Getting Started with the CRSP Database at Washburn University — Rob Weigand

The *Center for Research in Security Prices* (CRSP) database contains stock prices, returns, trading volume, market capitalization, dividend and other data from 1962-2005. Data such as these are available for individual stocks. Our CRSP subscription also covers a wide variety of stock indexes. Among these are value- and equal-weighted indexes for the NYSE, NASDAQ and AMEX (separately and together), and indexes sorted into deciles by market cap, beta, and standard deviation. Surprisingly, some basic items are left out of the entry-level CRSP index subscription, such as a plain-vanilla S&P 500 index (CRSP can charge more for it).

Data Coverage. The WU SOBU subscription is for the *daily* CRSP database, vs. the *monthly* file which goes back to 1925. The designation "daily" and "monthly" refer more to the time period spanned by the data than to the frequency of the data. For example, although we subscribe to the daily database, data items can be sampled daily, monthly, quarterly or annually.

Why CRSP? One practical benefit of a CRSP subscription is that these are the stock market data that editors and reviewers at quality journals in the fields of finance and accounting expect to see in papers submitted for review. Another benefit is that the CRSP reporting tools make it (relatively) easy to sample the database and condition the output for the types of things that researchers in business usually do, such as conduct event studies (including lining up matching market index returns with stock returns) or obtain data for a portfolio of stocks using fixed dates. CRSP outputs data in a text format that is easily read by Excel or statistical programs such as SPSS and SAS (although sometimes a little more data conditioning is required to read into SAS).

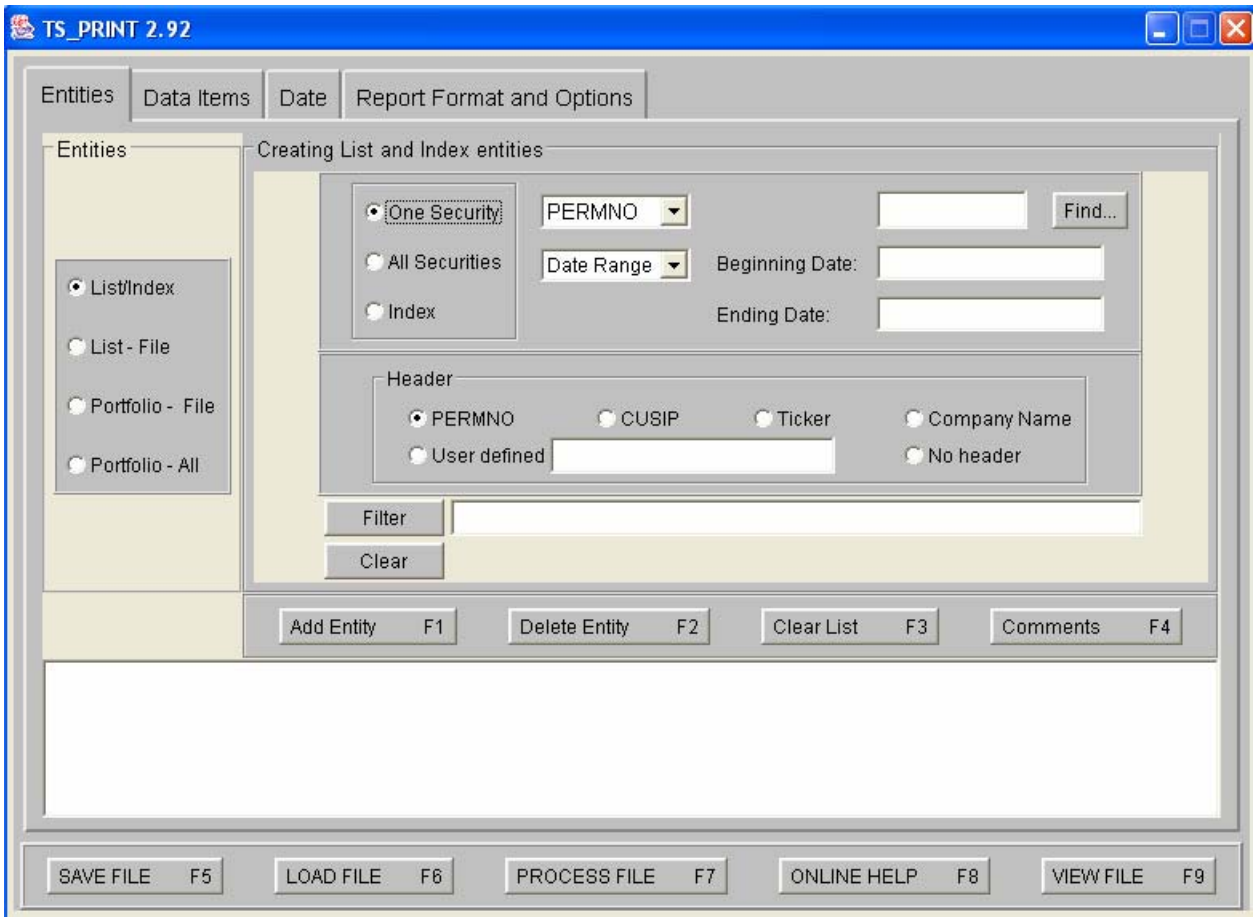
Accessing Data with TS Print. Our CRSP subscription comes with a menu-driven interface called TS Print. For a quick start with TS print, watch the PowerPoint tutorial tsprint_2005.ppt provided by CRSP. You'll be able to pull data such as prices, returns, volumes, dividends and market capitalizations using the menu-driven interface within 30 minutes.

Learning More. To learn more about the CRSP utilities, refer to the CRSP Utilities Guide (stock_indices_utilities.pdf). To gain a deeper understanding of the data covered by CRSP and the structure of the data, study the Data Description Guide (stock_indices_data_description.pdf).

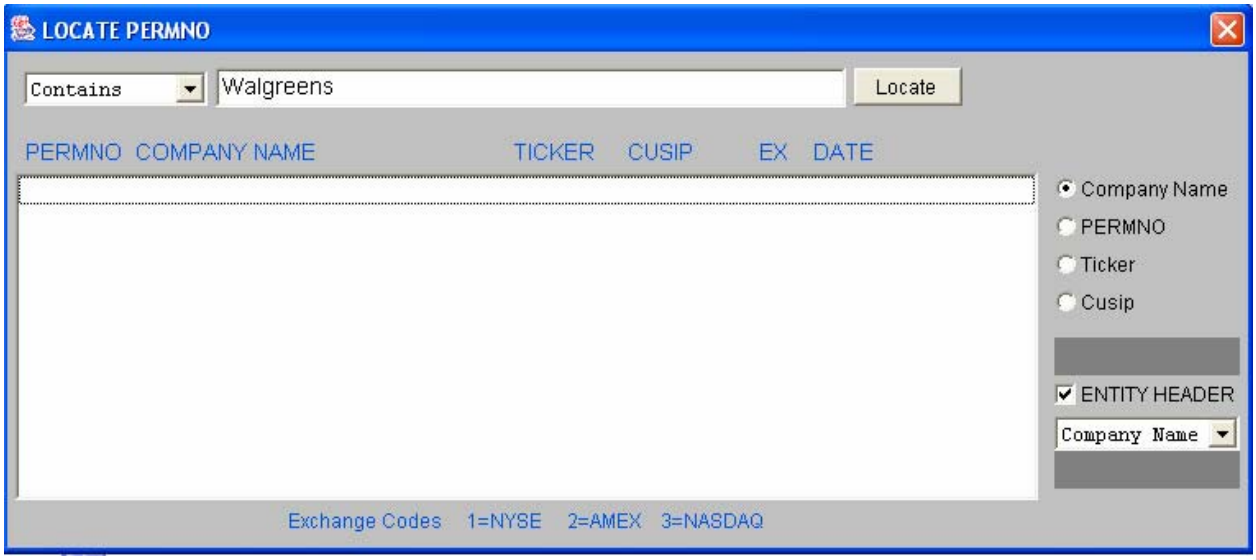
Just Do It. The best way to learn about CRSP is to access the data, either by following the TS Print PowerPoint tutorial, or by following the directions I've laid out below (this is the fast lane). The TS Print menu builds a file that contains directions for accessing the CRSP data. In what follows below I will demonstrate how to directly edit the access file without building it from the TS Print menu. Working directly in the file is the best way to learn CRSP syntax.

Permnos, Cusips and Tickers. One difference between CRSP and Compustat is the way the user identifies firms in the database. Compustat likes cusips and tickers; CRSP prefers an identifier known as a permno (permanent number). Accessing CRSP data using the permno is better because when a firm changes names or merges with another firm its cusip or ticker may change, but the permno stays the same. (The only thing we'll use the TS Print menu for in these examples is to look up a couple of permnos and then load and run our data access file.)

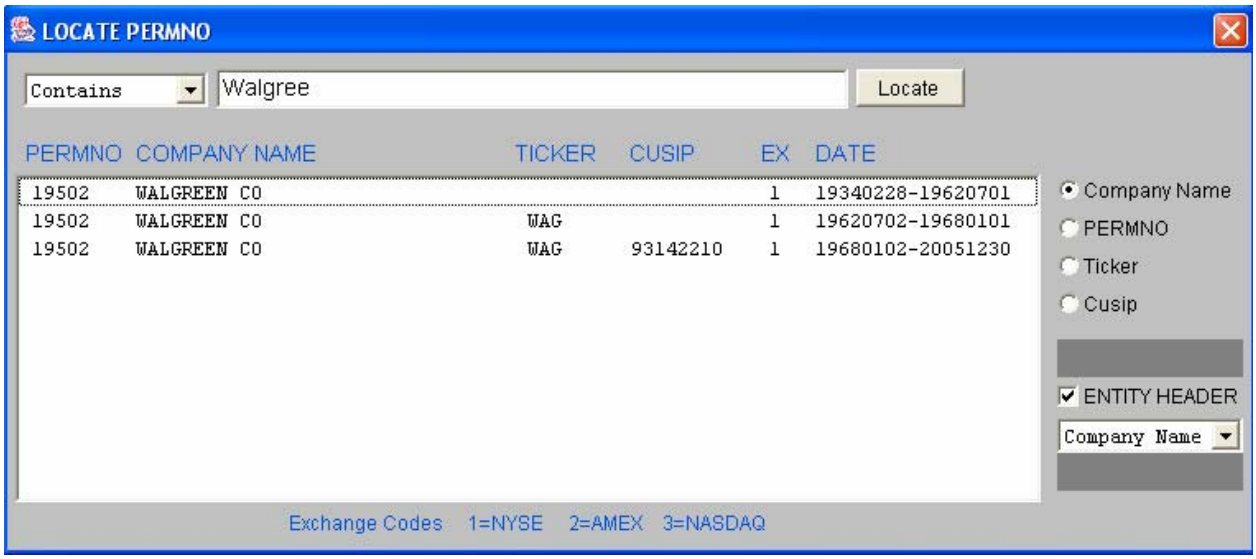
Open up the TS Print application. You'll see four tabs at the top — Entities, Data Items, Dates, and Report Formats and Options. The application should open on the Entities tab. Select List/Index, One Security, and Permno (see screenshot below), and click on "Find." Type in as much of the name of the firm you're looking for as seems practical. We'll look for Walgreens.



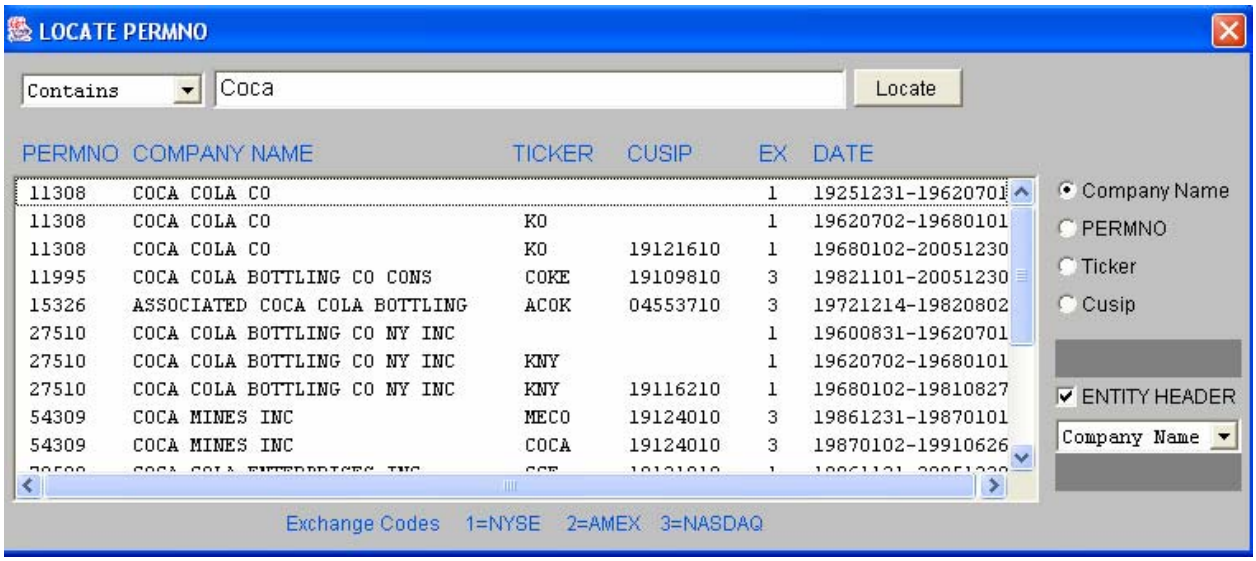
In the "Contains" box type in Walgreens and click on "Locate." Notice that nothing comes up. That's because the actual name of the company is Walgreen Co. This is typical of the rigidity of the CRSP database. In this case you would have made more progress by typing in "Walgree."



As shown below, this brings up three items. Notice they all have the same permno, but tickers and cusips can vary. The dates covered by each entry are also shown. Thinking through it for a moment we can infer that the permno "19502" covers continuous data for Walgreens from 1962-2005. (CRSP forces you to consider these sorts of details, there's no easy way around it.)



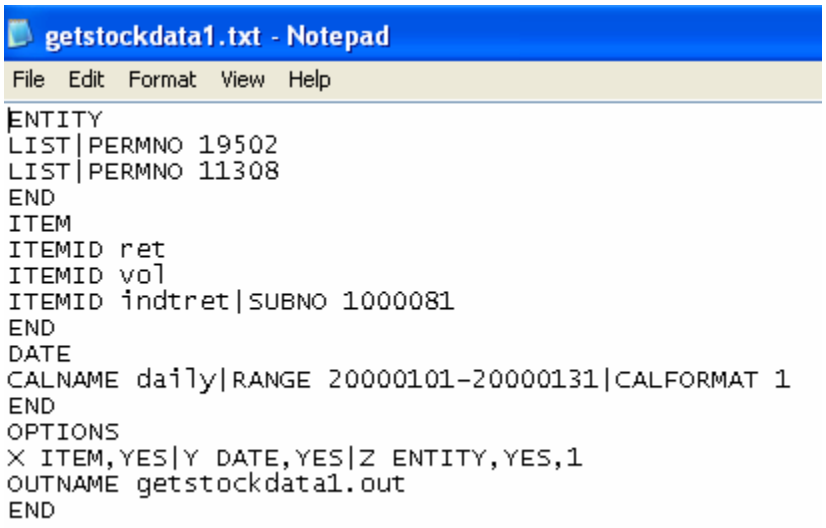
Let's include one more firm in our sample. How about Coca-Cola? If you type in Coca-Cola with a hyphen and click "Locate", CRSP returns nothing. That's because it lists the name as Coca Cola, without the hyphen. If you type in "Coca" and click "Find", you get the following:



No shortage of firms with "Coca" in the title. In this case it's helpful to also consider Coke's ticker (KO), scan all 3 listings for COCA COLA CO, and infer that the permno "11308" covers data from 1962-2005. (As above, CRSP is not well-known for just handing you your data without a minor struggle.) Next we'll examine an access file that will pull stock returns and trading volume for Walgreens and Coke, and match the returns with returns from a market index.

Choice of Market Index. The choice of market index is up to the individual researcher. In finance the "best" index to use has changed over time as various econometric issues arise, and are later trumped by newer econometric issues. For example, in the late 1980s-early 1990s everyone was using value-weighted indexes instead of equal-weighted indexes, as value-weighted indexes would be more closely correlated with the true changes in wealth that occurred in the macroeconomy when stock prices rose and fell. In 1998, however, a group of researchers published a paper that showed value-weighted indexes produced a bias in long-term compounded returns, so everyone switched back to equal-weighted indexes. In the last few papers I've published I've used an equal-weighted index of all the stocks on the NYSE, AMEX and NASDAQ (think of it as a "Wilshire 10,000"). The CRSP code for this index is 1000081. (The codes for all the CRSP indices begin on page 31 of stock_indices_data_descriptions.pdf).

The access file is a text file; it can be created or edited in Microsoft Notepad or some other text editor. The first file I'll use is named "getstockdata1.txt." It will pull daily stock returns and trading volume for Walgreens and Coke and the market index described above using fixed dates, January 1, 2000 to January 31, 2000. **Very important** — this file needs to be in the directory C:\CRSP\work on your PC or it won't run!



The first line, "ENTITY," tells the program that what follows is the section of the file identifying the individual stocks for which we would like to access data. The LIST|PERMNO notation is one way of telling CRSP which firms we want. Another, shown below, is to read in the permnos and dates from a separate file. "End" tells the program that the list of firms (entities) is complete.

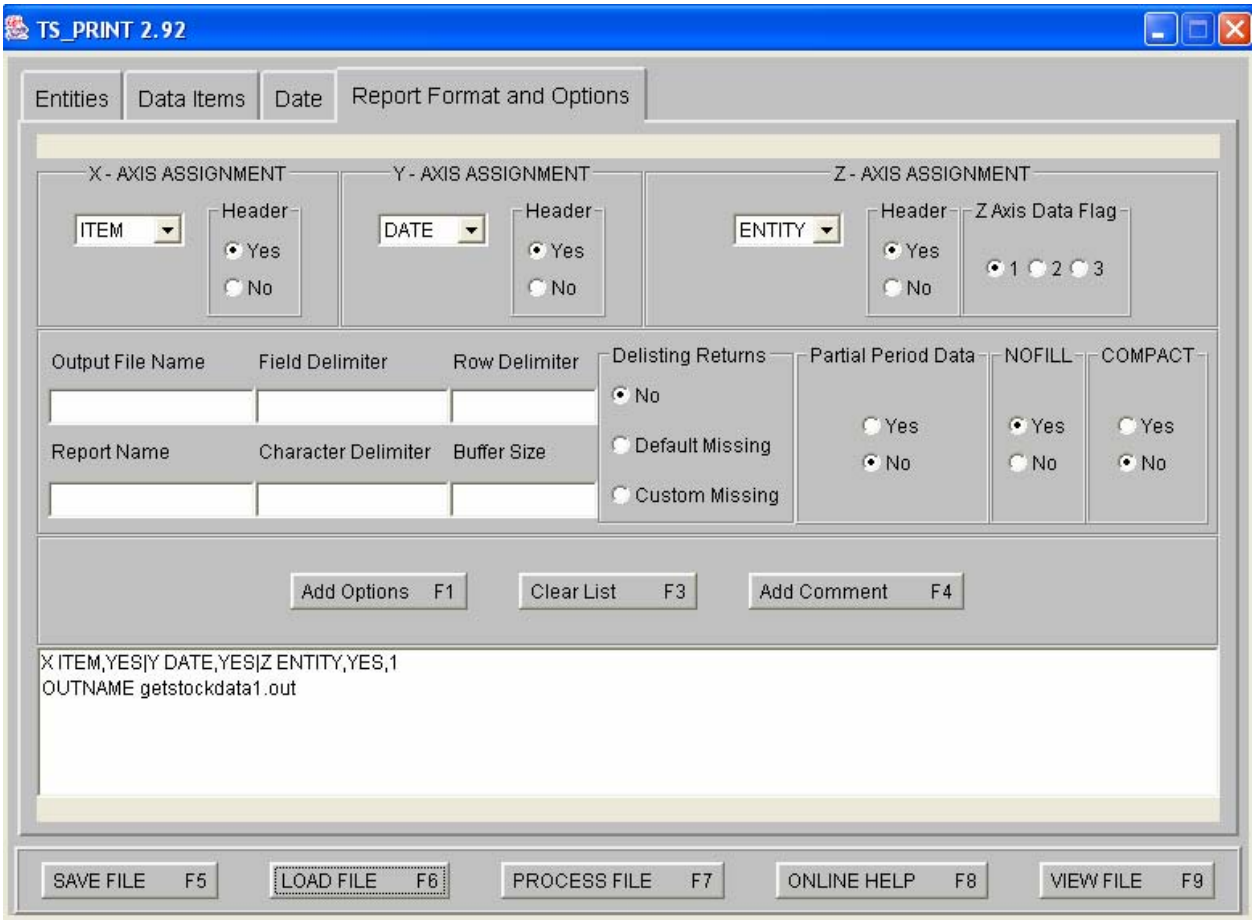
The next section of the program is the "ITEM" section, which identifies the data items we want. After each ITEMID listing we give the CRSP code for the item. Here "ret" is stock returns, "vol," is volume, and "indtret" is the return on a matching stock index. Interestingly (which means for reasons I can't explain), when CRSP is matching stock index returns to individual stocks, it views the index returns as a data item (rather than an entity). The syntax "indtret|SUBNO 1000081" is necessary for CRSP to properly line up the stock index return with the returns of the individual stocks (because we won't always pull data using fixed dates; sometimes we'll conduct an "event study," where each firm's data will be centered around a date

that's unique for each firm). As above, "END" tells the program that the ITEM section of the program is concluded.

Next we have the "DATE" section. Here the CALNAME items are the frequency of the data (DAILY in this case; could also be MONTHLY, QUARTERLY, or ANNUAL), the fixed RANGE of dates, and the format in which we would like the dates to be written.

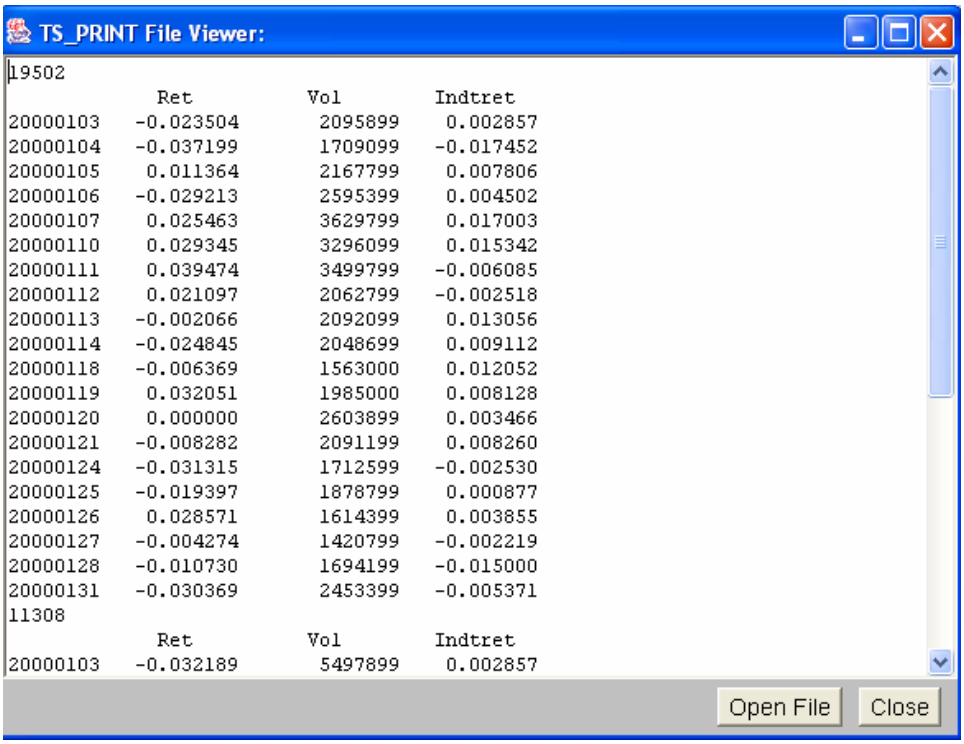
The last section, "OPTIONS," allows us to write out the data in different X, Y, Z aspects, name the output data file, and include other titles in the output data set (not shown because I don't want additional text labels in my dataset, just column headings).

Save the file using Notepad — I've named it "getstockdata1.txt." Then go to the last tab in TS Print (Report Format and Options). Choose "LOAD FILE" and browse the C:\CRSP\work directory to find your file. CRSP will default to looking for *.rqt ("request") files. Just override that and tell the program to look for *.txt files specifically, and it will. Load the file "getstockdata1.txt."



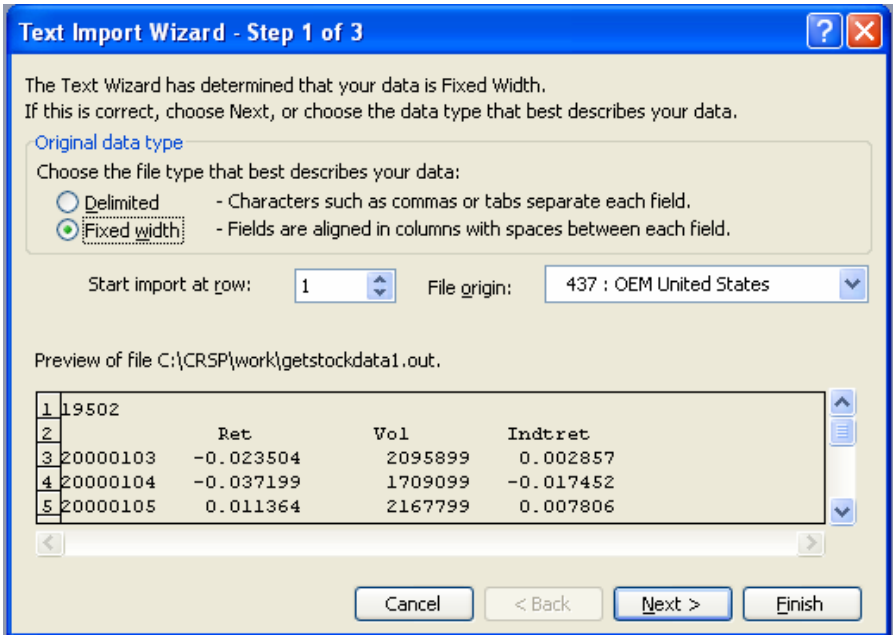
You get a partial look at your file in the TS Print window. Next select "PROCESS FILE" and choose the "PROCESS SCREEN DATA" option. If you get a message back that says "PROCESSING COMPLETE," you're in business. (If you get an error message, remember that debugging your own programs is a great learning experience.) Then select "VIEW FILE" and verify that you're happy with the data pulled by CRSP and the way the data were written out.

Here's what our data look like:

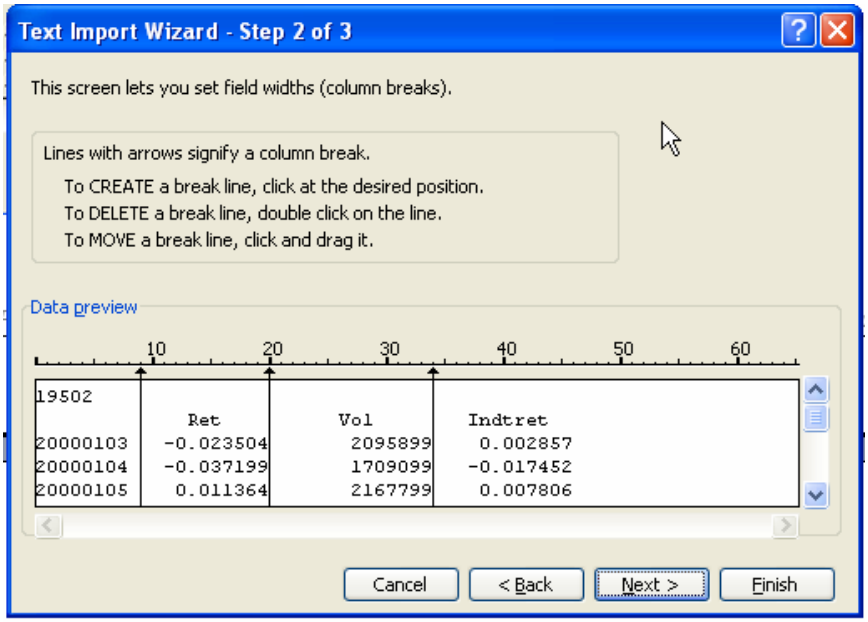


Only the partial file is shown in the window, but you can scroll through the entire file. The data are ready to be read into Excel (shown below). Notice that data for each firm is separated by the permno. (If you wanted to load the data file into SAS and run statistics using SAS's "BY" option, you could have also printed out each stock's ticker symbol as an ITEM.)

If you double-click on the output file "getstockdata1.out," Excel will open it, but it won't parse the data into columns. It's better to first open Excel, choose "File-Open," select "getstockdata1.out," and in the "Text Import Wizard — Step 1 of 3" window, for "Original data type," choose "Fixed width."



Click "Next" and Excel will make a suggestion regarding where to break the columns. You can drag the lines around if you don't like Excel's suggestions.



Click "Next" again and Excel gives you the chance to set the column formats (decimals, currency, percentages, etc.) from inside this window. Bypassing this option by clicking "Finish" does not preclude you from changing the formatting later. When you click "Finish" you'll see something like below (which I've truncated a bit, but you get the idea).

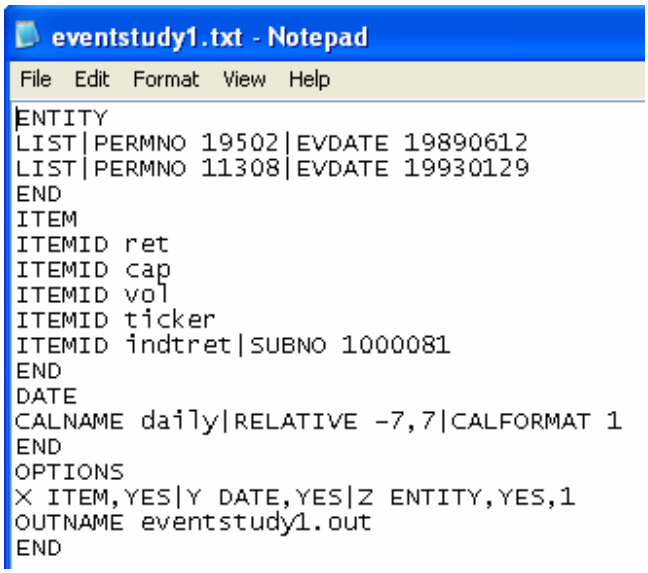
	A	B	C	D
1	19502			
2		Ret	Vol	Indtret
3	20000103	-0.0235	2095899	0.002857
4	20000104	-0.0372	1709099	-0.01745
5	20000105	0.011364	2167799	0.007806
6	20000106	-0.02921	2595399	0.004502
7	20000107	0.025463	3629799	0.017003
8	20000110	0.029345	3296099	0.015342
9	20000111	0.039474	3499799	-0.00609
10	20000112	0.021097	2062799	-0.00252
11	20000113	-0.00207	2092099	0.013056
12	20000114	-0.02485	2048699	0.009112
13	20000118	-0.00637	1563000	0.012052
14	20000119	0.032051	1985000	0.008128
15	20000120	0	2603899	0.003466
16	20000121	-0.00828	2091199	0.00826
17	20000124	-0.03132	1712599	-0.00253
18	20000125	-0.0194	1878799	0.000877
19	20000126	0.028571	1614399	0.003855
20	20000127	-0.00427	1420799	-0.00222
21	20000128	-0.01073	1694199	-0.015
22	20000131	-0.03037	2453399	-0.00537
23	11308			
24		Ret	Vol	Indtret
25	20000103	-0.03219	5497899	0.002857
26	20000104	-0.00887	3653599	-0.01745
27	20000105	0.019016	4728699	0.007806
28	20000106	0.001098	3564599	0.004502
29	20000107	0.065789	5735899	0.017003

The data are ready to be analyzed or otherwise manipulated inside Excel.

A Critical Step — Handling Missing Values. Despite CRSP's claim that they have the cleanest, most accurate data in the world, you will occasionally get missing values. You'll know you have a missing value when you get a value of −88 or −99 instead of a stock return, trading volume, price, or some other data item you were seeking. For this reason, you need to "clean" your CRSP data. You can do this in Excel by running your data through another spreadsheet tab with formulas in the cells such as =IF(OR(A1=−88,A1=−99),"",A1), which tells Excel to write nothing if the error −88 or −99 is encountered (the double quotes with nothing in between). If you're conditioning the data to read into SAS you could have put a dot (period) in between the quotes in the IF statement, since SAS likes it when missing values use the dot as a placeholder.

Just Clean It. If you don't clean your CRSP data you're going to make a mess. Let's say you're analyzing stock returns, with values like 0.014, and −0.007. Just think what it would do to your results to have a few −88's or −99's randomly strewn about. (If cleaning your data seems like an unreasonably onerous task just remember that since graduate school we've been conditioned to endure many disappointments, both large and small.)

Setting up an Event Study. Next we'll use a slightly different input file format to conduct an event study. The main difference here is that instead of using fixed dates for each firm, we're going to sample data before and after an "event date" that's unique for each firm in the sample. In the example below we'll pull some different data items for 7 days before and after each firm's event date. The file is called "eventstudy1.txt."

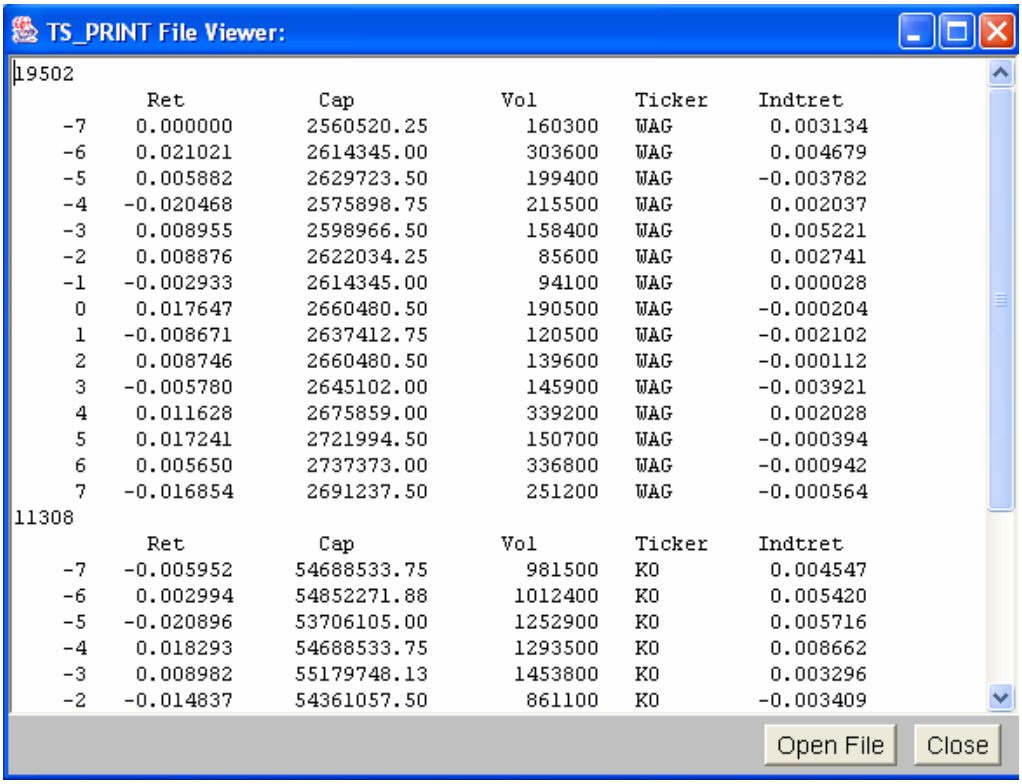


Notice how, in the "ENTITY" section of the program, we've added an additional item — an event date (EVDATE) for each firm. This is matched with a change in the CALNAME section. Instead of specifying a fixed RANGE as in the first program (same dates for all firms), we're telling CRSP we want our dates to be RELATIVE to the EVDATEs (which are unique to each firm). Specifically, we want daily data from days −7 to +7 relative to the EVDATEs. (If we had specified "monthly" after CALNAME instead of daily, and had −12,12 after RELATIVE, we would pull monthly data for the months −12 to +12 relative to the EVDATEs.)

In this file we're asking CRSP to access returns, market capitalization, and trading volume for our 2 stocks. We're also asking for something I alluded to above — the ticker symbol to be printed out in a column in case we want to run stats "BY" a variable in SAS (or SPSS). We're

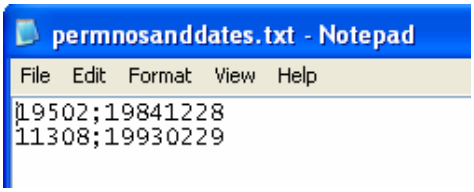
also asking for returns on a market index to be written out for days −7 to +7 relative to each firm's event date. Additionally, we're naming our output file "eventstudy1.out."

As above, Load the file, Process the file, choose the Process Screen Data option, and click OK when CRSP tells you that processing is complete. You can preview the file with CRSP's View File option:

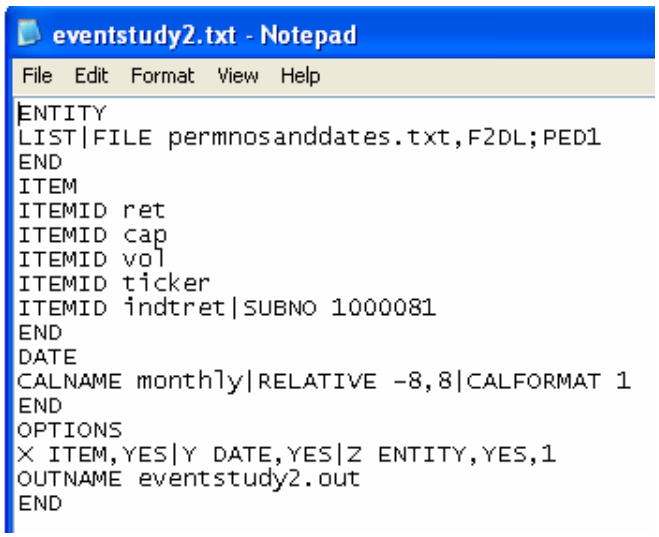


The file "eventstudy1.out" is ready to read into Excel as described above.

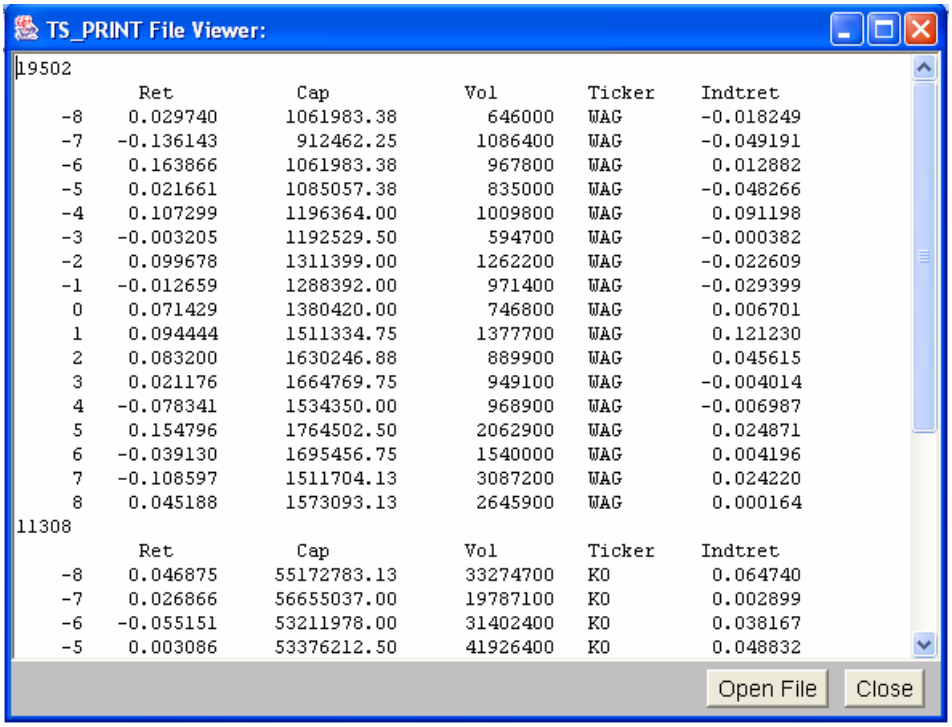
Reading in Permnos and Dates from an External File. When you're pulling data for a large number of firms, you might not want to clutter up your access file with all the permnos and dates. CRSP allows you to type out the firm information in an external text file and read it into the access file. I've created a separate file called "permnosanddates.txt" with permnos and event dates separated by a semi-colon:



I've edited the file "eventstudy1.txt" we used above to accommodate reading this info from a file instead of having it contained in the access file itself. It's reproduced below as "eventstudy2.txt." The only other change to the file is that now I've asked CRSP to access end-of-month data, 8 months before and after some new event dates. If you're motivated to ask what "F2DL;PED1" means … please don't! (Because I just copied it out of the CRSP manual and it worked.)



The output file is shown below. Notice that the returns are larger than the daily returns, as the time period spanned is one trading month (there are 21 trading days in an average month) instead of one day.



As was the case with all the other data files, these data are ready to read into Excel.

Where Do I Go From Here? The first thing I would do is get TS Print installed on your PC and replicate this tutorial. Next I would browse the PPT tutorial "tsprint_2005.ppt," which encourages you to meander through the TS Print interface. This will familiarize you with other variable names and options. The *best* next step, however, would be to conceptualize how market data such as these would complement a research project you're working on or contemplating.