**TCLINK: TAQ-CRSP link table**

1. GENERAL INFO
   * All TAQ master files are concatenated into \_mast1 with the vintage date (see below) that basically groups by file.
   * TAQ vintage dates, i.e. last day of the yyyymm which the master files belongs to, are used in the Ticker match, instead of the entry specific FDATE (line 141).
   * The date field in the TAQ master files for months in [199601-199612, 200407-201412] is DATEF instead of FDATE. ACTION: rename (line 38).
   * They do not use CUSIP8 fields from the master files, but extract their own (lines 39, 54).
   * COMNAM (CRSP) and NAME (TAQ) might have multiple blanks and might differ by case sensitiveness. ACTION: consolidate multiple blanks into one take upper case (lines 65, 98, 111)
   * Some company may have more than one TICKER-PERMNO link within vintage date.
   * NAMEDIS is the smallest spelling distance of the NAME from COMNAM or viceversa. The spelling distance spedis(query, keyword) is a function that maps the number of edits necessary to convert the keyword to the query string, into some cost (0 is better).
2. PRE-PROCESS
   1. Data editing
3. Missing FDATE\* is set to the variable date (they call it vintage date), i.e. last day of month since on line 27 date = %sysfunc(intnx(MONTH,&date1,&m,E))
4. Remove the ' .' from CUSIP and strip leading/trailing blanks, i.e. line 52 CUSIP = strip(compress(CUSIP," ."));
5. CUSIP8 is generated as the first 8 chars of CUSIP
6. DELETE all records without CUSIP AND NAME

\* Master files affected are 200002-200003. The FDATE is corrupt and cannot be recovered (as per response by WRDS support). We simply drop those two files with the assumption that name events in those two months are minor. If I recall correctly, only one entry is lost from the final table.

* 1. Sorting

All TAQ master entries (rows) are sequentially sorted in ascending order by DATE, SYMBOL, FDATE and CUSIP and stored into the \_mast2 table. Duplicate entries (qualified by the sorting fields) are removed.

1. LINKING
   1. Link by CUSIP
2. **Retrieve from CRSP the entries with latest available COMNAM corresponding to unique pairs of PERMNO, NCUSIP (for non-missing NCUSIPs) and sort the created \_msenames table (triplets) by PERMNO and NCUSIP.   
   They use** CRSP Monthly Stock Event - Name History **table, i.e. /wrds/crsp/sasdata/#\_stock/msenames.**
3. **Left match the CUSIP8 (TAQ \_mast2) to the NCUSIP (CRSP \_msenames).\***

\* No date conditions since it is implicit in the NCUSIP design.

* 1. Link by Ticker
     1. Separate TAQ entries into unmatched (\_NoMap1) and matched (\_Match1). The match is scored as 0 (line 87).
     2. Keep SYMBOL, COMNAM, vintage date and FDATE if COMNAM is not null (into \_NoMap2), i.e. dropping null names.
     3. Retrieve unique pairs of PERMNO, SMBL\*, COMNAM and min/max DATES\*\* (line 114) into table \_CRSP2.
     4. Keep only most recent COMNAMs for each PERMNO, SMBL into table \_CRSP3.
     5. Match SYMBOL (TAQ \_NoMap2) to SMBL (\_CRSP3) while TAQ vintage date is within CRSP’s NAMEDT and NAMEENDDT.
     6. Score ticker-name matches as 2 (implicitly NAMEDIS <= 30 evaluates to false and leaves the match to 2, line 157).   
        I am missing where the score 1 goes. It was given to missing PERMNOs in the CUSIP match on line 87 (I probably don’t understand the syntax). Then 1 is carried over for entries which have a NAME.
     7. Assign match type 3 if the NAMEDIS is bigger than 3 (line 157).

\* If it exists, they pick the trading ticker TSYMBOL, otherwise the CRSP TICKER, which can be null.   
\*\* The symbol might change forth and back. We sort by CUSIP, FDATE and SYMBOL, which clearly outlines the following case:

|  |  |  |  |
| --- | --- | --- | --- |
| CUSIP | SYMBOL | NAME | FDATE |
| 00088E10 | IATV | ACTV INC | 19930104 |
| 00088E10 | ACTV | ACTV INC | 19950503 |
| 00088E10 | IATV | ACTV INC | 19980630 |

Here, you cannot simply consolidate the range of existence of a SYMBOL by min(FDATE) and max(FDATE), or you will end up trying to match IATV on [19950503 – 19980630], when ACTV instead should be used.

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* W R D S R E S E A R C H M A C R O S \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* WRDS Macro: TCLINK \*/

/\* Summary : Create TAQ-CRSP Link Table \*/

/\* Date : September 20, 2010 \*/

/\* Author : Rabih Moussawi, WRDS \*/

/\* Variables : - BEGDATE and ENDDATE are Start and End Dates in YYYYMM format \*/

/\* - OUTSET: TAQ-CRSP link table output dataset \*/

/\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

%MACRO TCLINK (BEGDATE=199301,ENDDATE=201012,OUTSET=WORK.TCLINK);

/\* Check Validity of TAQ Library Assignment \*/

%if (%sysfunc(libref(taq))) %then %do; libname taq "/wrds/taq/sasdata/"; %end;

%put; %put ### START . ; %put ;

/\* IDEA: Use VINTAGE-SYMBOL as TAQ Primary Key \*/

/\* Then Link it to PERMNO using CUSIP and Ticker Info \*/

options nonotes;

%let date1= %sysfunc(inputn(&begdate,yymmn6.));

%let date2= %sysfunc(inputn(&enddate,yymmn6.));

%if &date1<&date2 %then %let NMONTHS=%sysfunc(intck(MONTH,&date1,&date2));

%else %let NMONTHS=0;

data \_mast1; set \_null\_; run;

/\* Begin Loop To Construct a 'Master' TAQ Master Dataset \*/

%do m=0 %to &NMONTHS;

%let date = %sysfunc(intnx(MONTH,&date1,&m,E));

%let yymm = %sysfunc(putn(&date,yymmn6));

/\* Make Sure that dataset Exist \*/

%if %sysfunc(exist(taq.mast\_&yymm))=1 %then

%do;

%put ### Processing Master Dataset for &yymm ### ;

data \_mastm; format DATE date9.;

set taq.mast\_&yymm;

date=&date;

%if (&yymm>=199601 and &yymm<=199612) or (&yymm>=200407 and &yymm<=200412) %then

%do;

rename DATEF=FDATE;

drop CUSIP8;

%end;

run;

%if &m=0 %then %do; data \_mast1; set \_mastm; run; %end;

%else %do; proc append base=\_mast1 data=\_mastm force; run; %end;

proc sql; drop table \_mastm; quit;

%end;

/\* End Loop \*/

%end;

/\* Clean TAQ Master Dataset Information \*/

data \_mast2; format CUSIP8 $8.;

set \_mast1 (keep=DATE FDATE CUSIP SYMBOL NAME SHROUT TYPE);

CUSIP = strip(compress(CUSIP," ."));

if missing(FDATE) then fdate=date;

if not missing(CUSIP) then CUSIP8=substr(CUSIP,1,8);

if missing(CUSIP) and missing(NAME) then delete;

run;

/\* Sort Data using DATE-SYMBOL Key \*/

proc sort data=\_mast2 nodupkey; by date symbol fdate cusip; run;

/\* Step 1: Link by CUSIP \*/

/\* CRSP: Get all PERMNO-NCUSIP combinations \*/

proc sql;

create table \_msenames

as select distinct permno, ncusip, upcase(compbl(comnam)) as comnam

from crsp.msenames where not missing(ncusip)

group by permno, ncusip

having nameendt=max(nameendt);

quit;

proc sort data=\_msenames nodupkey; by permno ncusip; run;

/\* Map TAQ and CRSP using 8-digit CUSIP \*/

proc sql;

create table \_mast3

as select b.permno, a.\*, b.comnam

from \_mast2 as a left join \_msenames as b

on a.cusip8=b.ncusip;

quit;

/\* Step 2: Find links for the remaining unmatched cases using Exchange Ticker \*/

/\* Identify Unmatched Cases by Splitting the Sample into Match1 and NoMap1 \*/

proc sort data=\_mast3 nodupkey; by date symbol permno fdate; run;

data \_Match1 \_NoMap1;

set \_mast3;

by date symbol permno fdate;

if last.symbol;

SCORE=(missing(permno));

NAMEDIS=min(spedis(name,comnam),spedis(comnam,name));

if not missing(permno) then output \_match1;

else output \_NoMap1;

run;

/\* Add the Matches by Ticker \*/

data \_NoMap2;

set \_NoMap1;

where not missing(name);

symbol=strip(symbol);

name = upcase(compbl(name));

drop permno comnam score namedis;

run;

/\* Get entire list of CRSP stocks with Exchange Ticker information \*/

/\* Arrange effective dates for link by Exchange 'Trading' Ticker \*/

/\* Use CRSP Ticker if Trading Ticker is missing \*/

data \_CRSP1;

set crsp.msenames;

if not missing(tsymbol) then SMBL = tsymbol;

else SMBL=ticker;

smbl=strip(smbl);

if not missing(smbl);

COMNAM=upcase(compbl(comnam));

run;

/\* Get date ranges for every permno-ticker combination \*/

proc sql;

create table \_CRSP2

as select permno, smbl, comnam,

min(namedt)as namedt,max(nameendt) as nameenddt

from \_CRSP1

where not missing (smbl)

group by permno, smbl

order by permno, smbl, namedt;

quit;

/\* Label date range variables and keep only most recent company name \*/

data \_CRSP3;

set \_CRSP2;

by permno smbl;

if last.smbl;

label namedt="Start date of exch. ticker record";

label nameenddt="End date of exch. ticker record";

format namedt nameenddt date9.;

run;

/\* Get PERMNO for Unmatched Stocks using Ticker-DATE Match\*/

proc sql;

create table \_NoMap3

as select a.\*, b.permno,comnam,

min(spedis(a.name,b.comnam),spedis(b.comnam,a.name)) as NAMEDIS

from \_NoMap2 as a, \_CRSP3 as b

where strip(a.symbol)=strip(b.smbl) and a.date between namedt and nameenddt

order by date,symbol,namedis;

quit;

/\* Assign all Ticker Matches a Lower Score than CUSIP Matches \*/

data \_NoMap4;

set \_NoMap3;

by date symbol;

if first.symbol;

SCORE=2;

run;

/\* Score links using company name spelling distance: 0 is Best \*/

/\* Consolidate Link Table \*/

data \_TAQLINK1;

set \_Match1 \_NoMap4(in=b);

SCORE=SCORE+(NAMEDIS>30);

label SCORE="0.CUSIP+Names, 1.CUSIP, 2.Ticker+Names, 3.Ticker Only";

label NAMEDIS="Spelling Distance between TAQ and CRSP Company Names";

label DATE="TAQ Vintage Date";

label CUSIP8='8-digit CUSIP';

label CUSIP ='Full CUSIP Number: 9-digit CUSIP + 3-digit NSCC Exchange ID';

rename CUSIP=CUSIP\_FULL CUSIP8=CUSIP;

label SYMBOL="Stock Symbol in TAQ";

label NAME = "Company Name in TAQ";

label COMNAM = "Company Name in CRSP";

label FDATE = "Effective Date of Current TAQ Name Record";

run;

/\* Some companies may have more than one TICKER-PERMNO link, \*/

/\* Can Clean the link additionally for one observation per permno-date \*/

/\* Final Sort \*/

proc sort data=\_TAQLINK1 out=&outset nodupkey; by date symbol; run;

/\* House Cleaning \*/

proc sql;

drop table \_Mast1,\_Mast2,\_Mast3,\_msenames,\_CRSP1,\_CRSP2,\_CRSP3,

\_Match1,\_NoMap1,\_NoMap2,\_NoMap3,\_NoMap4,\_TAQLINK1;

quit;

%put; %put ### DONE . ; %put ;

options notes;

%MEND TCLINK;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* \*\*\*\*\*\*\*\*\*\*\*\*\* Material Copyright Wharton Research Data Services \*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* All Rights Reserved \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*/