Paper SY01

"Methods for Minimizing Errors in Address Data"

Milorad Stojanovic
RTI International, RTP, NC

ABSTRACT

We live in an imperfect world and we work with imperfect data. Data often contain small and large errors. This is especially true when collecting data from mail surveys that are sometimes filled out in a hurried or careless manner. Correct address data is critical to achieving high mail survey response rates. This paper presents methods to improve the quality of the street address data obtained from mail surveys and other instruments that collect address data. Improvements are made in the quality of the information typically stored as ADDR1 and ADDR2 by using: standardization techniques, correction methods, and business rules.

SAS/Base and SAS/Macro are used to implement the ideas presented in this paper.

INTRODUCTION

In his book Larry English (1) asked, "Why is an organization – or should it be – interested in information quality?" His answer, "It can be summed up in one word: **Profit.**". In the case of survey research and marketing organizations, we are naturally interested in achieving the highest possible response rates. One way to improve response rate is to improve the quality of information in all segments of survey.

Current data quality problems cost U.S. business more than **$600 billion per year** (2).

In the recent article, "Report: 25% of Critical Data is Flawed", Sharon Gaudin (3) stated that according to the analysts from Garthner, a major industry research firm based in Connecticut, "more than 25 percent of critical data within Fortune 1,000 companies is incomplete and inaccurate",

"What's causing the flaws in the data is not a computer inaccurately calculating data", says Gordon Haff, a senior analyst with Illuminata, a New Hampshire based analyst firm. "It is typically going to be more an issue of what data is being collected and how it's being collected as opposed to it being incorrectly processed once it's in the system."

The old and much known principle, "Garbage In Garbage Out" (GIGO), is still valid.

After those serious and not very positive statements, what can we do to change the situation? We think we can do a lot at each and every step in the project and the **SOONER** the **BETTER.**

In this paper, primary attention is given to personal addresses in the US and Canada. To cover the whole world is a tremendous task and this paper doesn't have an answer for that. Different countries have different standards for the Data Address Line (DAL). Many times those standards are contradictory to the standards in US and Canada.

WHAT WAS THE TRIGGER FOR THIS PAPER?

Approximately a year ago we got a complaint from a telephone interviewing group. One telephone interviewer said he had the same data in two consecutive roster lines. This seemed strange because the data had been checked and records that had identical addresses were removed.

What was 'the same' for the human eyes and mind was not 'identical' for the computer program. We checked the rosters and found this particular case. The two rosters did indeed contain the same information but the addresses were slightly different. The two address lines were something like:

115 Franklin Street South Apt #38
115 Franklin St South Apt. 38

These are obviously the same delivery address, but they are not identical. If USPS standards were followed, the address would be:

115 FRANKLIN ST S APT 38

USPS 'Postal Addressing Standards' Publication 28 clearly states how an address should be presented. In addition, there are guidelines to prevent too much information which can lead to confusion and many times to late or failed delivery.

WHAT IS AN ADDRESS?

"Addresses provide a means of locating people, structures and other spatial objects".

In countries all around the world, specialists from many areas are working on various problems regarding addresses. In the USA, work on an "Address Data Content Standard" was done by the Subcommittee on Cultural and Demographic Data (4). The standard does not apply to addresses of entities that lack a spatial component and specifically excludes electronic addresses, such as e-mail addresses.

All addresses consist of at least three address lines. Usually an address consists of:

1. Recipient Line
2. Delivery Address Line (DAL)
3. Last Line

WHAT ARE THE ADDRESS PROBLEMS?

In organizations which deal with surveys, the goal is to have high response rates while keeping costs to a minimum. If a survey can't be delivered due to address problems, the response rate is reduced. Response rates are improved when questionnaires are successfully delivered to the survey participant on the first try. Questionnaires and supporting documents are delivered by United States Postal Services (USPS), FedEx, UPS, and other mail carriers. Mail carriers will be more efficient and happier if recipient's addresses are valid and clear. Many times addresses are valid, but the sequence of data is 'free style', or it may contain unnecessary additional information that causes confusion and failure to deliver.

Another source of many undelivered surveys is changing addresses due to people moving. There is a constant need to update the addresses. Keeping addresses up to date is a tremendous task when you consider that over 40 million Americans change addresses annually (5). The National Change of Address (NCOA) service, provided by the USPS, makes this address change information readily available and it significantly reduces undeliverable mail pieces. This service needs to be fully utilized in maintaining address data bases.

Errors are often generated in writing Street number, Apartment number, and/or Post Office Box number. It is the habit of some respondents to give – Street name and number as well as Post Office Box number. This is too much information and an obvious source of confusion for a mail carrier.

The variety of ways to enter Rural Route, Highway Contract Routes etc. produces even more mess.

Recently we had an opportunity to read the paper "What is my Address?", by Dan Tasker (6). He asks the question, "Who knows best what my address is – me or you?" We agree with his opinion completely. Many times respondents are frustrated and fill in meaningless or too much information. It is like trying to install software on a personal computer. Often an answer is required before you can go forward, so any sequence of meaningless characters is entered in the box like, "Your company name", just to proceed further. If you don't enter those meaningless letters, you can't proceed. Something similar can happen when the designer of an interactive form knows the proper format for your address better than you.

This paper does not discuss the handling of invalid or misspelled street names or numbers. This is beyond the scope of this paper because it requires more sources of data like the Street Register for each state.

HOW CAN ADDRESS PROBLEMS BE REDUCED?

Each element of an address should be improved and standardized to conform to the USPS Postal Standards (7).

Here we like to point out the paper by David Loshin (8). He made a very good remark about the use of delivery address line 1 and line 2. He states, "It is that the use of two fields to capture a mailing address allows data entry personnel to input information that can be subsequently 'lost' between those two fields".

The trickiest line for standardization and unification is the Delivery Address Line (DAL). What are the most frequent mistakes or imprecision in DALs?

STANDARDIZING:

Using a few hundred thousand DALs, it is possible to extend the following three check lists given by USPS:

1. Geographical Directional like: North, South, East, West
2. Street Abbreviations like: Street, ST, Alley, Drive, DR
3. Secondary Unit Designators like: Rural Route, Post Office Box, Apartment, and Lot etc.

We call this the learning capability of the program. Under learning capability, we classify improperly written parts of a DAL into proper categories like:

ALLIE → ALLEY -> ALY (proper abbreviation)

There are a number of such examples. Their number is correlated to the number of input records. Also the large number of input DALs allows us to test robustness and correctness of algorithms used in the program.

When we are sending questionnaires to respondents we are usually using:

1. USPS/FEDEX/UPS and other companies they are essential services for delivering survey questionnaires to respondents.
2. We MUST do our best to provide mail carriers with correct and up to date address.

TOKENS AND DELIMITERS

The first thing in solving address problems is in separating all data in the DAL into tokens. To do that first we must choose separating characters. The blank character is not the only one which separates each part of the DAL. We found many special characters exist in DALs. Their presence is a consequence of many different factors from free usage of those characters to inaccurate transcription, etc. Carefully looking at all USPS standards and exceptions we developed a list of delimiters.

Some special characters are not delimiters – those characters should be removed and the space which they occupied should be compressed. The apostrophe in the example below is one example of these special characters that are removed in the process of standardizing addresses.

113 SIMMON'S ROAD should be **113 SIMMONS ROAD**

STEPS IN STANDARDIZATION AND UNIFICATION OF DALs

1. Save the original DAL.
2. Change all letters to upper case.
3. Replace double or more spaces between words with a single space.
4. Shift content of DAL to the left.
5. Apply the macro for standardization and unification.
6. Compare results.

What the macro does:

In preparation, we made an extensive library list of Geographical Directional, Street Abbreviations, and Secondary Unit Designators. It includes USPS suggested words and our extensions (i.e. misspelled words).

- The macro parses addresses for unusual presentation of Rural Route, Post Office Box, Apartment, and Lot etc. These unusual presentations are replaced with properly standardize abbreviations. If space is not available, the text is shifted to the right as needed.
- All characters which are understood as delimiters are replaced with space character (blank).
- The new DAL is compressed, trimmed, and its length is computed.
- The SCAN function is used to parse each token of the DAL variable. If the DAL has redundant elements the macro chooses which elements should be removed.
- Special attention is given to the exemptions and unusual situations.
- The macro removes leading zeros from street numbers and from rural route numbers.
- If rural route numbers are attached to the text RR, a blank character is inserted. USPS standard for rural routes is followed completely – RR N BOX NN.
- Similar rule is applied for Post Office Box – PO BOX NNNN.
- Highway Contract Route addresses are standardized in the form - HC N BOX NN.
- Standardization for County, State, and Local highways require FULL text of corresponding words and it is done as follows. In the case of abbreviation, full text is applied. For example: CR 1230 will become COUNTY ROAD 1230
- The macro handles DALs with two suffixes as follows:. For example: 500 MAIN ROAD WEST becomes 500 MAIN ROAD W
- Special attention is given to addresses in Puerto Rico an US Virgin Islands (Spanish speaking territories).
- Algorithm in the macro avoids the use of abbreviations when the two-word directional is the primary street name or when directional is part of the street name.
- The macro presents just one designation in the case when more than one is present. For example: HC 30 BOX 2 MILLER WAY becomes HC 30 BOX 2
- Designations like CALLER, FIRM CALLER, BIN, LOCKBOX, or DRAWER are changed to PO BOX.

- Special attention is dedicated to all exceptions.

OTHER SOLUTIONS

There are software products on the market, some of them claim to do all standardization even including international addresses. We would like to mention the FIXADDRESS macro (9), by Mike Zdeb. In our work, we applied a slightly different approach.

CONCLUSION

Improving quality of addresses is not an option - it is imperative for success in survey research. The development of this method and the use of the macro described in this paper, yielded significant improvements in address quality, especially when unedited addresses were processed.

LIMITATIONS

This paper is limited in two ways. The topic is relatively modest – improvement of delivery address lines (DALs). Also this paper does not deal with DALs outside the USA and Canada. Many foreign countries have different ways of presenting DAL – often quite different to the USPS guidelines. The prospect for improving foreign addresses was enhanced with the mid 2003 Universal Postal Union (10) publication "Approved International Address Standard UPU S42-1" deals with International Postal Address Components and Templates.

FUTURE WORK

In the future, we hope to design and implement a complex macro which will deal with complete addresses. It will include usage of ZIP+4 file and improved logic for comparing pieces of data.

REFERENCES

1. Larry English, "Improving Data Warehouse and Business Information Quality", John Willey & Sons, Inc., 1999

2. "Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data.", The Data Warehousing Institute. Report Serious 2002.

3. Sharon Gaudin, "Report: 25% of Critical Data is Flawed", May 19 2004
URL: http://itmanagement.earthweb.com/datbus/print.php/3356211

4. "Address Data Content Standard", Subcommittee on Cultural and Demographic Data, Federal Geographic Data Committee, April 17, 2003
URL: http://www.census.gov/geo/www/standards/scdd/AddApril2003h2.htm

5. National Change of Address, USPS
URL: http://www.usps.com/ncsc/products/ncoa.htm

6. Dan Tasker, "What's my Address?", Business Rules Journal, Vol.4, No. 9 (Sept. 2003)
URL: http://www.BRCommunity.com/a2003/b165.html

7. "Postal Addressing Standards", USPS, Publication 28, November 2000

8. David Loshin, "Knowledge Integrity: MID – Missing In Data", DM Review, October 2003

9. Mike Zdeb, from "Combining Data", appendix B - Address Standardization (macro FIXADDRESS)

10. "International Postal Address Components and Templates – Standard S42-1", Universal Postal Union, June 17 2003,

URL: http://xml.coverpages.org/ni2003-06-17-a.html

ACNOWLEDGEMENTS

I would like to thank all my colleagues in RTI for their patience and helpful comments. I would especially like to thank Laura Burns from RTI Education Survey Division, for her time and willingness helping me to make this paper available for publication.

AUTHOR CONTACT INFORMATION

Milorad Stojanovic
RTI, International
Research Computing Division
RTP, NC, 27709
(919) 541-7376
E-mail: **milorad@rti.org**