COMPARATIVE ANALYSIS OF RECORD LINKAGE DECISION RULES

William E. Winkler, Rm 3000-4, Bureau of the Census, Washington, DC 20223

Keywords: EM Algorithm, Loglinear, Dependence, String Comparator Metric

This paper provides an empirical comparison of decision rules in the Fellegi-Sunter model of record linkage. Using files for which true linkage status is known, the results of applying various parameter-estimation/decision-rule strategies for designating links and nonlinks are compared. The Expectation-Maximization Algorithm provides estimates of parameters for loglinear models of latent classes in situations where the underlying probability distributions of agreements on identifiers such has surname, house number and age satisfy a conditional independence assumption and in situations where more general interactions are allowed.

1. INTRODUCTION

This paper describes methods of estimating probability distributions for latent class models and applications of associated record linkage decision rules. The best previous decision rules were generally based on extensive modelling using training sets for which true matching status was known. The best method of this paper does not require training sets and may allow automatic creation of optimal decision rules for relatively extensive sets of files.

Smith and Newcombe (1975) first observed that, when linking files of individuals, agreement on family (or household) identifiers such as last name, house number and street name is not necessarily independent of agreement on individual identifiers such as first name, age and sex. When all identifiers were used in the basic decision rule (e.g., Newcombe et al. 1959, Newcombe 1988), then different persons (from a pair of records from the two files) in the same household could get weights (or scores) that are higher than weights of a single person (again from a pair of records from two files) that is listed as residing in two different locations. Their solution (which improved on the basic decision rule) was to develop a scoring and decision rule mechanism that created scores based on family and individual identifiers independently and then to combine the two scores in the decision rule that determined which pairs were links.

An alternative to the Smith-Newcombe procedure is to have a general fitting mechanism that estimates underlying probability distributions when agreements on identifiers are dependent. In such situations, one weight would result and the original decision rule of Newcombe (which had been shown to be optimal by Fellegi and Sunter 1969) could be applied. The advantages of the general methods are that they would work with arbitrary sets of identifiers and would automatically determine dependencies where none were suspected to exist.

The outline of the paper is as follows. The first section gives background on record linkage and Expectation-Maximization (EM, e.g., Dempster, Laird, and Rubin 1977, Haberman 1975, 1979) techniques for estimating parameters. For latent class loglinear models, appropriate EM references are Winkler (1989) and Meng and Rubin (1992). The second section describes the various models and decision rules for which empirical results are provided. While the EM fitting procedures are applied to all pairs, the decision rules force 1-1 matching (i.e., each record can be matched with at most one other) using a linear sum assignment procedure (Jaro 1989). In the third section results about the estimated probability distributions and associated decision rules are presented. The fourth section discusses why the fitting procedures that yield estimated probability distributions that are quite close to the true distributions do not necessarily yield the best decision rules. The last section is a summary.

2. BACKGROUND

2.1. Fellegi-Sunter Model of Record Linkage

The record linkage process attempts to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files A and B into M, the set of true links, and U, the set of true nonlinks. Making rigorous concepts introduced by Newcombe (e.g., Newcombe et al. 1959), Fellegi and Sunter (1969) considered ratios of probabilities of the form:

$$R = Pr(\gamma \in \Gamma \mid M) / Pr(\gamma \in \Gamma \mid U) \qquad (2.1)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on surname, first name, and age. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific surnames, such as Smith or Zabrinsky, occur.

The decision rule is given by:

If $R >$ UPPER, then designate pair as a link.
If LOWER $\leq R \leq$ UPPER, then designate pair as a possible link and hold for clerical review. (2.2)
If $R <$ LOWER, then designate pair as a nonlink.

Fellegi and Sunter (1969, Theorem) showed that the decision rule is optimal in the sense that for any pair of fixed upper bounds on the rates of false links and false nonlinks, the clerical review region is minimized over all decision rules on the same comparison space Γ. The cutoff thresholds UPPER and LOWER are determined by the error

bounds. We call the ratio R or any monotonely increasing transformation of it (such as given by a logarithm) a matching weight or total agreement weight.

In actual applications, the optimality of the decision rule (2.2) is heavily dependent on the accuracy of the estimates of the probabilities given in (2.1). The probabilities in (2.1) are called matching parameters. One way to evaluate how accurately estimates of probabilities in (2.1) agree with the truth is to plot the cumulative estimated conditional probabilities given links and nonlinks against the corresponding cumulative conditional probabilities based on the truth. The cumulation is by the ordering of the ratio R in (2.2) induced by the estimated probabilities.

2.2. Expectation-Maximization Algorithm

For each $\gamma \in \Gamma$ and each pair of subsets C_1 and C_2 that partition $\mathbf{A} \times \mathbf{B}$ consider

$$P(\gamma) = P(\gamma \mid C_1)\, P(C_1) + P(\gamma \mid C_2)\, P(C_2), \qquad (2.3)$$

where C_1 and C_2 might represent M and U, respectively. We can observe the proportion of pairs having representation $\gamma \in \Gamma$. If γ represents a simple agree/disagree pattern, we can estimate the probabilities on the right hand side via the Expectation-Maximization (EM) Algorithm (see e.g., Dempster, Laird, and Rubin 1977).

If we assume that agreements on different characteristics are conditionally independent within C_1 and C_2, then the maximization step is in closed-form and the EM Algorithm is quite straightforward to apply (Jaro 1989). More generally, to account for dependencies of the agreements of different matching fields (e.g., Thibaudeau 1989), we apply a variant of an algorithm of Haberman (1975, see also Winkler 1989). As there are ten matching variables, we only have sufficient degrees of freedom to fit all 3-way interactions (see e.g., Bishop, Fienberg, and Holland 1975, Haberman 1979).

We also partition $\mathbf{A} \times \mathbf{B}$ into three sets of pairs C_1, C_2, and C_3 using an equation analogous to (2.3). The EM procedures are then divided into 3-class or 2-class procedures. When appropriate, two of the three classes are combined into either a set which represents M or U with the remaining class representing its complement.

For probabilities computed under the independence assumption and with data from record linkage settings, the 2-class EM Algorithm typically converges to a unique limiting solution over a wide range of plausible starting points (Thibaudeau 1989, Winkler 1989). Our experience is that the independent, 3-class EM also converges to a unique limiting solution which we, in turn, use as the starting point for the 3-class, 3-way interaction EM.

The disadvantage of any of the EM procedures is that they may divide $\mathbf{A} \times \mathbf{B}$ into two sets that differ significantly from the desired sets of links M and nonlinks U.

An enhancement to the basic EM procedures is to put additional convex (affine) constraints on some of the conditional probabilities and proportions to assure that the solutions are closer to the known true values. For instance the proportion in the first class of the 3-class model might be bounded above by 0.1 or the probability of disagreeing on first name conditional on being in the first class might be bounded above by 0.05. The intuitive idea of applying convex constraints is that the EM procedures might be given a predisposition toward placing certain pairs into different classes based on prior knowledge of the characteristics of links and nonlinks.

2.3. Decision Rules

For comparative purposes, we consider several elementary decision rules and a variety of increasingly more complicated rules for designating links and nonlinks. For every decision rule we force 1-1 matching via a linear sum assignment procedure introduced by Jaro (i.e., each record from one file can be linked with at most one record from another) and we use string comparator metrics (see e.g., Jaro 1989, Winkler 1990). The reason we consider 1-1 matching methods is that they can dramatically lower the size of the clerical review region. For instance, within a household say, the father-father, mother-mother, son-son, and daughter-daughter pairs might be kept. The remaining twelve pairs (which might be clerically reviewed if 1-1 matching were not forced) might be designated as nonlinks.

As the rate of typographical error for identifiers (e.g., Smith versus Smoth) among links is quite high for the empirical files of this paper (25% of first names and 15% of last names, Winkler 1990), we use the string comparators to get better decision rules in those cases where identifers agree almost exactly but not exactly. For the best decision rules we also incorporate frequency-based (value-specific) enhancements that account for the relative frequency of occurrence of specific fields such as last name or first name (Fellegi and Sunter 1969). In all cases the relative frequency weights are precisely scaled to the basic yes/no agreement weights obtained from the fitting software or by guesses.

The basic rules are listed in Table 1. For the independence model, the rules essentially differ in how the marginal probabilities are obtained. The rules listed under em-methods get probabilities via the em algorithm. The rule listed under the stc-method uses the marginal probabilities based on the known truth. In the ideal situation such iterative fitting methods that require extensive human intervention at intermediate steps as used in Statistics Canada's matching system GRLS-V2 (Hill 1992) could yield the true marginal probabilities.

Table 1. Decision Rules

Independent
(1) 3-class, em, freq replace yes/no probablities in class 1 for first and last name with relative frequency ones.
(2) stc, freq use m- and u-probs based on truth, relative frequencies for first and last name.
3-Way Interaction
(3) 3-class, em, double double count incremental distinguishing power of first and last name.
(4) 3-class, convex same as (4) but apply additional convex constraints.

For 3-way interaction models, we apply rules that use basic probabilities from EM procedures, that use an enhancement that double-counts the incremental distinguishing power of first and last name, and that fit with additional convex constraints. By incremental distinguishing power, we mean the conditional probability of agreement on first or last name in class one given the probabilities associated with the remaining variables. The additional convex constraints restrict the proportion of pairs in the first of the three classes to be less than 0.088 and restrict the probability of disagreement on first name given the pair is in class one to be less than 0.07.

2.4. Matching Fields and Data Files

The ten fields available for matching are the six individual identifiers: first name, age, sex, marital status, relationship to head of household, and race, and the four family (or household) identifiers: last name, house number, street name, and telephone number.

The file sizes are approximately 12,000 and 15,000. Slighty less than 9,900 pairs of records are true links and are identified in the files. The identification was based on extensive manual review and field followup for a set of blocks in St Louis, Missouri.

3. RESULTS

3.1. Distributions

Results of fitting with either two or three classes under various independence and interaction assumptions show that the basic 3-class, 3-way interaction model gives by far the best fit (Table 2). Approximate chi-squares values are computed according to Haberman (1979, p. 562) and Z-values via normal approximation.

The estimated cumulative distribution conditional on links (Figures 1, 2, and 3) is much closer to the truth when the 3-class, 3-way interaction model is used than when various independence models are used. The corresponding curves for nonlinks (not shown) also yield that the 3-class, 3-way interactions model gives the best fit.

Table 2. Chi-Square Fits, Degrees of Freedom, and Z-Values under Various Models

	Chi	DOF	Z
2-class Independent	55796	1002	1224.0
3-class Independent	26987	991	584.0
3-class 3-way	517	495	0.7
3-class 3-way, convex	501	468	1.3

If we were to consider all pairs (rather than the subset obtained under the 1-1 matching restraint), then the 3-class, 3-way interaction probabilities would yield the best decision rules and reasonably accurate estimates of error rates.

The last fit illustrates how fits degrade when convex constraints are imposed that cause only mild deviations from the fit under no constraints. While convex fits can yield fits in different classes that are closer to the true proportions and most of the true probabilities, they have not yet improved the decision rules and are not considered further.

3.2. Decision Rules under 1-1 Matching

The decision rules ((1)-3-class independent EM, (2)-independent with margins based on truth, & (3)-3-class, 3-way EM) are roughly equivalent at error levels of 0.005 and above (Table 3). At the error level of 20 false links (0.2 percent false link rate), rules (1), (2), and (3) designate 9808, 9813, and 9601 pairs as links, respectively and, at error level 50 false links (0.5 percent false link rate), the rules designate 9875, 9881, and 9802 pairs as links, respectively.

If we adjust the probabilities of section 3.1 to the subsets of pairs considered by the 1-1 matching rules (the subsets are dependent on the weight estimates), then the 3-class, independent em-probabilities still deviate quite substantially from the truth (Figures 4 and 6). While the 3-class, 3-way interaction probabilities generally deviate from the truth for links (Figures 5), they remain very close to the truth at error levels of less than 10%. They also are reasonably close to the truth for nonlinks (Figure 7).

Table 3. Number of Pairs by Matching at Different Levels of False Links
1-1 Matching, Frequency-Based

# False	Number of Pairs em indep (1)	stc indep (2)	em 3-way (3)
12	9760	9751	9290
20	9808	9813	9601
30	9829	9836	9709
40	9864	9860	9771
50	9875	9881	9802
70	9909	9911	9851
100	9950	9954	9908
120	9976	9975	9937
150	10006	10006	9976
1/	9859	9859	9861

1/ Highest number of true links achievable with the 1-1 matching restriction.

4. DISCUSSION

4.1. Pairs Via 3-class and 2-class EM Procedures

When 2-class EM procedures are applied to pairs of files having both individual identifiers such as first name, age, and sex and family (household) identifiers such as last name, house number, and street name, the set of pairs naturally divides into those agreeing on household identifiers and those that do not.

The 3-class EM creates a more natural partitioning because it basically divides the set of pairs into (1) links within the same household, (2) nonlinks within the same household, and (3) nonlinks outside the same household. For the decision rules, the probabilities associated with classes (2) and (3) are combined to yield the probabilities for the set of nonlinks U.

4.2. Decision Rules

On an absolute basis, the best two decision rules that require neither knowledge of the truth nor training sets ((1)-3-class independent EM and (3)-3-class, 3-way interaction EM) both work very well. At the 0.5% false link rate rule (1) yields 99.7% of the true links ((9875-50)/9859) while rule (3) yields 98.9% of the true links ((9802-50)/9861)).

The best 3-way interaction model works slightly worse than the best of the independence models because the 3-way interaction model places two additional types of pairs in Class C_1 that the independence model does not. The first type basically consists of husband-wife pairs which agree on age and which agree on a miskeyed sex code. The second type includes father-son pairs that agree on name. The fields that best allow us to distinguish individuals within a household are first name, age, and sex. Each of these two types of pairs agree on two of the three fields. If 30 pairs of these two types (representing approximately 0.03% of the 116,305 pairs used in the em-estimation procedures) were shifted from Class C_1 to Class C_2, then rule (3) would work as well as rule (1).

At present, we suspect more careful modelling using selected subsets of interaction terms greater than the third order or careful application of combinations of convex constraints will yield an improvement to rule (3). The disadvantage of using selected subsets of the higher order interaction terms is that such modelling is quite difficult in the best of circumstances (Bishop, Fienberg, and Holland 1975) and may be somewhat specific to the pairs of files being matched. The advantage of the convex constraints is that they can easily be applied based on prior matching situations. For instance, we might restrict the probability of simultaneous disagreement on first name and sex in the first class to be less than 0.01.

4.3. Probability Distributions

When decision rules that do not force 1-1 matching are applied, probability distributions obtained under the 3-class, 3-way interaction models are sufficiently accurate that they can be used to estimate true error rates. They have the additional intuitive feature that if a pair agrees on house number and street name, then the incremental weight associated with last name is very small. If the pair disagrees on household components such as house number and street name, then the incremental distinguishing power of last name is quite large.

Error rates for decision rules based on probability distributions estimated under the independence assumption (whether or not 1-1 matching is forced) are sufficiently inaccurate that they are unusable (e.g., Figures 1, 2, 4, and 6). For estimating error rates in the independent case when an 1-1 decision rule with good distinguishing power is used, we would use a method of Belin and Rubin (1991). Unlike the EM modelling method of this paper, the Belin-Rubin method generally requires that a representative training set be available for modelling certain parameters in their model. For the weights arising from the 3-class, 3-way interaction model (whether or not 1-1 matching is forced), Belin-Rubin fitting software will not always converge due to the fact that the curve of natural logarithm versus weight is not clearly bimodal as it was in other applications for which the software had been developed. We note that the Belin-Rubin

procedures use only the summarizing information contained in the matching weight while the 3-class, 3-way interaction models use all the information from the various agreement patterns.

4.4. Different Starting Points for 3-Way Interaction Model

Using a variety of different starting points, there appear to be at least three local maxima, all of which give the same value of the likelihood function to five significant digits. Among the limiting solutions, the proportions assigned to different classes vary substantially and the conditional probabilities generally vary.

If starting points were chosen fairly close to the solution of the 3-class, independent model (which appears to yield unique solutions), then convergence was always to the first local maxima. We note that solutions can, at best, be unique up to permutation (Haberman 1979, Chapter 10).

4.5. General Applicability of Methods

The use of the 3-class, independent EM procedure should generally yield good results in the type of 1-1 matching decision rules employed in this paper. In four other pairs of files of individuals in which household identifiers were present and for which true matching status known, the 3-class, independent EM performed at least as well it did for the pair of files in this paper.

5. SUMMARY

This paper considers methods for matching individuals using a combination of individual identifiers such as first name, age, and sex and family identifiers such as house number and street name. At present, the best decision rules (based on either false link rate or size of clerical review region) allow at most one individual from one file to be matched against one from another (i.e., force 1-1 matching) and use probability distributions that satisfy an independence assumption.

If all pairs are considered (i.e., 1-1 matching is not forced), then the best decision rules use probability distributions fit under 3-way interaction models and allow estimation of error rates. When 1-1 matching is forced, however, the 3-way interaction decision rules perform somewhat worse than the best of the independent rules. For 1-1 matching 3-way interaction probabilities for all pairs can be adjusted to yield reasonably accurate estimates of error rates but independent probabilities can not.

*This paper reports views of the author that do not necessarily represent those of the Bureau of the Census. The author thanks Yves Thibaudeau, Carl Konshnik, and Phillip Steel of the Bureau of the Census and Michael Larsen of Harvard University for comments.

REFERENCES

Belin, T. R. and Rubin, D. B. (1991) "Recent Developments in Calibrating Error Rates for Computer Matching," Proc. of the 1991 Census Annual Research Conf., 657-668.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), Discrete Multivariate Analysis, MIT Press, Cambridge, MA.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. Royal Stat. Soc. B, 39 1-38.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," Journal of the American Statistical Association, 64 1183-1210.

Haberman, S. J. (1975), "Iterative Scaling for Log-Linear Model for Frequency Tables Derived by Indirect Observation," Proceedings of the Section on Statistical Computing, American Statistical Association, pp. 45-50.

Haberman, S. (1979), Analysis of Qualitative Data, Academic Press. New York.

Hill, T. (1991), "GRLS-V2, Release of 22 May 1991," (Available from General Systems, R.H. Coats Bldg, 14-O, Statistics Canada, Ottawa, Ontario K1A 0T6.)

Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," Journal of the American Statistical Association, 89, 414-420.

Meng, X. and Rubin, D. B., (1992) "Maximum Likelihood Via the ECM Algorithm: A General Framework," Biometrika, to appear.

Newcombe, H. B. (1988) Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business. Oxford: Oxford Univ. Press.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," Science 130 954-959.

Smith, M. E., and Newcombe, H. B. (1975), "Methods of Computer Linkage of Hospital Admission-Separation Records into Cumulative Health Histories," Meth. Inform. Med. 14 118-125.

Thibaudeau, Y. (1989), "Fitting Log-Linear Models When Some Dichotomous Variables are Unobservable," in Proceedings of the Section on Statistical Computing, American Statistical Association, pp. 283-288.

Winkler, W. E. (1989), "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage," Proceedings of the Fifth Census Bureau Annual Research Conference, 145-155.

Winkler, W. E. (1990), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," Proc. of the Section on Survey Research Methods, Amer. Statistical Assoc., pp. 354-359.

Figure 1. Estimates vs Truth
Cumulative Distribution of Matches
1– Independent EM

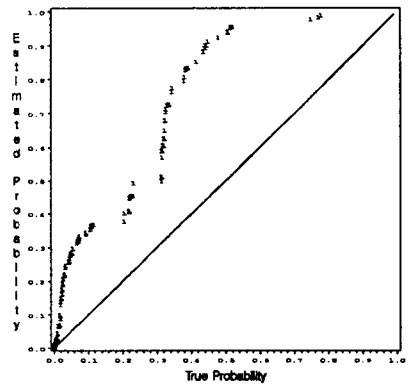


Figure 2. Estimates vs Truth
Cumulative Distribution of Matches
2– Independent, Statistics Canada

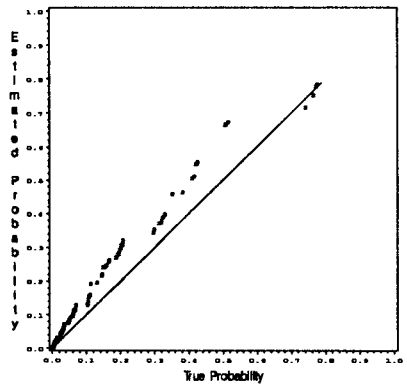


Figure 3. Estimates vs Truth
Cumulative Distribution of Matches
3– 3–way Interaction EM

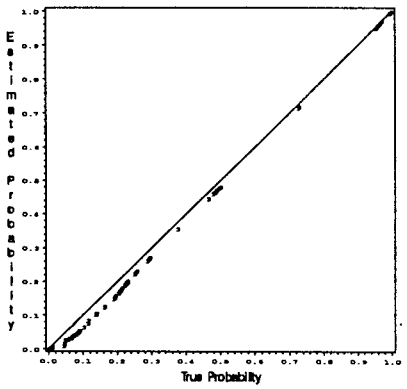


Figure 4. Estimates vs Truth
Cumulative Distribution of Matches
1– Independent EM, 1–1 Matching

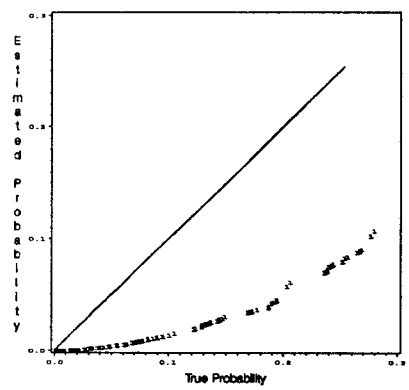


Figure 5. Estimates vs Truth
Cumulative Distribution of Matches
3– 3–way Interaction EM, 1–1 Matching

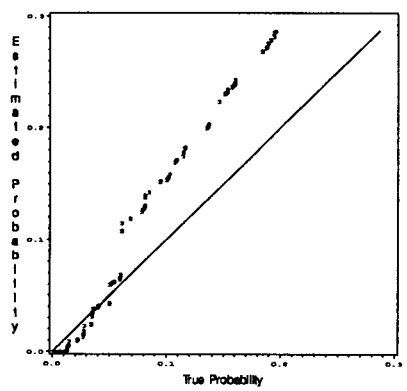


Figure 6. Estimates vs Truth
Cumulative Distribution of Nonmatches
1– Independent EM, 1–1 Matching

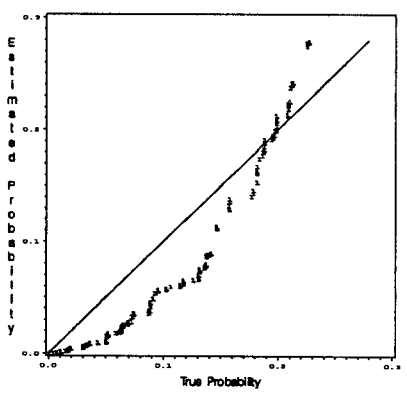
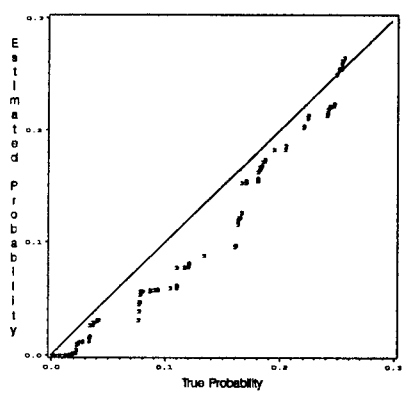


Figure 7. Estimates vs Truth
Cumulative Distribution of Nonmatches
3– 3–way Interaction EM, 1–1 Matching



Cumulation is by decreasing estimated weight
Small probabilities correspond to small error rates
45 degree line as reference