PC Matchmaker:
A General Personal Computer-Based Matching Program for Historical Research.
USER'S GUIDE

Jeremy Atack
Vanderbilt University and NBER

Fred Bateman
University of Georgia

Mary Eschelbach Hansen
American University

Programmer/Consultant
Timothy Gregson

Program conceived by Jeremy Atack and Fred Bateman,
developed by Mary Eschelbach Gregson (now Hansen),
programmed by Timothy Gregson,
and funded by the National Science Foundation
under grants SES-8914861 and SES-9001066

PC MATCHMAKER USER'S GUIDE

(In addition to reading this documentation, users are urged to read Jeremy Atack, Fred Bateman and Mary Eschelbach Gregson, "Martchmaker, Matchmaker, Make Me a Match," *Historical Methods*, Spring 1992, Vol. 25, No. 2, pp. 53-65.)

INTRODUCTION

The following is a "how-to" guide for PC Matchmaker, a simple to use (i.e. menu- and prompt-driven) program for record linkage using PC-DOS-based personal computers. It can also be run in a DOS “CMD” window under 32-bit versions of Microsoft’s Windows operating system. Depending upon the OS version in use, this Window is accessed by typing “cmd” into the “search” box under the Windows 7 Start Menu or under “Accessories” in the “Programs” listing under the Start Menu.

PC Matchmaker draws upon the accumulated experiences of historians, genealogists, medical professionals and government agencies. Without their work, we could not and would not have devised this program.

SYSTEM REQUIREMENTS

PC Matchmaker operates in a DOS (MS-DOS or IBM PC-DOS) environment. The main portions of the program are written in Microsoft QuickBASIC 4.5. Some utilities and subroutines are in Microsoft C 6.0. QuickBASIC was chosen because it is widely known. Modifications to the program can thus be made inexpensively by inexpert users should they be needed. No extraordinary hardware or software is required to link files with PC Matchmaker. However, considerable disk space is necessary to link even moderately-sized files.

As described above, it can also be run within a DOS Window on a 32-bit Windows OS computer. To run on a 64-bit machine requires recompiling the program and its subroutines.

It is currently being adapted and rewritten for use on the Web by Mary Eschelbach Hansen.

CONVENTIONS

In this guide, text centered on a separate line in UPPERCASE are commands you may enter at the DOS prompt. Plain courier indicates prompts displayed by PCM that you will see on-screen. Text in brackets ([]) indicates optional parts of commands. Text or two dots inside greater-than and less-than signs (<xx> or <..>) indicates information that is specific to any run of PCM. This is usually path or file names. DO NOT type the [] or <> as part of the command.

RUNNING PC MATCHMAKER

We distribute PC Matchmaker as a self-extracting archive file named PCM.EXE. Copy PCM.EXE to the directory where you would like the program to be. Then execute PCM.EXE. The program files will appear in the directory. The file PCMCODE.EXE is a self-extracting achive of the code. You DO NOT need this to run PC Matchmaker. It should be used only by experienced computer programmers to modify PC Matchmaker

PCM.BAT

PC Matchmaker is invoked though the batch file PCM.BAT. This batch file takes as parameters two path names (PATHA and PATHB) and two file names (FILEA and FILEB) plus an optional Matching Instruction File name (SCRFILE). The syntax is

```
PCM PATHA PATHB FILEA FILEB [SCRFILE]
```

The paths tell PC Matchmaker the location of the two data files to be linked. The two paths may be identical, but we suggest you keep data files in separate directories.

Some Caveats:

PC Matchmaker expects that the files to be linked will be dBASE files. File names must be given without the ".dbf" extension. The files need not have the same structure or field names. PC Matchmaker can compare any two fields of the same dBASE type (i.e. numeric, text, or logical) and the fields need not have the same name.

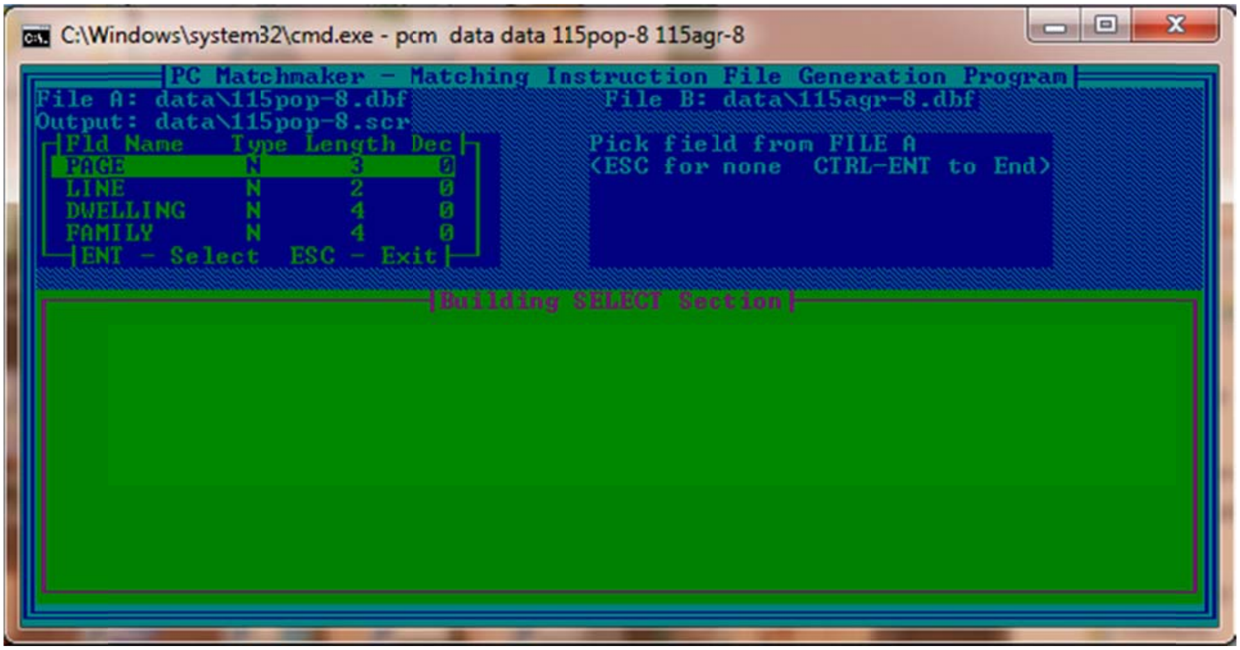
There are, however, two caveats:

(1) Memo fields are not explicitly supported and should be stripped from data files prior to using the PC Matchmaker and

(2) if dBASE indexing is desired data files must be reindexed after processing with the PC Matchmaker.

THE MATCHING INSTRUCTION FILE

The Matching Instruction File (SCRFILE) that is optional on the command line contains the criteria that PC Matchmaker will use to make links. From the user's point of view, developing the Matching Instruction File is the most important stage of the linking process.



```
C:\Windows\system32\cmd.exe - pcm  data data 115pop-8 115agr-8

PC Matchmaker - Matching Instruction File Generation Program
File A: data\115pop-8.dbf                File B: data\115agr-8.dbf
Output: data\115pop-8.scr
 Fld Name      Type Length Dec          Pick field from FILE A
 PAGE           N      3    0           <ESC for none   CTRL-ENT to End>
 LINE           N      2    0
 DWELLING       N      4    0
 FAMILY         N      4    0
  ENT - Select  ESC - Exit

                    Building SELECT Section
```

SCREENSHOT OF “MATCHING INSTRUCTION FILE” CREATION

If you do not specify a Matching Instruction File, PC Matchmaker will prompt you to create one. PC Matchmaker will present you with the "Matching Instruction File Generation Program" screen. Usually you will build a Matching Instruction File with PC Matchmaker's menu-driven generator. Then, if you are linking many pairs of files you can bypass construction of a new matching instruction file for every run by specifying (on the command line) the matching instruction file that you created using the generator on earlier runs. Likewise, if you want to experiment with different criteria for matching, you can load the matching instruction file into a text editor and make changes. When PC Matchmaker generates the matching instruction file it is named <FILEA>.SCR and can be found in DIRA.

To edit this file, you may have to change the file extension from “SCR” to “TXT” (and remember to change it back to SCR after making changes as the batch program searches for SCRFILE.SCR although you do **not** specify the extension).

An example of a Matching Instruction File:

```
SELECT
        AGE   NULL >15    0

BLOCK
        NYSIIS        LASTNAME  LASTNAME  LASTNYS    LASTNYS

MATCH
        NAMELASTNAME  LASTNAME
                EXACT
                FIRSTXLETTERS    5
                BLOCK
                BLOCK
        NAMEFIRSTNAME FIRSTNAME
                EXACT
                FIRSTXLETTERS    3
        NAMEINITIAL       INITIAL
                EXACT
```

PC Matchmaker needs your instructions for four basic functions: SELECT, BLOCK, FILTER and MATCH. This is the information in the matching instruction file. After running PC Matchmaker for the first time, we suggest you examine the matching instruction file using your text editor or file viewer.

The SELECT function lets the researcher link subsets of the Data Files. PC Matchmaker itself imposes no restrictions on the SELECT function. It is up to the researcher to determine the records he wishes to SELECT. The BLOCK function partitions the data and restricts searches to within blocks. FILTER further limits PC Matchmaker's search for links by setting minimum standards for likeness between records. Note that select, block and filter all restrict the number of comparisons made in some way. None of these functions is mandatory, but using them increases the speed and efficiency of PC Matchmaker. On the other hand, the MATCH function tells PC Matchmaker the specific questions it should ask when it compares two records. PC Matchmaker cannot function unless the MATCH function is defined.

THE MATCHING INSTRUCTION FILE GENERATION PROGRAM

The "Matching Instruction File Generation Program" is automatically called when no matching instruction file is specified on PC Matchmaker's command line. For all four functions, the box in the top left-hand corner of the screen displays the fields in File A. The box in the right-hand corner displays the fields in File B. As you define each function, PC Matchmaker displays your choices in the bottom half of the screen. When PC Matchmaker displays a list of choices or fields in a box, you can move the highlight bar with the arrows or page up/down keys. Pressing ENTER chooses the highlighted line.

Defining SELECT

Use the SELECT function if you want to link a subset of your data files. You can tell PC Matchmaker to SELECT only those records that fit your description. To select records from filea, move the highlight bar to the appropriate field name and press ENTER. PC Matchmaker enters the chosen field in the bottom half of the screen and prompts you for the SELECT value for this field, i.e. the expression that describes the records you want to SELECT. If a value is given PC Matchmaker assumes it should SELECT records that have the specified field equal to this value. The user can override the assumption of equality by specifying greater than ">" or less than "<" at the beginning of the expression. "Not equal to" is not an available option, but the SELECTs can be compounded with the logical expressions AND and OR.

Pressing ESC toggles between the File A and File B windows.

CTRL-ENTER completes the SELECT function and moves you to BLOCK.

SELECT is not mandatory for either data file. If you wish to consider all records in both data files, press CTRL-ENTER. PC Matchmaker will then skip SELECT and move to BLOCK.

Defining BLOCK

Use the BLOCK function if you want PC Matchmaker to compare only those records that are similar in some way. BLOCKing is like dividing the records into piles according to your idea of similarity, and then only making comparisons between piles. We strongly recommend that you BLOCK all but the smallest data files. PC Matchmaker permits multiple (nested) blocking.

If you wish to block the data files, PC Matchmaker offers a choice of NYSIIS, Soundex, First X Letters, or COPY your own block field. PC Matchmaker generates NYSIIS and Soundex codes if one of these options is taken. The COPY and FIRSTXLETTERS options allow you to block on the entirety or first x letters of fields that already exist in Files A and B. Regardless of the option chosen PC Matchmaker prompts you to choose the field from each data file on which to BLOCK. The program than asks for a (unique) field name for the BLOCK output.

> Important Note
>
> If you define the MATCH function (discussed below) on a field you use for blocking, you must select Block as your last MATCH subtype on that field. Failing to do so will invalidate the weights calculated from the data in the BLOCK field.

Pressing CTRL-ENTER moves you from BLOCK to FILTER.

Defining FILTER

Use the filter function to limit comparisons within BLOCK(s). You can filter on NAME (character) or NUMERIC fields. If you FILTER on a character field PC Matchmaker throws out comparisons that do not have a defined number of common letters. The common letters must appear sequentially, but they need not be contiguous. After you have chosen the fields to FILTER on, PC Matchmaker prompts you to give the number of common letters (FILTER characters). Numeric filters are always exact: comparisons are thrown out unless the value of the fields is equal.

Pressing CTRL-ENTER moves you from FILTER to MATCH.

Defining MATCH

The MATCH function is mandatory. The MATCH function tells PC Matchmaker to ask a series of questions that can be answered about each comparison that passes FILTER. The questions are nested in the sense that when a question is answered in the affirmative, the program records the answer and skips to the next field on which the MATCH function is defined. No records can be linked if MATCH is not specified.

MATCH types are NAME, NUMERIC and TEXT. Choose the MATCH type and two fields to be considered. Define only one MATCH type for each pair of fields. NAME and NUMERIC MATCH types have subtypes. The NAME subtypes are EXACT, FIRSTXLETTERS, and BLOCK. More than one subtype may be defined for each MATCH type.

MATCH subtypes should always be defined in decreasing order of importance. That is, always have PC Matchmaker ask if the fields are exactly alike before it asks if the first four letters are the same.

> Important Note
> The BLOCK subtype must be the last subtype when
> comparisons are made on the blocking fields. Do not use the
> BLOCK subtype otherwise.

NUMERIC subtypes are in terms of +/- ranges. The TEXT type has only one (implied) subtype, EXACT. Thus the TEXT type is the same as NAME type, EXACT subtype.

RECORD LINKAGE

CTRL-ENTER brings up a copy of your Matching Instruction File. Pressing ESC at this time aborts PC Matchmaker. ENTER accepts the Matching Instruction File and begins record linkage.

At this time the program calculates summary statistics on the occurrence of values in each MATCH field. These frequencies are used for weighting the comparisons. That is, the result of each question asked by the MATCH function is combined with the frequencies to give a statistical weight to the likelihood that the two records compared are for the same person.

LINKING FILES WITHOUT CALCULATING TRANSMISSION WEIGHTS

PC Matchmaker calculates the weights and outputs a report of all the comparisons. When the program has completed these tasks, it exits to DOS with this message:

```
WEIGHT REPORT IS IN FILE <..>.
AT THIS TIME, PLEASE DETERMINE THE CUTOFF VALUE.
TO COMPUTE TRANSMISSION WEIGHTS AND RERUN WEIGHT
CALCULATIONS, RUN
ITERATE DIRA DIRB FILEA FILEB CUTOFF [SCRFILE]
TO APPLY DEFINITE LINKS AND CREATE
INDETERMINATE FILE, RUN
LINKER DIRA DIRB FILEA FILEB CUTOFF [GRPA] [GRPB]
```

The rest of this section assumes you do not want compute transmission weights (transmission weights are used to adjust the initial weights by calculating the probability of error in the data). Details about computing transmission weights are in the next section.

The report file PC Matchmaker gives the contents of the fields used for the MATCH function, the weights computed for each field and the total weight for each comparison. In every matching system there is a tradeoff between completeness and accuracy. For some projects it is necessary that all links be true links. For other projects it is more important that the greatest number of records are linked automatically. PC Matchmaker leaves this choice to the researcher.

You should examine the report file PC Matchmaker has created. Decide at what weight the links become unlikely based on your judgement of the project and the data. This will be the cutoff value for the run.

Any non-zero value can be chosen as the cutoff value. A value of zero is treated as null.

LINKING

When you have decided on a cutoff value, type

```
LINKER <DIRA> <DIRB> <FILEA> <FILEB> <CUTOFF VALUE> [GRPA  OR NONE]
[GRPB OR NONE]
```

Note that this is the command that you were prompted to give when PC Matchmaker finished writing the report to disk. LINKER is another batch file that applies the cutoff value you gave in its command line.

The parameters GRPA and GRPB refer to "grouping" fields. This is useful when the records being linked have other records that need to be associated with the linked records. For example, much of our work involves the linkage of population census records between census years. The "group" is the family though we link on heads of household.PC Matchmaker is setup for one-to-one record linkage. Thus, when it applies the cutoff value, it chooses that comparison with the highest total weight that exceeds the cutoff. Other comparisons, even if the total weight exceeds the cutoff value, are discarded.

LINKER.BAT outputs two dBASE format files named FILEA.DEF (for definite links) and FILEA.IND (for indefinite comparisons). FILEA.DEF has the record numbers of the comparisons that were above the cutoff value. FILEA.IND contains the record numbers of the remainder of the comparisons that PC Matchmaker made. To view or use these files in dBASE, you should rename them to <..>.dbf.Sometimes the same weight is assigned to more than one comparison for a record. PC Matchmaker lets you decide which comparison will make the best link. Such comparisons appear in FILEA.IND even if their weight exceeds the cutoff value.

LINKER.BAT also adds a field called LINKCODE to each of the original data files and enters an indentification number in that field for each comparison above the cutoff value.You can merge FILEA.IND with the data files to make decisions about the remaining comparisons.

OPTIONAL COMPUTATION OF TRANSMISSION WEIGHTS

Errors can enter data files in many ways. To the extent that the data files do not contain perfectly correct information on the population, the weights computed thus far overestimate the probability that a comparison is a true link. In effect PC Matchmaker has assumed that the data files contain only true information. If a comparison is above the cutoff value (you have determined that it is very likely a true match) but doesn't match exactly on every field, there are most likely errors in the data. The weights can be adjusted downward to account for the probability of such errors. PC Matchmaker computes the proportion of comparisons above the cutoff value that have discrepancies. This is the estimated error rate. During iteration the weights are adjusted by (1-error rate). Iteration can be continued until the adjustment to the weights is arbitrarily small.

Recall that when PC Matchmaker finishes its initial run, it displays the following

```
WEIGHT REPORT IS IN FILE <..>.
AT THIS TIME, PLEASE DETERMINE THE CUTOFF VALUE.
TO COMPUTE TRANSMISSION WEIGHTS AND RERUN WEIGHT
CALCULATIONS, RUN
ITERATE DIRA DIRB FILEA FILEB CUTOFF [SCRFILE]
TO APPLY DEFINITE LINKS AND CREATE
INDETERMINATE FILE, RUN
LINKER DIRA DIRB FILEA FILEB CUTOFF [GRPA] [GRPB]
```

The previous section described how to link files assuming no transmission weights were to be estimated. If you do want to adjust the weights by the probability of data error, instead of running LINKER run ITERATE. First, examine the report file and determine the cutoff value to use for this run. Then type

```
ITERATE DIRA DIRB FILEA FILEB CUTOFF [SCRFILE]
```

ITERATE computes a new set of weights for all comparisons. As it calculates new transmission weights, ITERATE displays them alongside the transmission weights from the previous run so you can decide if further iteration will be beneficial (i.e. you can see how much they have changed). ITERATE essentially controls a loop back through the weighting section of the program and leaves you at the same point it began, with the message:

```
NEW WEIGHT REPORT IS IN FILE <..>.
AT THIS TIME, PLEASE DETERMINE THE CUTOFF VALUE.
TO RECOMPUTE TRANSMISSION WEIGHTS AND RERUN WEIGHT
CALCULATIONS, RUN
ITERATE DIRA DIRB FILEA FILEB CUTOFF [SCRFILE]
TO APPLY DEFINITE LINKS AND CREATE INDETERMINATE FILE, RUN
LINKER DIRA DIRB FILEA FILEB CUTOFF [GRPA] [GRPB]
```

At this point you should either iterate through the weighting section of the program again with newly computed transmission weights or choose a final cutoff value and link records as described in the previous section.

PLEASE ADDRESS ANY COMMENTS TO:
Jeremy Atack
Department of Economics
Box 1819 Station B
Vanderbilt University
Nashville, TN. 37235

(615) 434-2467

Jeremy.atack@vanderbilt.edu