

Probabilistic Record Linkage: A SAS® OOP Solution

Gary W. Griffin, Whitehurst Associates Inc., Atlanta, GA

ABSTRACT

The need to match records from two separate files with no unique identifiers, such as social security number, is a problem often encountered in both public and private organizations. One solution to that problem is to use probabilistic record linkage techniques. The basic process involves calculating a probability for each comparison made with information from the record pairs and then accumulating those probabilities to determine the overall odds for a given pair of records. The odds provides evidence which can be used to make decisions in favor of or against the fact that the pair of records identifies the same individual, business, or other entity. In this presentation, I will discuss the development of a PC based probabilistic record linkage system using the SAS System's object oriented programming paradigm. This system exploits the SAS System's Frame technology and OOP to provide users with a point and click graphical interface which simplifies an otherwise complicated and somewhat abstract process.

INTRODUCTION

In the late nineteen fifties, Dr. Howard Newcombe pioneered a technique for matching together record pairs which identified the same entity when neither record contained a unique identifier. Today that technique is known as Probabilistic Record Linkage and its importance has been realized in both the public and private sectors. State and Federal public health agencies use it for surveillance of select target populations of interest. Marketing departments in national and multi-national corporations use the technique to match their customer databases to other data sources thereby further defining marketing segments. In sum, Probabilistic Record Linkage provides a powerful, unique methodology that takes two unrelated secondary data sources and creates a primary data source. This new data source provides additional information about individuals, families, and/or businesses that would otherwise be unknown.

With so much interest in this technique, to say that developing a probabilistic record linkage system is not a new concept should not be a surprise. Although systems have been developed specifically for the purpose of probabilistic record linkage, they appear to suffer from two primary weaknesses. First, these systems do little else. I propose that a probabilistic record linkage system not only

needs to do record linkage, but it should also provide you with other capabilities such as data management, data security, and evaluation tools. Second, while probabilistic record linkage is an extremely powerful tool, it is also a very complex process that requires considerable skill and knowledge. Current systems do not simplify this process in any significant way for end users.

Thus, my purpose in writing this paper is twofold. First, I will discuss some of the important capabilities that are needed for a more COMPLETE probabilistic record linkage system and how the SAS System is ideally suited as a development tool to provide these enhancements. Second I will present DATALINK 1000. DATALINK 1000 is a PC based probabilistic record linkage system which I designed and developed using the SAS System. This system has the extra capabilities described in this paper.

PROBABILISTIC RECORD LINKAGE AND SAS

A theoretical model that outlines the components of a probabilistic record linkage system is presented in Figure 1. This model details the flow of data from its point of origin (i.e. its original data source) to its ultimate destination which is the consumption of the information by interested users. This model also demonstrates why the SAS System is ideally suited as a development tool for a record linkage system.

As shown in the model, the starting point of probabilistic record linkage involves bringing together at least two data sources. One of the major strengths of a complete record linkage system is the ability to read data and link it no matter what format the data reside in. Current record linkage systems require the data to be in an ASCII format in order to use the system. If the data are in an ASCII format, then SAS makes it easy to read and use. However, this need not be the case. In SAS, you do not have to be concerned with the format of the data which you are intending to use in the record linkage system. Databases such as Oracle, Sybase, and DB2 can be easily accessed by using SAS/ACCESS® software.

Once two data source are linked, some type of DBMS is needed as a repository of the linked data. Linked databases should be easily accessed and managed. SAS data sets are an excellent tool for this task.

The importance of Probabilistic Record Linkage is not in the process itself, but the result of the process (i.e. a new primary data source). It is from this new data source that you can begin to formulate questions and gain invaluable information which would otherwise be unavailable. Of course, to get at this information, questions must be asked and the data must be analyzed as shown in Figure 1. Again, the SAS System delivers an abundance of analytical tools to tackle this task. From descriptive statistics to multivariate statistical modeling, SAS/STAT® software is some of the best in the world.

The final component of the record linkage system is to put information extracted from the analyzed data in presentation form for consumption by users. Using tools provided by the SAS System such as the many graphic options of the SAS/GRAPH® software and report generating procedures such as PROC REPORT and PROC TABULATE make this very easy to do.

Multiple Platforms

An additional strength gained from using the SAS System as the development tool for a record linkage system is the ability to port the code to multiple platforms. With transport files, a record linkage system created using SAS can be moved to multiple platforms. Only minor changes to the software would then be required in order for the system to run on a particular platform.

ABOUT DATALINK 1000

DATALINK 1000 is a PC based probabilistic record linkage system designed and developed to run under the Windows operating environment. The system was developed using SAS/AF® software, Screen Control Language, and Frame entries from SAS System versions 6.08 and 6.10.

Features of this system include a graphical user interface which utilizes point and click technology to make record linkage of two separate data sources a simple and easy process to perform, a linked file viewer to assess the validity of matched record pairs, canned reporting procedures for evaluating the success of the linkage, and system utilities which provide data security and data management for system users.

The frame which is displayed for you to perform probabilistic record linkage requires minimal effort of the user. In this screen, you are required to only select the search and master files to be linked, the search and master file variables that are to be used in the linkage, and a linkage type for each pair of corresponding variables from the

search and master files. The final step is to select "Link Data" from the menu and initiate the process. Again, this frame is intended and does provide you with a simple and straight forward ability to link two files with minimal knowledge of record linkage.

The screen which allows you to view linked records provides two basic capabilities. First, you can visually examine linked records with this ability being further refined to allow you to select all records or only duplicate records. Second, after visually inspecting the matched record pairs, you can then delete unwanted linked record pairs before saving the linked file.

The canned reporting consists of two standard reports and one graph. One report provides you with a listing of all linked records and the other gives a listing of duplicate linked records. The graph displays a pie chart with the number and percentage of records for predetermined score categories.

DATALINK 1000 provides security at two different levels. First, you are given system security which checks user access by requiring you to log into the system with a password and a user id. The second level of security is provided by this system by tracking and controlling user access to frames and certain functionality. Access levels are determined by the System Administrator when new users are added to the system security data set. For example, a record linkage transaction can be deleted in the system module. This functionality is obviously not something you want all users to be able to do.

Management of linked data is a vital necessity that is also provided in the DATALINK 1000 system. First, you can save and track each record linkage project. Second, authorized users can delete unwanted/unneeded linkage projects. Third, you can create an analysis database with only the fields that are needed. This particular data management function is very easy to use. In this frame you are required to select a specific linkage project. The variables from the search and master files that were used in this linkage operation are then displayed in list boxes with the variables that were used to match record pairs being highlighted. To create the analysis database, you need only to select those variables from each file that are wanted in the final data set and specify an output data set name. The system then creates the data set.

System Architecture

DATALINK 1000 has two main libraries. The SYSCATS

library which contains all the system catalogs and the SYSDATA library contains all the system data sets.

All frames, SCL entries, system object classes and associated methods are contained in five catalogs -- Datalink, Manager, Opencode, Output, and Pmenus. The Datalink catalog contains all those entries that are involved in the actual record linkage process. The Opencode catalog has Source entries that are DATA step routines needed by the system. The Output catalog handles all the reporting processing. And finally, the Pmenus contains the menus for the entire system.

The DATALINK 1000 system uses the widget classes supplied by the SAS System in developing the graphical interface. However, in order to simplify complex processes by encapsulating behavior such as the record linkage itself, three new object classes were created -- the Linkage Class, the System Class, and the Output Classes. Each class contains all the methods and data necessary to implement the functionality of the class.

The Linkage Class is used specifically for the record linkage. The System Class is used to for system level behavior such as checking user access. The Output Class includes the methods and instance variables needed to produce the canned reports.

An example of how one step of the linkage process was broken down and placed in a method within the Linkage Class is shown below. This method is used for calculating positive global weights.

PGLBWGHT:

```
method;
  classid = symgetn('classid');
  mvlistid = getnitemn(classid,'mvlistid');
  svlistid = getnitemn(classid,'svlistid');
  mfname = getnitemc(classid,'mfname');
  sfname = getnitemc(classid,'sfname');
  sfdsid = getnitemn(classid,'sfdsid');
  mfdsid = getnitemn(classid,'mfdsid');
  numsfrec = getnitemn(classid,'numsfrec');
```

* Construct keep statements to select only needed vars from datasets;

*Master File List;

```
n=listlen(mvlistid);
do I = 1 to n;
  cite = getitemc(mvlistid,i);
  if I = 1 then do;
    mvars=citem;
```

```
end;
  else do;
    mvars = mvars|| ' '||citem;
  end;
end;
mvars = mvars|| ' '||'nobmnum';
```

*Search File List;

```
n=listlen(svlistid);
do I = 1 to n;
  cite = getitemc(svlistid,i);
  if I = 1 then do;
    svars=citem;
  end;
  else do;
    svars = svars|| ' '||citem;
  end;
end;
svars = svars|| ' '||'nobsnum';
rc=close(mfdsid);
rc=close(sfdsid);
```

*Add unique number to each dataset;

```
call symput('fsdata',mfname);
call symput('sfdata',sfname);
```

```
rc=preview('copy',syscats.opencode.cntlnum.source');
submit continue;
endsubmit;
```

*Create index on search file;

```
sfdsid = open(sfname,'v');
rc=icreate(sfdsid,'srchindx',svars,'unique
nomiss');
```

```
rc = close(sfdsid);
```

*Create index on master file;

```
mfdsid = open(mfname,'v');
rc=icreate(mfdsid,'mstrindx',mvars,'unique
nomiss');
```

```
rc = close(mfdsid);
```

*Open datasets with applied keep statements;

```
mfdsid =open(mfname||'(Keep='||mvars||')','u');
sfdsid = open(sfname||'(Keep='||svars||')','u');
rc = setnitemn(classid,mfdsid,'mfdsid');
rc = setnitemn(classid,sfdsid,'sfdsid');
```

```
mvarlsid = makelist();
svarlsid = makelist();
```

```
pgwtlist = makelist();

*Start creation of positive global weights;

*Create random pairs for 2000 records;

if numsfrec > 2000 then nob = 2000;
else nob = numsfrec;

do I = 1 to nob;
  rannum = round(ranuni(12345)*nob,1);
  if rannum > attrn(sfdsid,'nob') then
    rannum = attrn(sfdsid,'nob');

*Fetch observation from search file and put values in list;
rc = fetchobs(sfdsid,rannum);
n=listlen(svlistid);
do j = 1 to n;
  citem = getitemc(svlistid,j);
  vtype = vartype(sfdsid,varnum(sfdsid,citem));
  if vtype = 'N' then
    rc = insertn(svarlsid,getvarn(sfdsid,
      varnum(sfdsid,citem)),-1,citem);
  else rc = insertc(svarlsid,getvarc(sfdsid,
    varnum(sfdsid,citem)),-1,citem);
end;

*Fetch observation from master file and put values in list;

rc = fetchobs(mfidsid,rannum);
n=listlen(mvlistid);
do j = 1 to n;
  citem = getitemc(mvlistid,j);
  vtype = vartype(mfidsid,varnum(mfidsid,citem));
  if vtype = 'N' then
    rc = insertn(mvarlsid,getvarn(mfidsid,
      varnum(mfidsid,citem)),-1,citem);
  else rc = insertc(mvarlsid,getvarc(mfidsid,
    varnum(mfidsid,citem)),-1,citem);
end;

*Compare values and get counts;

n=listlen(mvlistid);
do j = 1 to n;
  out1cnt = 0;
  out2cnt = 0;
  out3cnt = 0;
  out4cnt = 0;
  mitem = getitemc(mvlistid,j);
  sitem = getitemc(svlistid,j);

ltype = itemtype(mvarlsid,nameditem(mvarlsid,mitem));
```

```
if ltype = 'C' then do;
  if getnitemc(mvarlsid,mitem) =
    getnitemc(svarlsid,sitem) then out1cnt+1;
  else if getnitemc(mvarlsid,mitem)=' ' and
    getnitemc(svarlsid,sitem) ne ' ' then out2cnt+1;
  else if getnitemc(mvarlsid,mitem) ne " and
    getnitemc(svarlsid,sitem) = ' ' then out3cnt+1;
  else
    out4cnt+1;
end;
else do;
  if getnitemn(mvarlsid,mitem) =
    getnitemn(svarlsid,sitem) then out1cnt+1;
  else if getnitemn(mvarlsid,mitem)=. and
    getnitemn(svarlsid,sitem) ne . then out2cnt+1;
  else if getnitemn(mvarlsid,mitem) ne . and
    getnitemn(svarlsid,sitem) = . then out3cnt+1;
  else
    out4cnt+1;
end;
```

*Insert counts into list;

```
found = nameditem(pgwtlist,'comp'||j);
```

if not found then do;

```
list = makelist();
list = insertn(list,out1cnt,-1);
list = insertn(list,out2cnt,-1);
list = insertn(list,out3cnt,-1);
list = insertn(list,out4cnt,-1);
pgwtlist = insertl(pgwtlist,list,-1,'comp'||j);
end;
else do;
  list = getiteml(pgwtlist,found);
  list = setitemn(list,getitemn(list,1)+out1cnt,1);
  list = setitemn(list,getitemn(list,2)+out2cnt,2);
  list = setitemn(list,getitemn(list,3)+out3cnt,3);
  list = setitemn(list,getitemn(list,4)+out4cnt,4);
end;
end;
```

end;

```
rc = clearlist(svarlsid,'Y');
rc = clearlist(mvarlsid,'Y');
end;
rc = dellist(svarlsid);
rc = dellist(mvarlsid);
```

*Get list counts, calculate weights and reinsert weights;

```
n=listlen(pgwtlist);
do j = 1 to n;
```

```
listid = getniteml(pgwlist,'comp'lj);
do k = 1 to 4;
weight=abs(round(log2(getitemn(listid,k)/ nobl)*10,1));
if weight = . then weight = 0;
listid = setitemn(listid,weight,k);
end;
end;
rc = setnitemn(classid,pgwlist,'pgwlist');
```

endmethod;

CONCLUSION

Awareness of Probabilistic Record Linkage and its usefulness continues to grow in the user community. As this awareness grows, so will the need for software that allows you to easily and quickly perform probabilistic record linkage. Furthermore, the need will not only be limited to the ability to conduct linkages, but you will also need other system capabilities such as data management and data analysis. The SAS System furnishes an excellent array of development tools that make the task of providing users with such a system very feasible.

DATALINK 1000 is a prototype developed with limited resources. This system has some of the features that make a probabilistic record linkage system much more useful and powerful as an analytical tool. However, clearly further development of the system is warranted.

With the release of version 6.11, SAS/AF software has many new enhancements that will add functionality to DATALINK 1000 when incorporated into the system. Future releases will exploit these new features. Furthermore, new and improved tools will be developed and added to DATALINK 1000.

TERMS

record linkage: The bringing together of pieces of information from two separate data sources which are believed to relate and identify the same entity.

probabilistic record linkage: Calculating the likelihood that a pair of records represent the same entity.

positive global weights: These are the logarithms of the frequencies of the various outcomes of the identifier comparisons which are derived from a file of unlinkable pairs of records that are brought together at random.

BIBLIOGRAPHY

Martin, James and James J. Odell (1992), *Object-Oriented Analysis and Design*, New Jersey: Prentice Hall.

Newcombe, Howard B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administrative, and Business*, New York: Oxford University Press.

SAS Institute Inc. (1993), SAS/AF® Software: FRAME Entry Usage and Reference, Version 6, First Edition. Cary, NC: SAS Institute, Inc.

SAS Institute Inc. (1994), SAS® Screen Control Language: Reference, Version 6, Second Edition. Cary, NC: SAS Institute Inc.

Winblad, Ann L., Samuel D. Edwards, and David R. King (1990) *Object-Oriented Software*, USA: Addison-Wesley Publishing Company, Inc.

SAS, SAS/ACCESS, SAS/AF, SAS/GRAPH, and SAS/STAT are registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

The author may be contacted at:

Gary W. Griffin
Whitehurst Associates, Inc.
341 Eureka Drive
Atlanta, GA 30305
(404) 240-0729
Internet: ggriffin@mindspring.com

Figure 1: Probabilistic Record Linkage Model

