THE ZIPSTAN STANDARDIZATION SYSTEM

George L. Farnsworth
University of Southern California

It is curious that computer-assisted cartography is not more widely used in the urban setting, considering its quite long history and highly developed technical state. Conventional statistical techniques as represented by such computer packages as SPSS and BIOMED* are far more widely used in urban applications than equivalent statistical mapping systems. Indeed, while it would be very difficult today to find an urban statistician willing to calculate regression statistics on an old Frieden calculator, many cities still have draftsmen generating statistical maps with the techniques of twenty years ago. One result of this situation is that statistical analysis has found rapidly increasing applications but at the same time geographic analysis has remained relatively static.

It is my feeling that a major cause of this phenomenon is that preparation of a given data file for non-spatial statistical analysis is substantially easier than preparation of the same data file for computer cartography. We encounter numerous examples of police departments, welfare agencies and others who seem willing to convert their records to machine-readable form for analysis but who omit the street address or any but the simplest geographic identifiers (ZIP code, precinct, etc.) from the conversion or analysis. The U. S. Manpower Administration, for example, sponsors a system called ESARS (Employment Security Automated Reporting System) used by almost all states for statistical analysis of characteristics of the unemployed but which does not contain the applicant's address or any geographic level below county.

By now it should be well known that in urban areas, street addresses are convertible into geographic coordinates or other kinds of location codes (census tract, block) since such methodology at least in rudimentary form, has a long history. The truth is, however, that until recently actually performing the conversion from street addresses to other location codes was quite troublesome and involved substantial manual work. Since the early 1970's the Census Bureau's GBF/DIME System has provided a machine-readable index to the conversion and the ADMATCH system a software product to assist in making the conversions.

The ADMATCH system itself was an advance over some earlier techniques used at Census and elsewhere but is finally being replaced by the UNIMATCH system for actually comparing addresses to the DIME records. Without going into detail about UNIMATCH, which has already been widely described, it is enough to say that it is a very fast and flexible system for implementing various matching rules to achieve very accurate and complete address conversion.

---

* SPSS(Statistical Program for the Social Sciences) and BIOMED (Biomedical Statistical Program) are two sets of computer programs which produce various types of statistical analyses.

Proper use of UNIMATCH, however, frequently requires some prior processing of the records to ensure that the house number, street name and other address components are in known positions and formats. The ZIPSTAN system is designed to automate this process, known as "standardization." Operating under a user command language, the system parses and performs syntax analysis of the input addresses to recognize, isolate and standardize the various address components. Thus a typical address such as "918 N. Washington St." might use any one of several variants for words as "north," "Washington," "street" and might also contain apartment indicators, and various punctuation symbols.

The following forms would all be standardized to the identical output address as shown:

918 N Washington St.	Los Angeles CA 90065
918 North Washington St.	LA Calif 90065
918-B No. Washington Str.	L. A. Cal. 90065
918½ N. Washington St.	Los Angeles 90065 CA
00918 N. Washington St.	LA CA 90065
918-920 No. Wash. Strt.	Los Angeles, California 90065
918 NORTH WASHINGTON STREET	LOS ANGELES, CA90065
918 N. WASH. ST. APT 17	L.A. 65 Calif. 90065

Standard output address would be generated in fields as shown:

FIELD	LENGTH	CONTENT
House No.	7	918
Prefix	2	N
Street	25	WASHINGTON
Suffix	2	ST
City	20	LOS ANGELES
State	2	CA
ZIP	5	90065

Not shown are fields for additional prefix and suffix abbreviations, apartment numbers, house number suffixes and status codes.

Provision is easily made for translating to or from street names, variants or street codes and city or town codes. The ZIP code alone may even be used to generate a city name and state.

In standardizing the Census GBF/DIME-Files for coding addresses, it is possible to skip records for non-street features, convert city codes to names and drop selected fields not needed, all in one pass of the input file.

The ideas behind ZIPSTAN are not particularly new or unusual but the implementation is successful in providing an easy system to use and one which is flexible enough to handle a wide variety of situations. The "standardized" address picked out by ZIPSTAN can be as short as 5 digits (when only ZIP code is

to be used) or as long as 100 characters when a full address including house and apartment numbers, street name, city, state and ZIP are all to be included. Components can be in fixed or variable positions within the input records and as much or as little data may be picked out of the input records and added to the standardized address for output.

In addition to its use for standardizing addresses, ZIPSTAN may be used to convert other fields, select or reject certain records, perform various selective file copy and editing functions and print portions of the input file.

The system, like UNIMATCH, is written in IBM Assembly Language for the 360/370 series of machines. A complete User's Manual is in final preparation and the combination ZIPSTAN/UNIMATCH system is available from the Center for Census Use Studies at the U. S. Bureau of the Census. Along with UNIMATCH the system has been widely used around the U. S. and in France (with some modification).

Although I myself am no longer at the Census Bureau, I am continuing development of ZIPSTAN and UNIMATCH at the University of Southern California where we are in the process of joining with several users in Europe to produce a new integrated system in COBOL. I invite comments from users of ZIPSTAN and similar systems which could be incorporated in any new versions.