# Duplicate FHA Single-Family Mortgage Records and Related Problems*

Thomas N. Herzog, PhD, ASA
Office of Evaluation
U.S. Department of Housing and Urban Development

*This work represents the personal views of the author rather than those of HUD and/or the government of the United States. A number of the examples used in this work have been modified to protect individual privacy and/or to simplify the exposition. In particular, none of the borrower names and addresses is intended to represent those of actual FHA-insured mortgages.

**Abstract**

The Federal Housing Administration (FHA) insures mortgages against the risk of foreclosure. Since its inception in 1934, FHA has insured over 37 million mortgages on single-family homes. This requires FHA to store data on a large number of mortgages. Because of the large number of mergers/acquisitions among mortgage lenders, it is sometimes difficult for the surviving lenders to maintain accurate databases. As a consequence, such lenders do not always transmit accurate/timely data to FHA on the termination of FHA-insured single-family mortgages that they are servicing. This, in turn, means that FHA's database has many mortgages listed as "active" that have in fact been terminated. In addition, FHA made a decision many years ago to omit the property address of insured mortgages from its databases on single-family mortgages because of the high cost of computer storage at that time. Although this decision was later reversed as such costs declined, even in the year 2008, FHA had a substantial number of "active" mortgage records without a corresponding property address.

In order to improve the quality of FHA's databases in these two aspects, we have applied a number of internal consistency checks and record linkage techniques. The first approach was to use a variety of internal consistency checks to identify "active" mortgage records whose underlying mortgages had in fact terminated. The second approach involved matching FHA records with corresponding records of the Government National Mortgage Association (GNMA). This second approach allowed us to (1) obtain property addresses from the GNMA database and add them to the FHA database as well as (2) identify additional "active" FHA mortgage records whose underlying mortgages had terminated.

We have employed a variety of internal consistency checks to identify and subsequently remove "duplicate" mortgage records from this database.

## 1. INTRODUCTION

### 1.1 Background

The Federal Housing Administration (FHA), an agency within the U.S. Department of Housing and Urban Development (HUD), insures mortgages against the risk that the borrower, for whatever reason, will be unable to continue making payments on his/her mortgage. The FHA's mortgage

*This work represents the personal views of the author rather than those of HUD and/or the government of the United States. A number of the examples used in this work have been modified to protect individual privacy and/or to simplify the exposition. In particular, none of the borrower names and addresses is intended to represent those of actual FHA-insured mortgages.

guarantee insurance programs are partitioned into four separate insurance funds. FHA's Mutual Mortgage Insurance Fund (MMIF) is its largest fund. The MMIF insures mortgages on single-family homes consisting primarily of single-family detached houses and townhouses. FHA reported that as of June 30, 2009, the MMIF had 4.9 million mortgages insured with an aggregate face amount of $604 billion. FHA's General Insurance Fund (GIF) also insures mortgages on single-family homes. In addition, the GIF insures loans on individual condominium units as well as on apartment buildings, nursing homes, hospitals, and mobile homes. The Cooperative Management Housing Insurance Fund (CMHIF) insures mortgages on cooperative apartment buildings. The Special Risk Insurance Fund (SRIF) insures mortgages on single-family homes, excluding condominiums, and apartment buildings.

Since its inception in 1934, the FHA has insured over 37 million mortgages on single-family homes. The bulk of these mortgages have been insured under FHA's Mutual Mortgage Insurance Fund (MMIF). The mutuality feature of this fund means that dividends (also known as distributive shares) may be paid to certain borrowers when they terminate their insurance. The amount of the dividend depends on the mortgage amount, the year the mortgage began amortizing, and the amortization plan of the mortgage. On September 1, 1983, FHA instituted a one-time premium collection system whereby the entire premium is paid in advance, and unearned premium refunds are paid to those who successfully terminate their loans prior to maturity.

## 1.2 Purpose

When a borrower refinances or prepays an FHA-insured single-family mortgage, the lender servicing that mortgage is supposed to notify FHA, and FHA is supposed to make the appropriate changes to its databases. Unfortunately, this process does not always work as intended. As a consequence, FHA has hundreds of thousands of mortgage records in its single-family data warehouse with a termination status of "active" when in fact the underlying mortgage has been refinanced, paid in full, or terminated for some other reason. Because (1) FHA only insures "first" mortgages, i.e., those with a primary lien on the underlying property, and (2) no condominium units[1] are supposed to be insured under the MMIF, it follows that there should be at most one "active" MMIF-insured mortgage per property address.

[1] For purposes of this paper, it is useful to assume that FHA-insured condominium units have never been insured under FHA's MMIF. In actuality, this was only true through September 30 2008, after which newly-originated mortgages on condominium units were insured under the MMIF.

In addition, a relatively small number of additional mortgages have been entered into FHA's single-family data warehouse under two or more identification numbers,[2] usually with only slight differences between the identification numbers and few, if any, differences in the case data. When such records are displayed together, it is usually apparent that both represent the same mortgage; however, because the identification numbers are different, the data warehouse treats both records as unique, individual mortgages. Our intent here is to describe some of the data problems we have investigated as well as the efforts we have made to improve the accuracy of the affected records. The consequences of these data problems will be discussed in Section 3.4.

### 1.3 FHA Single-family Data Warehouse

Currently, FHA's primary single-family database is known as the FHA single-family data warehouse. While the vast majority of its records do not have serious data problems, we have identified some with severe problems. Because so many cases are involved here, an error rate of only one or two percent could result in hundreds of thousands of incorrect records.

### 1.4 Outline of Paper

In the next section we discuss the concept of an FHA case number in order to facilitate the rest of the discussion of the paper. In Section 3, we discuss the problem of the duplicate mortgage records. In Section 4, we discuss the problem of the mortgage records with incorrect termination statuses. In both Sections 3 and 4, we begin by illustrating the nature of the problem, we then describe the procedures we used to identify and correct these problems, and finally we list some of the consequences of these errors. In Section 5, we summarize a scheme used to obtain property addresses on FHA-insured single-family mortgages from a database of mortgages maintained by the Government National Mortgage Association (GNMA).Finally, in Section 6, we describe schemes for estimating the number of mortgages records having the problems considered here.

## 2. FHA CASE NUMBERS ON SINGLE-FAMILY MORTGAGES

Beginning in January 1962, all FHA case numbers consisted of a 3-digit State and Office code prefix (SSO-) followed by a dash, a six digit serial number, and a final check digit (SSO-DDDDDDC). Within each HUD field office, the serial numbers are assigned chronologically. The use of the check-digit was

[2] These identification numbers are known as "FHA case numbers" and are discussed in detail in Section 2.

intended to improve the accuracy of the FHA case number; unfortunately, it was easy to circumvent the protective function of the check digit by appending an "X" to the end of a questionable FHA case number. This caused FHA's previous single-family computer system to calculate and insert the appropriate check digit for any number entered, thus passing over the incorrect or missing check digit. This allowed the processing of the invalid FHA case to continue and even gave its case number an aura of legitimacy because it would then have a "correct" check digit attached from that point on. Fortunately, this problem was eliminated with the implementation of HUD's Computerized Homes Underwriting Management System[3] (CHUMS) in the mid-1980s. The CHUMS assigns FHA case numbers automatically, leaving no chance for manual error.

# 3. THE PROBLEM WITH THE DUPLICATE MORTGAGE RECORDS

## 3.1 Statement of Problem

The first problem we considered was the existence of thousands of mortgage records entered into HUD's single-family data warehouse under two or more FHA case numbers, usually with only slight differences between the FHA case numbers and few, if any, differences in the case data.

[3] CHUMS is still in use today.

## 3.2 Procedures for Identifying Duplicate Mortgage Records

In Section 3.3, we present six examples of (potentially) duplicate case records within FHA's single-family data warehouse. One scheme identified cases whose FHA case number was not consistent with the zip code on its property address. This worked well on cases having a zip code and turned up about 1,000 duplicates. Unfortunately, it was not more useful because millions of case records do not have a zip code in the data warehouse. Another scheme focused on cases with the highest serial numbers and the lowest serial numbers for each office. A related scheme made use of the fact that serial numbers are assigned in chronological order. Both of these are types of "range tests." A final scheme used deterministic record linkage techniques to identify pairs of records with identical serial numbers and mortgages amounts. Such record pairs were then investigated manually using an information-retrieval system to identify and delete duplicate case records.[4]

## 3.3 Examples of Problem Cases Found in the Data Warehouse

We now discuss some representative examples that illustrate the problems discussed in Section 3.1:

**Example 3.1:**

Consider the following case record.

| FHA Case Number | Street Address | City | State | Zip Code | Name(s) of Borrower |
|---|---|---|---|---|---|
| 441-1451573 | 704 Hand Ave | Cincinnati | PA | 45232 | Collins, Larry & Juanita |

Here we have a record with a Philadelphia, Pennsylvania office code of "441" but a Cincinnati, Ohio zip code. Further research showed that the office code should have been entered as "411" so that the actual record should have been:

| FHA Case Number | Street Address | City | State | Zip Code | Name(s) of Borrower |
|---|---|---|---|---|---|
| 411-1451573 | 704 Hand Ave | Cincinnati | OH | 45232 | Collins, Larry & Juanita |

**Example 3.2:**

Consider the following case record.

| FHA Case Number | Street Address | City | State | Zip Code | Name(s) of Borrower |
|---|---|---|---|---|---|

[4] Section 5.4 of Herzog, Scheuren, and Winkler [2007] contains a general discussion of deterministic tests used in data quality work while Section 8.3 describes deterministic record linkage techniques of the type employed here.

| | | | | | |
|---|---|---|---|---|---|
| 441-2760279 | 1309 Oakland Road | Richmond | PA | 23231 | McNelly, Robert |

Here we have a record with a Philadelphia, Pennsylvania office code of "441" but a Richmond, Virginia zip code. Further research showed that the office code should have been entered as "541" so that the actual record should have been:

| FHA Case Number | Street Address | City | State | Zip Code | Name(s) of Borrower |
|---|---|---|---|---|---|
| 541-2760279 | 1309 Oakland Road | Richmond | VA | 23231 | McNelly, Robert |

Examples 3.1 and 3.2 are examples in which internal consistency checks were used to identify erroneous entries within mortgage records that led to the identification of mortgage records as duplicate records.

**Example 3.3:**

We next consider the following six contiguous case records extracted from FHA's data warehouse in ascending order of their FHA case numbers:

| FHA Case Number | Initial Mortgage Amount | Begin Amortization Date | Contract Interest Rate | Name(s) of Borrower |
|---|---|---|---|---|
| 131-5132066 | $70,100 | Aug-1987 | 10.5% | Bopp, Jeffrey S. |
| 131-5132095 | $47,750 | Oct-1987 | 11.0% | Fudge, Robert W. |
| 131-5132116 | $76,500 | Aug-1987 | 10.5% | Woods, Sherrie D. |
| 131-5132151 | $68,800 | Jun-1984 | 14.5% | O'Hara Edward J Kim A |
| 131-5132180 | $47,600 | Aug-1987 | 10.5% | Epps, Sherri A. |
| 131-5132197 | $43,250 | Oct-1987 | 10.5% | Collins, Richard M. |

In the third column of the table, we observe that the fourth record appears to be out of sequence. The year that this loan began to amortize was 1984 versus 1987 for the five other loans shown. Moreover, its contract interest rate of 14.5% is much higher than those of the other five loans. Further examination revealed that the correct FHA case number for this mortgage was 131-3807331 and so this record was a duplicate entry that needed to be deleted from the FHA single-family data warehouse. In a sense then, the record on FHA case number 131-5132151 has an inconsistency between its case number and both (1) its begin amortization date and (2) its interest rate.

**Example 3.4:**

In this example, we examine, in ascending order of the FHA case number, the first six case records on the FHA single-family data warehouse within office "372" having a begin amortization date on or after January 1, 1975.

| FHA Case Number | Property Address | Begin Amortization Date | Contract Interest Rate | Name(s) of Borrower |
|---|---|---|---|---|
| 372-0101064 | 295 Kings Highway<br>Amhurst, NY | Dec-1979 | 11.5% | Bruce, Ronald H |
| 372-0113867 | 160 Rand St<br>Rochester, NY 14615 | Apr-1983 | 12.0% | Stuart P D |
| 372-0116726 | 538 Spencer Road<br>Rochester, NY 14609 | Jun-1983 | 12.0% | Laudico M J |
| 372-0707605 | 256 Hazelwood Ave<br>Buffalo, NY 14215 | Jan-1975 | 9.5% | Richardson AF L |
| 372-0707736 | 52 Ackerman St<br>Rochester, NY 14609 | Jan-1975 | 9.0% | Bowers John P |
| 372-0708494 | 22 Worcester Place<br>Buffalo, NY 14215 | Jan-1975 | 9.0% | McDowell A L |

We note here that the first three records listed appear to be out of sequence – or, as Naus [1975] says, out of range. Further research revealed that the FHA case numbers corresponding to these three records should have been 372-101064x, 372-113867x, and 372-116726x, respectively. This is an example of what Naus [1975] calls a range test.

The approaches just described were relatively naïve. We developed a more sophisticated approach with the assistance of the staff in HUD's Office of Information Policy Systems. The idea was to use computerized (deterministic) record linkage techniques to match case records. The first scheme we ran successfully involved finding records with identical serial numbers and identical mortgage amounts. At the time this work was done initially, the database used had around 10 million records. The result was a file consisting of nearly 200,000 matches from which we identified about 5,000 duplicate records by doing clerical follow-up/review. In this process, we focused primarily on comparing the property addresses of the matching pairs of records. We might have been able to do this more efficiently if we

had had software that standardized and parsed our property addresses. At the time we were unaware of the existence of commercial software that performs these tasks although some of this software is expensive.[5]

We consider two examples to illustrate this work.

**Example 3.5:**

| FHA Case Number | Property Address | Initial Mortgage Amount | Begin Amortization Date | Interest Rate | Name(s) of Borrower |
|---|---|---|---|---|---|
| 372-1132617 | 10998 Mill Rd<br>Bethany, NY 14054 | $33,000 | Feb-1983 | 12.5% | Brown David O |
| 374-1132613 | 10998 Mill Road<br>Bethany, NY 14054 | $33,000 | Feb-1983 | 12.5% | Brown JR |

Here, we have two records that we matched on the serial portion of their FHA case numbers – 113261-—as well as on their initial mortgage amounts—$33,000. It is clear that they also match on their begin amortization dates, interest rates, property addresses, and borrower name(s). We note the variation in the spelling of "road" in the property address field as well as the variation of the names in the borrower name field.

**Example 3.6:**

| FHA Case Number | Property Address | Initial Mortgage Amount | Interest Rate | Name(s) of Borrower |
|---|---|---|---|---|
| 371-0912411 | 1111 Stratford Ave<br>Bronx, NY 10464 | $54,950 | 15.5% | Bynum William Randolph |
| 374-0912416 | 1111 Stratford Ave<br>Bronx, NY 10464 | $54,950 | 15.5% | Bynum William Randolph |

Here again, we have two records that we matched on the serial portion of their FHA case numbers—091241—as well as on their initial mortgage amounts --$54,950.

[5] See Section 19.2 of Herzog, Scheuren, and Winkler [2007] for more details about such software.

## 3.4 Consequences of Problem of Duplicate Mortgage Records

Some of the current consequences of these errors are as follows:

- Such errors have a deleterious effect on these important databases and adversely affect their use for analytical (e.g., statistical and actuarial) studies. In particular, it is hard enough to construct accurate claim and prepayment models for the MMIF, even when the data are 100% accurate. Such errors also lead to an overestimate of the amount of insurance-in-force and a corresponding underestimate of the fund's capital ratio.
- Some valid FHA cases may not be entered onto the single-family data warehouse because another case had previously been entered onto the system with that FHA case number.
- Some claim payment requests may be delayed because the data on the case cannot be found on the single-family data warehouse under the correct FHA case number. This might occur if the FHA case number is not entered onto the system correctly. Such delays can be expensive as they increase HUD's interest costs.

Some consequences of these errors, encountered in the recent past, but not of consequence today include:

- There was potential for problems with unearned premium refunds. For example, HUD could have paid both an insurance claim and an unearned premium refund on the same FHA-insured mortgage, or HUD could have paid two or more unearned premium refunds on the same mortgage.
- These problems could have lead to fraud as unscrupulous tracers[6] pressured borrowers to accept multiple unearned premium refund payments, or as HUD employees or contractors attempted to take advantage of the situation.
- These errors could have contributed to the unfavorable publicity HUD receives for not finding borrowers who are supposedly owed money by HUD. This is in contrast to the vigorous efforts that other U.S. government agencies (e.g., the Internal Revenue Service) make to collect money owed the U.S. government.
- These errors had a deleterious effect on the financial condition of the Mutual Mortgage Insurance Fund. Even when a check was not sent to a borrower for a refund payment, the fund was nevertheless debited when an unearned premium was declared on a mortgage record in the data warehouse.

[6] "Often referred to as a third-party tracer (because the government is not directly involved with the refund process), a tracer can be defined as an information broker who works to locate individuals due a refund, notify them of unclaimed monies owed to them, help them obtain it from HUD/FHA, and receive a percentage of that refund as a fee in exchange for these services." Source: http://www.webspawner.com/users/mrsauz/.

# 4. MORTGAGE RECORDS WITH AN INCORRECT TERMINATION STATUS

## 4.1 Statement of Problem

The second problem we considered was the existence of hundreds of thousands of mortgage records residing on the single-family data warehouse and having a termination status of "active" when the underlying mortgage has actually terminated, usually by prepayment.

## 4.2 Procedures Used to Identify Mortgage Records with Incorrect Termination Status

The first scheme we used was a naive internal consistency check within the data warehouse to identify addresses as identical. We simply paired records that agreed on both (1) the first 10 alphanumeric characters of the street address of the insured property and (2) the first four digits of the zip code of the property address of the insured property. (As discussed in Chapter 8 of Herzog, Scheuren, and Winkler [2007], this is a type of record linkage.) Thus far, this process has worked reasonably well in that it has generated tens of thousands of pairs of records to examine and the termination status of a vast majority of these records was in fact in need of correction. Because the FHA single-family data warehouse has over 37 million records and the computer[7] we are using does not have the ability to process that many cases at once, we have blocked the data by office code. This means that we had to partition our data into four or five groups according to office code in order to process all of the cases.

We used a second internal consistency type of record-linkage scheme to identify mortgage records whose termination status was "active" but that were in fact prepayments as follows. From all of the mortgage records that had an entry in the field "old FHA case number" we extracted the old FHA case number. We then created a file consisting of all of the records with the old FHA case numbers whose termination status was "active." This enabled us to identify slightly over 8,000 mortgage records whose termination status needed to be changed from "active" to "terminated" by prepayment. The hope here was that these changes could be done in batch via an automated process rather than manually on a case-by-case basis.

A third scheme we used matched FHA records to records in a database maintained by the Government National Mortgage Association (GNMA). GNMA, like FHA, is an agency within HUD.

[7] We did our computing using IBM's APL2 Version 2. This has a maximum workspace size of 2 gigabytes.

GNMA's main task is to package FHA and VA mortgages into mortgage-backed securities and to sell these to investors. This is a more typical type of record linkage in that it involved two distinct databases. Here we were able to identify over 32,000 pairs of mortgage records in which the loan was listed as "active" on the record in the FHA data warehouse but the matching record was listed as "terminated by prepayment" in the GNMA database. The matching process was aided by the presence of unique identification numbers – namely, FHA case numbers – in both databases. To complicate matters slightly, the FHA case numbers in the GNMA database were frequently in formats other than the one required. Specifically, the field for the FHA case number in the GNMA database is supposed to consist of 15 digits. The first two digits are each supposed to be zero; the next ten digits are supposed to be the actual FHA case number including the check digit; and the last three digits are supposed to represent the FHA "ADP Section-of-the-Act" code. Frequently, this field in the GNMA database consisted of the three-digit FHA State/Office Code, followed by a dash, and ending with the six digit serial number and the check digit. Less frequently, the field had additional leading zeros. Finally, in a number of instances, the field had non-numeric characters (besides the dash in the fourth position and blank spaces at the end). To deal with these problems, we did separate analyses for those records that had a dash in the fourth position. Otherwise, we decided to delete from our analyses all case records whose FHA case number fields had any non-numeric characters.

## 4.3 Examples of Mortgages with Incorrect Termination Status

**Example 4.1:**

| FHA Case Number | Property Address | Begin Amortization Date | Status of Loan | Interest Rate | Name(s) of Borrower |
|---|---|---|---|---|---|
| 371-1019310 | 109-07 211th Place<br>Queens Village, NY 11429 | Mar-1982 | Active | 16.5% | Smith John Paulette |
| 374-4413730 | 109-07 211th Place<br>Queens Village, NY 11429 | Sep-2004 | Active | 5.5% | Smith, Paulette |

In this example, the data warehouse lists two active mortgages on a property located in Queens Village, New York, where in reality the first mortgage that was originated in March of 1982 at an interest rate of 16.5% has been refinanced at least once, i.e., during September 2004.

**Example 4.2:**

| FHA Case Number | Property Address | Begin Amortization Date | Status of Loan | Interest Rate | Name(s) of Borrower |
|---|---|---|---|---|---|
| 372-1221854 | 323 HIGHGATE AVE BUFFALO, NY 14215 | Apr-1984 | Active | 13% | J. & K. Falkides |
| 372-1519223 | 323 HIGHGATE AVE BUFFALO, NY 14215 | Jan-1987 | Prepaid | 9% | Falkides, John P |

In this example, the data warehouse lists one active mortgage and a later mortgage terminated by prepayment (denoted by "T") for a property in Buffalo, New York. It appears that the first mortgage was originated during April of 1984 at an interest rate of 13% and refinanced during January of 1987 at an interest rate of 9%.

**Example 4.3:**

| FHA Case Number | Property Address | Begin Amortization Date | Status of Loan | Interest Rate | Name(s) of Borrower |
|---|---|---|---|---|---|
| 131-5109339 | 14104 RTE 1750<br>COAL VALLEY, IL 61240 | Oct-1986 | Active | 9.0% | Wangel, Dale J |
| 131-7200510 | 14104 RTE 1750<br>COAL VALLEY, IL 61240 | Jul-1994 | Claimed | 7.5% | Wangel, Dale E |

In this example, the data warehouse lists one active mortgage and a later mortgage terminated by insurance claim for a property in Coal Valley, Illinois. Again, it appears that the first mortgage was originated during October of 1986 at an interest rate of 9% and refinanced during July of 1994 at an interest rate of 7.5%. The second loan eventually resulted in an insurance claim being paid by HUD.

## 4.4 Consequences of Problem of Incorrect Termination Status

We list below some of the consequences of this problem.

- The amount of insurance in force for the Mutual Mortgage Insurance Fund, the General Insurance Fund, and the Special Risk Insurance Fund are all overstated.
- Some lenders are paying periodic mortgage insurance premiums on mortgages that have already terminated.
- Some borrowers have not been paid unearned premium refunds or distributive shares to which they are (or were at one time) entitled.
- Such errors have a deleterious effect on these important databases and adversely affect their use for analytical (e.g., statistical and actuarial) studies. In particular, it is hard enough to construct accurate claim and prepayment models for the MMIF, even when the data are 100% accurate.

# 5. MORTGAGE RECORDS WITHOUT A PROPERTY ADDRESS

## 5.1 Statement of Problem

A number of years ago, some HUD staff made a decision not to have a field for property address on the HUD database of FHA-insured single-family mortgages. This decision has since been reversed, but HUD recently still had about 25,000 insured single-family mortgages on its single-family data warehouse

that lacked a property address – a critical data element. Although this is a large number of loans, it represents less than one percent of the FHA-insured single-family mortgages currently in-force.

### 5.2 Procedures Used to Find Property Addresses

We used the record linkage scheme of Section 4, in which we linked FHA mortgage records with GNMA records to append the property address from the GNMA record to the corresponding (i.e., matched) FHA record. Thus far, we have thereby obtained addresses for about 5,600 of the 25,000 FHA mortgage records.

### 5.3 Consequences of Missing Addresses

The following are some of the consequences of missing addresses:

- FHA could pay an insurance claim on a mortgage that it had not insured.
- FHA could pay a premium refund and/or a dividend on a mortgage that it had not insured.

## 6. ESTIMATING THE NUMBER OF PROBLEM MORTGAGE RECORDS

One scheme for estimating the number of duplicate records on the FHA single-family data warehouse involves the use of a procedure known as "capture-recapture." This is described in a number of texts, e.g., Bishop, Fienberg, and Holland [1975]. This involves drawing two or more independent samples from the data warehouse. If two samples—A and B—are used, for example, then this procedure entails identifying the number of duplicate records found (1) in both samples, (2) in sample A but not in sample B, and (3) in sample B but not in sample A.

At first glance, one might think that capture-recapture methods might work well for estimating the number of mortgages with the wrong termination status. However, this is not the case. We could only use capture-recapture methods here to estimate the number of mortgage records we could potentially correct using the scheme described in Section 4. This is primarily because some borrowers might terminate their mortgages either without refinancing with FHA or by selling their property to someone who does not take out an FHA-insured mortgage. So, we are faced with coming up with a different approach. One possible approach is to examine the contract interest rates of the "active" mortgages as well as the mortgages that are listed as "active" but have not paid any required periodic mortgage premiums during the last five years. The later type of mortgages can be identified because the servicing

lender identification number of such mortgages is given by "99995". The following table can be used to take a first, albeit naïve, approach to this problem.

**TABLE 1**

Number of "Active" MMIF Mortgages as of September 30, 2005

| Contract Interest Rate | Servicing Lender Number | | Total |
|---|---|---|---|
| | 99995 | Not 99995 | |
| ≥ 10% | 64,650 | 271,695 | 336,345 |
| ≥8% but <10% | 26,544 | 861,912 | 888,456 |
| < 8% | 4,583 | 2,704,051 | 2,708,634 |
| Total | 95,777 | 3,837,658 | 3,933,435 |

According to the table above, as of September 30, 2005, the data warehouse listed over one million mortgages as "active" with an annual contract interest rate of at least 8% even though such interest rates recently were as low as about 5%. Because of all of the work we have done using our naïve matching schemes, we suspected that a large proportion of these mortgages had, in fact, been prepaid. Moreover, the mortgages whose servicing lender has been assigned the number "99995" have not paid any required periodic premiums during the last ten years or so. Based on our extensive experience at examining these mortgage records, we felt brave enough to make the following estimate, albeit highly subjective, of the number of these nearly four million "active" mortgage records that we thought were in actuality "terminated."

## TABLE 2

Estimated Number of "Active" MMIF Mortgages as of September 30, 2005,
That Have Actually Been "Terminated"

| Mortgage Characteristics | | Number of "Active" Mortgages | Estimated Percentage of Mortgages that should have "terminated" | Estimated Number of Mortgages that should have "terminated" |
|---|---|---|---|---|
| Servicing Lender Number | Contract Interest Rate | | | |
| 99995 | Any | 95,777 | 95 | 92,000 |
| Not 99995 | ≥ 10% | 271,695 | 90 | 240,000 |
| Not 99995 | ≥8% but <10% | 861,912 | 50 | 430,000 |
| Not 99995 | < 8% | 2,704,051 | 2 | 54,000 |
| | | | TOTAL | 816,000 |

With more recent data, we have taken another look at this situation.

## TABLE 3

Number of "Active" MMIF Mortgages as of July 31, 2009

| Contract Interest Rate | Servicing Lender Number | | Total |
|---|---|---|---|
| | 99995 | Not 99995 | |
| ≥ 10% | 39,060 | 99,358 | 138,418 |
| ≥8% but <10% | 3,387 | 310,043 | 313,430 |
| < 8% | 2,801 | 4,598,217 | 4,601,018 |
| Total | 45,248 | 5,007,618 | 5,052,866 |

Again, we felt brave enough to make the following revised estimate, albeit highly subjective, of the number of these roughly five million "active" mortgage records that we thought were in actuality "terminated."

## TABLE 4

Estimated Number of "Active" MMIF Mortgages as of July 31, 2009
That Have Actually Been "Terminated"

| Mortgage Characteristics | | Number of "Active" Mortgages | Estimated Percentage of Mortgages that should have "terminated" | Estimated Number of Mortgages that should have "terminated" |
|---|---|---|---|---|
| Servicing Lender Number | Contract Interest Rate | | | |
| 99995 | Any | 45,248 | 95 | 43,000 |
| Not 99995 | ≥ 10% | 99,358 | 90 | 90,000 |
| Not 99995 | ≥8% but <10% | 310,043 | 50 | 155,000 |
| Not 99995 | < 8% | 4,598,217 | 1 | 46,000 |
| | | | TOTAL | 334,000 |

This revised estimate of 334,000 case records is a vast improvement over our previous estimate of 831,000. Something is going right. At the moment, we can only speculate as to what that is. Perhaps, the reason is that with the recent slowdown in the number of mortgages originated that began during the summer of 2007, lenders are finally getting a chance to catch up on their back-office tasks.

We note here that the "rule-of-thumb" in effect until about 10 years ago was to refinance one's mortgage when the current mortgage interest rate is 2% (i.e., 200 basis points) below the annual contract interest rate on one's current mortgage. We realize that some people do not take advantage of this opportunity because their outstanding balance is low and the costs of refinancing outweigh the savings from the lower interest rates. (On the other hand, FHA offers a "streamlined" refinancing option with minimal costs to the borrower.) Other people do not refinance because they are financially naïve or are undergoing major personal problems. So, some keen observers of the mortgage market may think that our estimates are too high while perhaps others may conclude that they are too low. In

any case, we hope that working with our colleagues at HUD, we can keep this process moving and continue to correct these data fields as expeditiously as possible without introducing any additional errors in the process.

## 7. REFERENCES

[1] Bishop, Y., S. Fienberg, and P. Holland, *Discrete Multivariate Analysis*, MIT Press, Cambridge, Mass., 1975.

[2] Herzog, T.N., F.J. Scheuren, and W.E. Winkler, *Data Quality and Record Linkage Techniques*, Springer, New York, NY, 2007.

[3] Naus, J., *Data Quality Control and Editing*, Marcel Dekker, New York, 1975.

### Acknowledgements

The author would like to thank William J. Eilerman, David A. Middaugh, Teri Hines, and Zenora Hines for their kind assistance with this work.

### Biography of the Author

**Thomas N. Herzog, Ph.D., A.S.A.**, is the chief actuary of the Federal Housing Administration in Washington, D.C. He has a Sc.B. in applied mathematics from Brown University and a Ph.D. in mathematics (with a major in statistics) from the University of Maryland. He is a Fellow of the American Statistical Association and an Associate of the Society of Actuaries. Dr. Herzog is the author or co-author of four books: *Introduction to Credibility Theory*, *Applications of Monte Carlo Methods to Finance and Insurance* (written with Prof. Graham Lord of Princeton University), *Models for Quantifying Risk*, (written with Robin Cunningham and Dick London) and *Data Quality and Record Linkage Techniques* (written with Fritz Scheuren and William Winkler). He is the winner of the 1990 AERF Practitioner's Award for a paper on *Home Equity Conversion Mortgages* (written with Theresa R. DiVenti).