confidentiality*, 1(2): 155-182.

De-identification treatments generally fall under two categories: synthetic treatments and non-synthetic treatments. Synthetic treatments involve creating synthetic data derived from the original data through multiple imputations, stochastic perturbation, and other model-based methods.[25] For the purpose of this document, the population size was too small and response rates were often too low that it was impossible to describe the distribution of the variables with perfect accuracy. Moreover, performing multiple rounds of data imputation as part of the data cleaning process added a degree of noise to the dataset.

Although we examined whether employing synthetic treatment was a viable option, in the end we determined that non-synthetic treatments were most appropriate for this particular dataset. There are myriad ways to de-identify microdata, e.g., global recoding, perturbation, local suppression, etc.[26] Global recoding involves reducing the amount of information provided in a variable. For instance, instead of reporting the county in which a respondent resides, the variable can be recoded to indicate a larger region (e.g., state). This method was not suitable for our purpose. Perturbation is essentially a partial synthetic treatment. Instead of creating an entirely new synthetic dataset, perturbation treats a subset of the dataset through random data swapping, where one record is replaced with another record from the dataset. This method, however, was not suitable for our purposes.

The de-identification model selected for this effort is a combination of micro-aggregation and local suppression. First, we treated the numeric tables. Rather than releasing individual level microdata, we combined groups of individuals into aggregate records.[27] [28] Although this lowers the overall data analytic utility, it does not require statistical modeling and thus does not involve distribution estimation.

Micro-aggregation was the optimal solution for de-identifying the market risk survey data. After weighing the benefits and risks of different levels of aggregation, we aggregated records by quartile, such that the aggregated dataset contained four observations. These observations would essentially be the averages of each quartile of the individual level microdata by AUM. For variables with very few responses, we introduced a second layer of confidentiality protection – local suppression, which entailed removing or suppressing data from the dataset. This procedure was employed to prevent data analysts from re-identifying individual records from the aggregates, given that they had the averages and the total number of observations.[29]

---

[25] Skinner, Chris. 2009.

[26] Sergio I. Prada et. al. 2010. Avoiding Disclosure: A Review of Policy, Statistical, and Access Issues. *Internal working policy paper jointly written by IMPAQ International and NORC at the University of Chicago*.

[27] Feige, Edgar L. and Harold W. Watts. 2005. Protection of Privacy Through Microaggregation. *Econometrics, EconWPA # 0502001*, 12pp. http://129.3.20.41/eps/em/papers/0502/0502001.pdf. Retrieved on January 11, 2011.

[28] Hansen, Stephen L. and Sumitra Mukherjee. 2003. *IEEE Transaction on Knowledge and Data Engineering,* 15(4): 1043-1044.

[29] We suppressed all quartile averages if the number of observations was less than three. We also added the standard deviation of each variable to the dataset to provide a better picture of the variables' distribution. Values were suppressed if the number of observations was less than three.

Conclusion

While the data gathered as a result of this survey are unique in their scope and subject matter, and we achieved a response rate in excess of 70 percent, analysts should be aware of a number of limitations in using the public use data file, e.g.:

- Observation level data may not consistently represent funds or parent organizations
- Sections of the surveys often were completed by different employees from the respondent institution. This may have caused some of the observed inconsistencies in responses in different tables.
- Some values have been suppressed because they were reported in an incorrect format (e.g. percentages reported in a dollar outstanding field).
- Data has been imputed to correct erroneous reported orders of magnitude.
- Some values may be dummy values from the original survey instrument.
- Zero ("0") values have been treated as missing data and were not counted in constructing quartile averages.
- Where negative values were reported, they are averaged with positive values. These variables should be understood as representing the relative long or short position of the quartile in the asset class, not the relative amount of activity in that asset class
- During the height of the crisis, some funds did not honor redemption requests within the normal timeframe. This led to some cases where funds would honor redemption requests from past periods, creating the impression that they were paying out redemptions that were larger than the requests.