Introduction to TAQ and ISSM

May 28, 2004

Patricia Ledesma Liébana

Agenda

- TAQ and ISSM – introduction
- Matching with other data sources
- Working with TAQ and ISSM
  - Should you learn SAS?
- Basic SAS skills and some tricks

TAQ and ISSM

- Trade and Quote (TAQ) database is a collection of intraday trades and quotes for all securities listed on the New York Stock Exchange, American Stock Exchange, Nasdaq National Market System and SmallCap issues. Data from 1993 – present in monthly files: trades, quotes, master table, dividends
- ISSM: Covers NYSE and AMEX between 1983 and 1992, and NASDAQ between 1987 and 1992. Each year of data is divided into two files, one for trades and one for quotes.
- Both TAQ and ISSM are available as SAS files in WRDS

Matching tick data to other data sources

Problems:

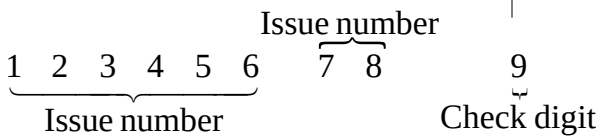
- Ticker symbols and CUSIP numbers are not constant throughout the history of a security.
- Ticker symbols may be recycled.

How to proceed:

- Fact: CRSP preserves historical CUSIP numbers, while Compustat keeps only the most recent CUSIP number
- Use CUSIPs and CRSP as a pivot: match by CUSIP, retrieve PERMNOs, use the merged CRSP-Compustat database

About CUSIPs (www.cusip.com)

- Three parts:



- Issuer number is unique with some exceptions
- Issue number: 10-88 for equity; 01 for options; fixed income always include one alphabetic character (I, 1 and O not used; 9Z is reserved)
- Check digit – modem legacy
- TAQ adds 3 digits that identify the exchange where the security was issued
- Compustat: CUSIP=issuer number; CIC=issue number + check digit

Matching TAQ with CRSP

- Match symbol to corresponding master table to retrieve TAQ CUSIP
- Use first 8 characters of TAQ's **CUSIP** to match with CRSP **NCUSIP** (*dsfnames* or *msfnames*) and retrieve the PERMNO
- If you need Compustat data, use the merged CRSP-Compustat files (using the **NPERMNO** from the *cstlink* file)

Matching ISSM to CRSP

- No CUSIP number provided.

Options:

- Match TAQ's SYMBOL with CRSP's TICKER. For cases with more than one match, assign the NCUSIP based on the date (must be between ST_DATE and END_DATE in *dsfnames* or *msfnames*)
- In skew3, there are "stats" files with matches based on end-of-year assignments. There are errors since exchanges reused tickers within the same year, contrary to their own rules.

Working with TAQ and ISSM

Issues:

- Logistics: amount of data
- Market microstructure: matching trades with quotes, trade direction, price impacts, etc.

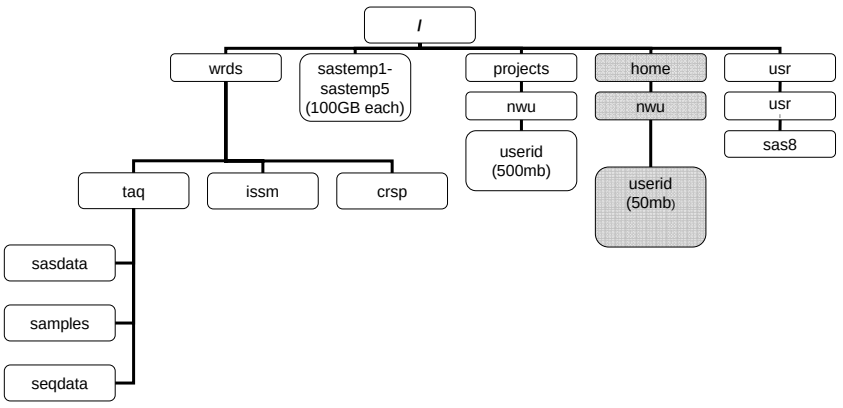
SAS in UNIX versus web + Matlab in a PC

- WRDS web interface is a perl script that writes a SAS program and runs it. It eliminates rows with missing values ➔ the number of rows you get depends on the variables you select. Not good for replication.
- SAS: steep learning curve; extremely efficient handling large amounts of data

Avoiding SAS

- Consolidated Trades (CT) files:
  - 132 files, 1.8GB on average (12/2003 takes 5.2GB)
  - Observations per month increased from average of 7.1 million in 1993 to 99.4 million in 2003.
- Consolidated Quotes (CQ) files:
  - 7.1GB on average (12/2003 takes 45.7GB)
  - Observations per month increased from average of 8.6 million in 1993 to 609.5 million in 2003.

WRDS structure



A simple program

```
data t1;
  set taq.ct0312;
  where symbol in ("IBM" "GE" "GM“)
    and ex eq "N"
    and (time ge '9:30:00't and time le '16:00:00't)
    and cond notin ("O" "Z“)
    and corr in (0 1);
  drop ex cond corr g127;
  format date yymmddn8.;

proc export data=t1 outfile="finc520.txt"
   dbms=csv replace;
```

Advice

- Use WHERE not IF when possible
- Place the conditions that eliminate the most observations at the beginning (speed)

From the December 2003 Consolidated Trades File
(111,060,168 observations)

	Time	Observations	Change
IBM, GE and GM with IF statement	04:38.9	526,554	
IBM, GE and GM with WHERE statement	3.71	526,554	
... and exchange (EX) is NYSE (N)		275,583	-250,971
.... and time between 9:30am and 4:00pm		275,304	-279
... sale condition (COND) is not opened last (O) or sold sale (Z)		275,293	-290
.... and correction code is (CORR) 0 (no correction) or 1 (corrected)		275,266	-27

Dates and time in SAS

- **Dates and times are numbers**:
  - Dates: Days elapsed since the SAS epoch (Jan. 1, 1960)
  - Times: Seconds elapsed since midnight
- What is displayed is a **format** – you can change the format, the underlying number remains the same.
- For example: **4:00pm** is **57600** for SAS.
  - With the time. format, it will be shown as 16:00:00.
  - With the timeampm11. format, it will be shown as 4:00:00 PM
- Another example: **Dec. 1, 2003** is **16040** for SAS
  - With the date9. format, it is shown as 01DEC2003
  - With the yymmddn8. format, it is shown as 20031201

Date and time functions

- Some useful date functions:
  - year(.)
  - month(.)
  - day(.)
  - weekday(.)
- Time functions:
  - hour(.)
  - minute(.)
  - second(.)