BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES
SRD Research Report Number: Census/SRD/RR-87/29

Geocoding Theory and Practice at the Bureau of the Census

by

R. Thomas O'Reagan
Alan Saalfeld
Statistical Research Division
Bureau of the Census

Recommended by: Tom O'Reagan

Report completed: September, 1986

Report issued: October 1, 1987

R. Thomas O'Reagan
Principal Researcher
Statistical Research Division
Bureau of the Census
Washington, DC 20233

Alan J. Saalfeld
Principal Researcher
Statistical Research Division
Bureau of the Census
Washington, DC 20233

GEOCODING THEORY AND PRACTICE AT THE BUREAU OF THE CENSUS

ABSTRACT. In several applications at the Bureau of the Census and elsewhere, it is necessary to link an uncontrolled and perhaps garbled address on an input record with the controlled and standardized representation of that same address on a master file of possible addresses. For the Bureau of the Census, the linking permits the association of a numerical geographic code--which is present in the master file, but not on the input transaction record--and that geographic code then allows aggregation of other input data on the transaction record at county, city, and block levels. The assignment of these geographic codes to the input records is called geocoding.

For some time now, the Bureau of the Census has used a set of computer programs to accomplish this geocoding. Its goal has been to automate the result that clerical linking at its best would produce, with the additional advantages of economy, speed, and consistency. The linking system has undergone substantial changes over time and continues to evolve. Early on, for example, the linking system gave equal import to each item of the address in reaching a coding decision; later programs applied some of the fundamentals of information theory, such as regarding unusual messages as more informative than common ones. New developments in information theory, statistical matching, rule-based systems, and artificial intelligence all offer additional opportunities for improving geocoding procedures. This paper will sketch the first fifteen years of development of geocoding principles and methods.

INTRODUCTION

1492 Santa Maria Way West, Yuma AZ 85364. This hypothetical address is so complete and perfect that one might think it identifies a geographic location. Because, however, the bigness of the community to which mail is addressed tends to rub off on nearby businesses and residences, that address might not even be in Yuma but out in the surrounding county. The Post Office need not be concerned; it can certainly find such an address on the ground and deliver mail. The Census Bureau finds the address a problem; it must aggregate data for this address with others in the same county, place, tract, or even block. Further, this sort of manipulation can be handled much more economically by computer if the location is pinpointed by fixed length numeric codes rather than by free form words and numbers in somewhat arbitrary and varying conventions.

For almost 30 years the Census Bureau has been seeking ways to use postal address information to place entities in their correct geographic areas by means of a computerized system. This paper will sketch the progress over the first half or so of that period.

SYSTEM REQUIREMENTS

The fundamental requirements for an automated system of associating addresses with specific geographic codes consist of:

(1) a geographic reference file with all levels of needed codes, and
(2) computer programs which:

    (a) identify the components of the addresses to be coded
    (b) link these components to the standardized components in the geographic reference file.

Our geographic reference file consists of street segments identified by a street name and house number range within a specified piece of geography (i.e., a side of a given city block or a street segment within a tract, etc.). The street segment file is generally organized by place, post office or ZIP code and is arranged by street name and house number range. Associated with each record are the necessary geographic codes identifying the location from the smallest geographic area that is identified in the system to the state level. The components of the street name, i.e., directions (North, Southeast), and street types (Boulevard, Court, etc.) are standardized within the record format.

To round out the system two essential computer programs are needed. The first program is to identify the components of the input address, standardize the components and arrange them into a record format consistent with the geographic base file. The second program is to provide the linkage to the proper geographic base

file record through computer matching techniques and assign the standard geographic location identifiers or codes.

The second computer program can be conceived of as blocking and weighting.

BLOCKING

Blocking is a process whereby potentially linkable records from the reference file are brought together as candidates for attempted linking with a particular input record. Considerations here are twofold: to reduce the number of occasions of omission of the true match reference record from the reference block, and to hold to a reasonable level the average number of potential candidates that will be considered with each input. As a notable amount of the computer time in an automated linkage system will be devoted to the detailed comparisons of pairs of records, a significant measure of the efficiency of the system in terms of cost per linkage for a specified accuracy must be this ratio of productive to unproductive comparisons. The comparison candidates must be reduced to but a small block from the original large reference file.

The 1967 Economic Census Address Coding System, in aiming to code to street segment or tract, compared an unscrambled or standardized version (wherein each component was identified) of the response address to that block of reference file records which had been sorted into the same three-digit ZIP code and SOUNDEX code (a phonetic translation of street name). That sort produced an average block size of 14 records. Further, the response house number was required to be within the range indicated on at least one of these reference file candidates. This brought the average (sub) block size to about two records. The blocks conceptually will always include the record which is equivalent to "none of these", and linkage with that record will be described as "inability to code".

A word about SOUNDEX is in order. One would not use SOUNDEX or any other compression scheme for that matter unless it had been determined that certain apparent distinctions were not real distinctions. Miller and Mueller will be blocked together under code M460. Carruthers and Carouthers will be blocked together as C636. Unfortunately, Carter will also be C636. So some information or discriminating power is willfully discarded as spurious. In terms of the entropy measures used in information theory, SOUNDEX has one half to two thirds the information content of the full name. But the technique of compression is introduced because it blocks together many of the common types of spelling errors and abbreviations that occur.

WEIGHTING

Actually, one <u>might</u> simply observe for any pair of records being compared, how many of the total components agree and how

many disagree. By inference, this would give equal import to each component, and was in fact the basis of the 1963 version of the coding system. That worked surprisingly well on the whole, coding 70 percent of the input with only 2.5 percent in error, but wasted the discriminating power of some fields and field values which are measurably more informative than others.

As an example, suppose that in the files for a census or other application 1/16 of the records have "Boulevard" in the STREET TYPE field, while 1/4 of the records show "Avenue" in that same field. The odds against a random or accidental agreement with those two particulars are 16:1 and 4:1 respectively. For "Circle", those odds might be 512:1. To convert such odds to weighting factors, it was convenient (and consistent with tradition in information theory) to use logs to the base 2, that is, a measure of "bits" of information. Agreement of the field values just listed would therefore deserve weights of +4, +2, and +9.

Disagreements might be expected to carry negative weight. And, of course, a field which is rarely in disagreement for truly matched records must receive a much greater negative weight than one which is known to commonly disagree even in true matches.

Positive and negative weights may be computed by the same formula, which we will detail later but which examines the relative frequencies in matched pairs and random pairs, respectively.

Each individual field within the record will have attached to it a measure of the probability of random agreement between two records, so that the probability of random agreement of any combination of the fields forming the comparison pair can be computed, and by subtraction, the probability of the pair referring to the same entity.

Unless the values of a field are evenly distributed throughout the population known to the system and encountered by it, the weight which can be attached to a particular field value will depend upon its distribution, since the more frequent its occurrence the more probable is agreement in records compared at random.

It might seem ideal to have a specific weight for every possible combination of field values for every field. As a practical matter that cannot be accomplished, and the theory is appropriate whether it is applied to agreements in the STREET NAME field without the name even being specified, or to a detailed table of weights for "Circle", "Plaza", and whatever other values occur as STREET TYPES.

Say that α_i is a particular value in the i^{th} field of a record from the input file. Likewise β_i is the corresponding value in a record in the reference block. We might know that if the i^{th} field is STREET NAME, $\alpha_i = \beta_i$ about 88 percent of the time in records which should truly match. It is not 100 percent

because errors occur in spelling on one or both files. Further, we know that 44 percent of the time records in the reference block agree (i.e. $_i = _i$) with the input even though they should not truly be matched. This is because "Tenley Court" and "Tenley Circle" are often in the same block. That is, they agree on STREET NAME by mere coincidence. When we observe agreement in the STREET NAME field we can assign the weight $\log_2 (.88/.44) = +1$.

Notationally, let:

M <=> the state of nature that the records are true matches
$\overline{M}$ <=> the state of nature that the records are not true matches

In decision theory terminology, we have performed an experiment in order to gain some insight as to whether M or $\overline{M}$ is the existing state of nature, and the observed result of our experiment is a particular outcome, $_i$. In this application our observed result is the comparison pair ($_i$, $_i$) and that equates to $_i$.

We are interested in what the conditional probability of this outcome might be, given that M is the state of nature. Symbolically that is:

$$P_{i1} = P(\ _i|M) \qquad (1)$$

Naturally, we will wish to compare that with the conditional probability of the same outcome given that $\overline{M}$ is the state of nature:

$$P_{i2} = P(\ _i|\overline{M}) \qquad (2)$$

P_{i1} is the probability that this particular pattern of paired components for a comparison pair (some components may agree, some disagree) would be observed were M true, and P_{i2} is the probability of obtaining those configurations were $\overline{M}$ the prevailing state of nature.

It is reasonable to compare these two probabilities in the form of a ratio:

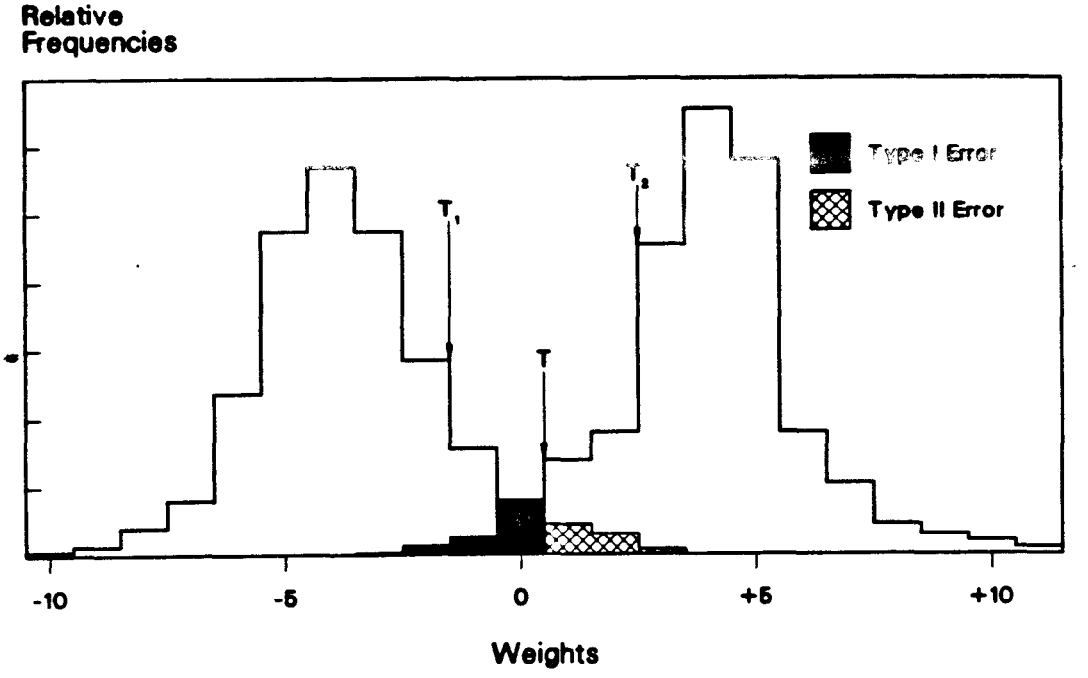
$$L_i = P_{i1}/P_{i2} \qquad (3)$$

Here we depart slightly from decision theory concepts into information theory by utilizing the $\log_2 L_i$ across all fields of the record to obtain an overall $\log_2 L$. Adding $\log_2$ weights assumes that the values in one field are independent of the values in every other field in the record, rather than that a STREET TYPE of "Court" implies somewhat strongly a DIRECTION of "North".

Then, according to the theory, if L exceeds some value T_2, ($T_2 > 1$), the hypothesis M is accepted. If the value of L falls below T_1 ($T_1 < 1$), the hypothesis $\overline{M}$ is accepted.

Note that there is a region between T_1 and T_2 where no decision would be made to favor either M or $\bar{M}$. When this non-decision area is allowed, it is possible to set T_1 or T_2 so that stipulated error levels of Type I and Type II will not be exceeded. A Type I error is committed in rejecting M when it is in fact the case, and a Type II error is in accepting M when it is not the true state of nature.

Figure 1
Thresholds



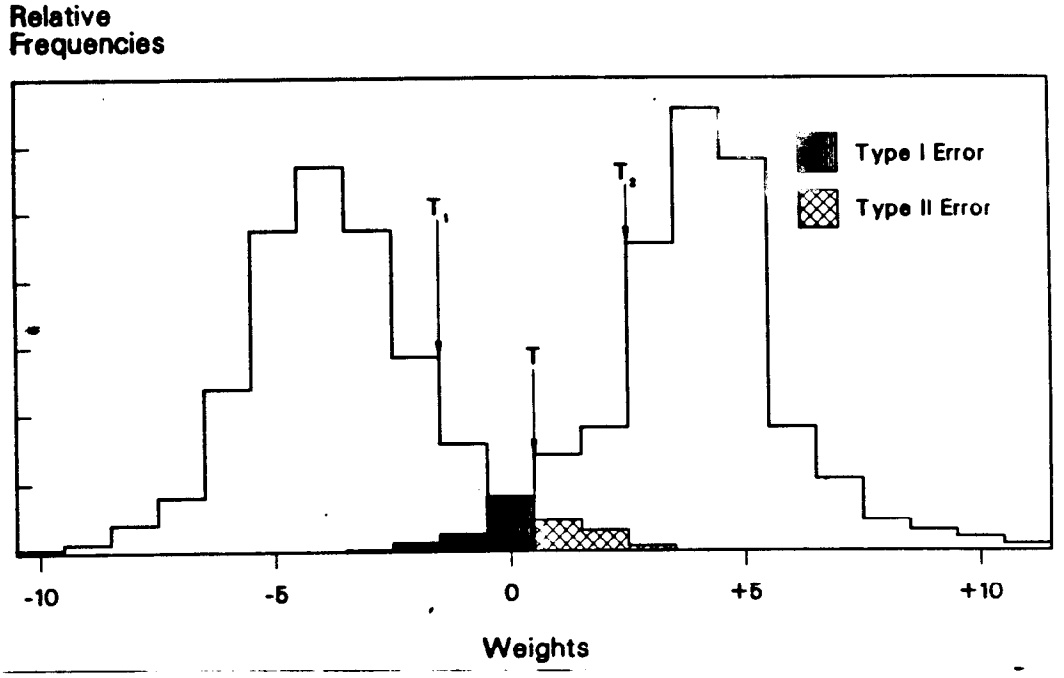
In practice, the luxury of this undecided area was not allowed and only one cut value, T, which was called the threshold was used. This threshold should be set on the basis of the cost function for the application, i.e. the cost of Type I error relative to a Type II error, and the average size block produced by the earlier file blocking. Specifically, the conditional probability that a comparison configuration, γ, is a valid match, is related to L and average block size, B, as follows:

$$P(M|\gamma) = \frac{1 + L}{B + L} \qquad (4)$$

This formulation assumes that each block does indeed contain the record which should be a valid link to the input. Say that our cost/risk computations had convinced us we could afford to link if the odds for a true match were 3:1 or better. Solving the equation above, we would conclude that condition is met (remember B = 2) whenever $L \geq 2$, that is, $\log_2 L \geq 1$. That provides us with a threshold of T = 1.

Note that there is a region between T_1 and T_2 where no decision would be made to favor either M or $\bar{M}$. When this non-decision area is allowed, it is possible to set T_1 or T_2 so that stipulated error levels of Type I and Type II will not be exceeded. A Type I error is committed in rejecting M when it is in fact the case, and a Type II error is in accepting M when it is not the true state of nature.

Figure 1
Thresholds



In practice, the luxury of this undecided area was not allowed and only one cut value, T, which was called the threshold was used. This threshold should be set on the basis of the cost function for the application, i.e. the cost of Type I error relative to a Type II error, and the average size block produced by the earlier file blocking. Specifically, the conditional probability that a comparison configuration, γ, is a valid match, is related to L and average block size, B, as follows:

$$P(M|\gamma) = \frac{1 + L}{B + L} \qquad (4)$$

This formulation assumes that each block does indeed contain the record which should be a valid link to the input. Say that our cost/risk computations had convinced us we could afford to link if the odds for a true match were 3:1 or better. Solving the equation above, we would conclude that condition is met (remember B = 2) whenever $L \geq 2$, that is, $\log_2 L \geq 1$. That provides us with a threshold of T = 1.

The comparison between the input and record 1 of the block meets the required threshold (total weight) value of +1, but the comparison with record 2 does not. Our decision is to link record 1 and its associated geographic codes with this input.

The disagreement between street types might give some pause, but recall that these are the only two explicit candidates; the alternative is to leave the record uncoded. If our computations were valid, we will be correct three times out of four in taking such a risk, and even more often if we observe higher totals than +1.

CHRONOLOGY AND SYSTEM VARIATIONS

1963 Economic Census System

We have already touched on this experience in passing. The major development in automating the geographic coding of postal addresses was accomplished during the 1963 Economic Censuses. An Address Reference File consisting of street segment records within census tracts was constructed for areas serviced by post offices located in cities of 25,000 or more population. In addition, a Building Reference File was developed as an adjunct to the Address Reference File for coding non-street name/house number type of addresses such as motels, hotels, buildings, etc.

In the unscrambler, each word in the city-state portion was identified, standardized and placed in the proper field of the formatted record. The computer then moved to the house number-street name portion and identified street type, direction, building type, house number and street name. The street types, directions and building types were standardized and all components were transferred to become part of the reformatted record. The logic for identifying the address components in the computer process involved a left orientation and recognition of each character on an address line, including such symbols as hyphens, periods, commas and spaces. The program accepted letters and numbers but discarded commas and periods in order to form separate words and identified each space between groups of characters as distinguishing the end or beginning of a word or component.

In the coding phase, the standardized and reformatted input address was matched to its corresponding street segment record in the Address Reference File or a Building Reference File. The 1963 rules for matching required that two words, street name and house number within a post office and postal zone match those of the Address Reference File. If the street name matched and the house number range was unique (i.e., it did not duplicate or overlap other house number ranges for the same street name of different records in the reference file) the computer assigned the geographic codes.

Failure to meet the condition of a unique identification resulted in the computer application of a point system. As the

computer compared each word on the street portion of the input address to the Address Reference File record, a point was assigned for match of each of the following components: (1) primary street type, (2) secondary street type, (3) primary direction, (4) secondary direction, (5) postal zone, and (6) side of street (odd or even). The record in the Address Reference File that received the highest number of points was used to assign the geographic codes. In case of a tie the input address was rejected for manual coding. In case of building name type addresses, the building name was matched to the Building Reference File and the house number and street name were transferred to the incoming record and then run against the Address Reference File for actual coding.

1967 Economic Census System

The Address Reference File for the 1967 Economic Censuses was modified to incorporate the ZIP code information and to include building names, thus eliminating a separate building name file. In addition, the Address Reference File was expanded to include street segments for all cities down to 2500 population size.

The computer logic of identifying, standardizing and formatting of input addresses was practically the same as that used in the 1963 operation described above.

The scoring system for matching components between the address and those of the reference file was improved over that used in the 1963 processing along the information theoretic lines shown in the prior section on weighting. Over 4,000,000 records were processed and 70 percent of the addresses were linked and coded. This is approximately the same proportion as in the earlier 1963 experience. However, Type I error was reduced by nearly 90 percent and the Type II error by over 80 percent.

1968 Developments in the New Haven Census Use Studies

As part of the test procedures carried out in the New Haven test census, an Address Coding Guide and DIME (Dual Independent Map Encoding) file were developed for the SMSA. Both of these files consisted of street segments (i.e., street name and house number ranges for block sides). The DIME file contained added information on intersection identification, block pairs and direction, necessary for computerized mapping. A generalized set of computer programs was also developed to facilitate the linkage of local addresses to either of the reference files to automatically code the location to block face, block, etc.

In creating the linkage programs (titled ADMATCH) for the New Haven applications, use was made of the 1963 and 1967 developments. The ADMATCH system also consisted of two basic programs, namely, (1) an unscrambler and (2) a linkage of the input address to the reference file record.

A feature of this unscrambler was the availability of street name variants. The user could opt to have street names of incoming addresses compared to all street name variants stored in the unscrambler. If the address contained a variant spelling of a given street which was stored in the unscrambler, this program inserted the standard street name for use in the linkage program.

The linkage or matching program was simpler than the unscrambler and also had options for the users application. These options ranged from complete and exact match to street name only match. Under the most stringent option, the coding rate for addresses from the various local administrative records was generally in the range of 70 to 80 percent. In those cases where all components of the incoming address exist, the coding resulting from the linkage operation was almost error free. As the level of requirements for determining a match was reduced, coding rate rose, but the degree of equivocation increased.

As ADMATCH was designed for local use, it is somewhat superior in that application, but for national scale use the Economic Census Geocoder is much more practical. In a broader test (called CAMEL) which took place in 1975-1976, and using a commercially acquired mailing list against the GBF/ DIME file, ADMATCH strict option coded 81 percent of the records with almost nil error, against 92 percent coding and a one percent error from the Economic Census Coding system. Six percent of the records in that input file were not even codable clerically.

Unimatch

At least passing note should be given to another matching system developed about 1971 at the Bureau, UNIMATCH. It was designed as a laboratory tool to simulate existing or contemplated matching algorithms on an IBM computer. It is so general in nature that it has emulated the performance of both ADMATCH and the Economic Census system quite accurately, but UNIMATCH can also do person-matching or many other types of matching. There is a standardizer companion called UNISTAN which we lack space to discuss here.

1972 and 1977 Economic Censuses

Pretests were run for the 1972 and 1977 Censuses and modifications in weights were made for the new files. There was something of a swing toward risk taking in 1972 coding and away from it for 1977. The coding was about 80 percent in 1972 with a 5 percent error. Also, a string comparator technique on Post Office name was introduced in 1976.

Recent Decennial Censuses

The block face reference file developed specifically for use in the 1970 Census was called the address coding guide (ACG). The

same contractor who provided the ACG supplied also the basic address mailing list with identical numerical codes across files for street names. Linking was quite straight-forward and did not require a probabilistic scheme as described in this paper.

The 1980 Decennial Census did not use the same files that had been used in 1970, so such direct numeric links were not feasible. However, deterministic coding was the policy. Address records which did not exactly match the new GBF/DIME file were clustered by common attribute and recycled (somewhat as in ADMATCH) after reference file or input had been perfected. Some addresses were recycled several times but the end quality of the data was high.

Because of both the quantity and sources of records involved and the requirement for fine level coding, the Decennial Census has chosen not to accept the risk of probabilistic computer coding.

CONCLUSIONS AND OBSERVATIONS

The basic technology for automatically linking street name and house number type addresses to geographic base files has been developed to a fair degree of sophistication. The programs give the appearance of making intelligent choices and produce outcomes comparable to those that trained human coders accomplish.

It is sometimes said that the Japanese are not original or creative in the fields of cameras, automobiles, and computers; they merely apply existing theory. That may be so, but many of us choose Japanese products in these areas. Similarly, no new theory has been developed for this application. It has been borrowed selectively from information theory, decision theory (or hypothesis testing if you are a statistician instead of an operations researcher), probability theory, and even phonetics. Nonetheless, the result is a useful one, even in other matching problems where it has been tried, such as person matching.

Our studies have repeatedly shown that the major number of both uncoded records and falsely coded records are explained by inadequacies of the reference and input files. We do not expect decisive improvement in either standardization or linking programs.

REFERENCES

(1) Dolleck, Sol, and Herman H. Fasteau. "Computerized Geographic Coding," Data Processing Magazine, Vol. 8., #10, October 1966, pp. 40-43.

(2) Hewes, William L. and Kent H. Stow. "Information Retrieved by Proper Name," Data Processing Magazine, 7, No. 6, June 1965, pp. 18-22.

(3) Jaro, Matthew A. "UNIMATCH - A Computer System for Generalized Record Linkage under Conditions of Uncertainty," Proceedings of the Spring Joint Computer Conference, 1972, pp. 523-530.

(4) Newcombe, Howard B. and James M. Kennedy. "Record Linkage," Communications of ACM, Vol. 5, #11, November 1962, pp. 563-566.

(5) U.S. Department of Commerce, Bureau of the Census. ADMATCH Users Manual (Census Use Study). Washington, D.C.: U.S. Government Printing Office, January 1970.

(6) U.S. Department of Commerce, Bureau of the Census. The Census Bureau's GBF/DIME System: A Tool for Urban Management and Planning. Washington, D.C.: U. S. Government Printing Office, September 1980

(7) U.S. Department of Commerce, Bureau of the Census. GBF/DIME System Description and Uses. Washington, D.C.: U.S. Government Printing Office, 1978.

(8) U.S. Department of Commerce, Bureau of the Census. Geographic Base (DIME) File System [illegible] (GBF-80-5). Washington, D.C.: U.S. Government Printing Office, 1974.

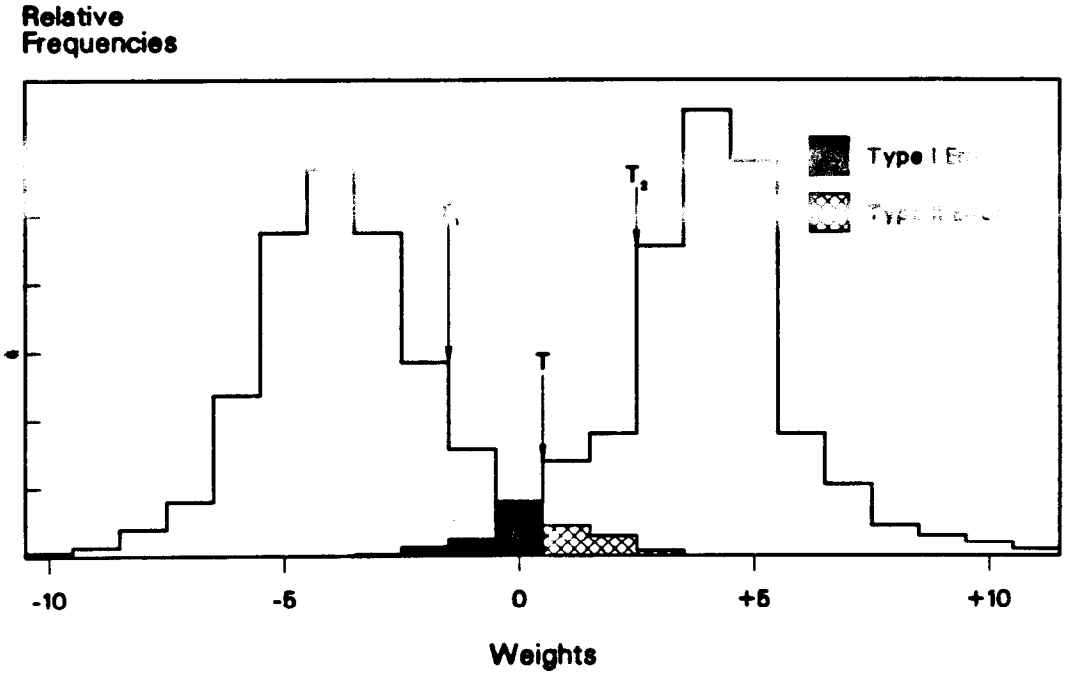
(9) U.S. Department of Commerce, Bureau of the Census. OS ADMATCH: An Address Matching System (Census Use Study). Washington, D.C.: U.S. Government Printing Office, 1972.

(10) U.S. Department of Commerce, Bureau of the Census. The Uses of GBF/DIME (Census Use Study Report No. 15). Washington, D.C.: U.S. Government Printing Office, 1974.

(11) U.S. Department of Commerce, Bureau of the Census. The Census Bureau, A Numerator and Denominator for Measuring Change (Technical Paper 37). (Washington, D.C.: U.S. Government Printing Office, June 1974).

Note that there is a region between T_1 and T_2 where no decision would be made to favor either M or $\bar{M}$. When this non-decision area is allowed, it is possible to set T_1 or T_2 so that stipulated error levels of Type I and Type II will not be exceeded. A Type I error is committed in rejecting M when it is in fact the case, and a Type II error is in accepting M when it is not the true state of nature.

Figure 1
Thresholds



In practice, the luxury of this undecided area was not allowed and only one cut value, T, which was called the threshold was used. This threshold should be set on the basis of the cost function for the application, i.e. the cost of Type I error relative to a Type II error, and the average size block produced by the earlier file blocking. Specifically, the conditional probability that a comparison configuration, γ, is a valid match, is related to L and average block size, B, as follows:

$$P(M|\gamma) = \frac{1 + L}{B + L} \qquad (4)$$

This formulation assumes that each block does indeed contain the record which should be a valid link to the input. Say that our cost/risk computations had convinced us we could afford to link if the odds for a true match were 3:1 or better. Solving the equation above, we would conclude that condition is met (remember B = 2) whenever $L \geq 2$, that is, $\log_2 L \geq 1$. That provides us with a threshold of T = 1.