Paper 167-2007

CRSP® Indices in SAS®

Taras Zlupko, CRSP, University of Chicago, Chicago, IL

ABSTRACT

About two hundred US stock and Treasury financial indices and benchmark portfolios are available from the Center for Research in Security Prices (CRSP). The breadth of application of CRSP indices in financial research as well as success of index-based financial products offered by the investment industry, stimulate further interest in the indices. SAS environment provides flexible access options and powerful analytical solutions when working with CRSP financial indices.

CRSP stand-alone index product has been available in SAS format for some time. The advent of the SASECRSP interface engine has introduced new perspective of the CRSP index data access and analysis. The engine allows working with data provided in the proprietary CRSPAccess format directly from SAS and enables additional options of data presentation. This flexibility can be employed to increase the efficiency of data processing and enrich overall research experience when working with CRSP indices and particularly when analysis is conducted in conjunction with security level data.

CRSP index data availability and access options in the SAS environment are discussed and contrasted. Existing SAS solutions applicable to indices analysis are exemplified. Some financial index performance statistics introduced.

CRSP INDICES

CRSP calculates broad range of indices which differ by the universe of securities used, rebalancing frequency, weighting type and other characteristics defined by particulars of methodology. Every index is related to more than one series of data items. This may include variations of index levels based on different types of returns and returns themselves, statistics related to securities counts included in the index and index rebalancing, index weights and other items.

Daily and monthly data frequencies of CRSP Indices are in many instances going back historically as far as to 1925 providing unique opportunity for analysis. CRSP documentation indicates that daily and monthly index returns are calculated based on daily and monthly security holding period returns respectively. These are reflective of the two versions in which CRSP stock and US Treasury databases are available, namely daily and monthly. Monthly index returns are further used in the compounding process to calculate quarterly and annual index values which are available in stand-alone index product from CRSP.

Twelve index groups and portfolio types are defined for CRSP stock indices. They differ on the basis of the securities used such as being restricted to the universe of NYSE, AMEX, NASDAQ or their combination, rebalancing calendar and type of statistics used for rebalancing. Annual or quarterly periods are used in indices rebalancing. Market capitalization, standard deviation or betas are used as rebalancing statistics. Other distinctions among indices may be drawn in that some indices are calculated in market capitalization value-weighted and equal-weighted formats and some in one of the above formats only. Some of the indices are calculated off both daily and monthly CRSP databases and some - using only one of the databases. For example, the US Treasuries indices and inflation series included in the index product are only available in monthly frequency. Another example of index differences is between stock file capitalization deciles and risk-based deciles. The former is a value-weighted index and the latter – the equal-weighted index.

CRSP indices are geared to perform a variety of analysis. Given the multitude of indices, it is important to make sure the nuances of index methodology is understood before selecting an index for further work. CRSP Data Description Guide mentioned in the references section of this paper is the source to be consulted for further details.

As mentioned in the Guide, major index types available from CRSP are:
CRSP Stock File Indices, including:
- CRSP Market Indices
- Published S&P 500 and NASDAQ Composite Index Data

CRSP Stock File Capitalization Decile Indices
CRSP Stock File Risk-Based Decile Indices
CRSP Cap-Based Portfolios
CRSP Indices for the S&P 500 Universe
CRSP US Treasury and Inflation Series

The ranges of indices related to individual exchanges are dependant on exchanges coverage in CRSP stock database. The New York Stock Exchange (NYSE) series begin on December 31, 1925, the American Stock Exchange (AMEX) - July 2, 1962 and the NASDAQ Stock Market (NASDAQ) - December 14, 1972. The series containing combinations of exchanges begin at the earliest point that data for any of the exchanges is available. Treasury Bill and the Consumer Price Index series begin on December 31, 1925. Treasuries of longer maturities begin in early 1940th.

ACCESS TO CRSP INDICES IN SAS

CRSP indices are available in SAS directly either as native SAS data sets or through the SASECRSP engine. The latter presents CRSPAccess formatted data in familiar SAS tables. The stand-alone index product from CRSP includes a version of the indices in SAS format. If CRSP indices are provided in conjunction with one of the stock market databases the data is in CRSPAccess format and accessible through the SASECRSP engine. Different versions of CRSP deliverables introduce material differences in index data presentation and contents which bears on the ways the data is accessed and further processed. Some of these differences are discussed below.

1. Stand-alone indices.
As mentioned earlier, CRSP stand-alone index product is available in SAS format. All SAS index datasets included in the stand-alone product are located in the same directory. Therefore setting up the library referencing physical location of data provides access to all of the indices. Most of the indices are available in daily, monthly, quarterly and annual frequency. Rebalancing information of CRSP cap-based portfolios is also included.

The following statement sets up the indices library in the stand-alone CRSP index product where c:\ind\ is the physical location of SAS data:

```
LIBNAME ind 'c:\ind\';
```

Figure 1 below shows the layout of data in SAS Explorer window.

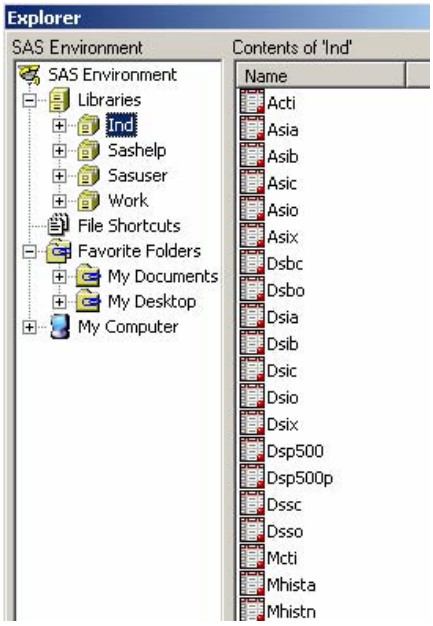

Figure 1. Stand-alone CRSP index product in SAS Explorer.

Most of CRSP indices in the stand-alone product are available in four pre-calculated frequencies – daily, monthly, quarterly and annual. Data frequency of a particular data set is determined by first character of its name. For example, 'A' indicates annual data, 'D' – daily, 'M' – monthly. Quarterly data set names start from 'Q' (not shown in Figure 1). Not all indices are available in all four aforementioned frequency formats. For example, risk indices are provided in daily frequency only, US Treasury indices - in monthly, quarterly and annual frequencies, and the cap-based portfolios are monthly. It is worthy noting that it is possible to arrive to lower frequency data from higher frequency indices but not vice versa.

As mentioned earlier, CRSP stock databases are produced in daily and monthly formats. Although the data is reflective of the same markets and same securities universe, the frequency difference influences some database content. Discussing this is beyond the scope of this paper. However, it is important to note that since the daily and monthly indices are produced off databases of relevant frequency, the indices may not exactly translate into each other. Quarterly and annual indices are calculated from monthly data. Therefore, although daily index series may not necessarily exactly compound into lower frequency data, quarterly and annual series are fully derivable from relevant monthly series.

The code below shows a way to arrive to quarterly returns from monthly returns for CRSP value-weighted total return index of NYSE for the entire history of the index. First observation in monthly dataset is omitted since index returns unlike index values are not available on first date. Then monthly returns are compounded over three-month periods. Only quarter-end return values are retained in the final dataset.

```
data mq_compret;set ind.msia(keep=caldt vwretd);
    retain q mcompret;
    if _n_=2 then do;q=0;mcompret=0;end;
       else; if q<3 then do;q=q+1;mcompret=(mcompret+1)*(vwretd+1)-1;end;
                    else do;q=1; mcompret=vwretd;end;
                           format caldt date9.;
                                  if q=3;
run;
```

Using SAS code such as shown above, it is possible to create lower frequency return series for other CRSP indices if necessary.

The cap-based portfolios data sets include both time-series data (names starting Mhist* in Figure 1) as well as quarterly rebalancing information of the portfolios.

Files containing Market Indices and Decile Portfolios as well as the risk indices files in the stand-alone index product represent particular universe of securities based on specific market or exchange or combination of thereof. Each data set contains fields related to the entire base universe as well as deciles. CRSP unique index identifier – INDNO® – is not included in the stand-alone index product. This and some other naming conventions require attention when processing stand-alone index product data in SAS.

2. Indices in CRSPAccess format.

When CRSP indices are provided together with the stock databases they are available in the CRSPAccess format and are accessible through the SASECRSP engine. The engine allows using CRSP unique index identifier, processing indices in series or groups and accessing rebalancing information not only for cap-based portfolios but for other indices. Some new features of the engine were reported by the author at recent SUGI meetings. Accessing indices warrants some additional discussion.

Either daily or monthly frequency of indices is available when accessing CRSP indices using the SASECRSP engine. As mentioned earlier, this is reflective of the two standard versions of CRSP stock database. The frequency of indices will depend on the version of stock database available to user. Access to different frequencies of indices in series or in groups using the SASECRSP engine requires setting up libraries with relevant SETID. Four SETIDs are available based on the combination of options to access the indices data in CRSPAccess format via the SASECRSP engine as follows:

400 Monthly Index Groups
420 Monthly Index Series
440 Daily Index Groups
460 Daily Index Series

The following statements will set up the daily and monthly Index Series libraries:

```
LIBNAME  dind  SASECRSP  'c:\dstk\'  setid=460;
LIBNAME  mind  SASECRSP  'c:\mstk\'  setid=420;
```

SAS data sets of the CRSP daily index series library are shown in Figure 2.

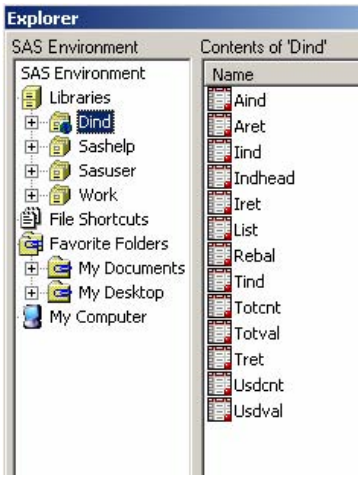

Figure 2. CRSP indices accessed through SASECRSP engine.

As seen in Figure 2, SAS data sets are different from those available in the stand alone index product shown in Figure 1. The data sets in Figure 2 are primarily organized by data item. For example, Aind data set contains capital appreciation (price) index values, Aret contains Aind returns, Tind contains total return index values, Tret contains total returns of Tind and so forth. In addition rebalancing information is available in the Rebal data set and includes rebalancing dates, count of securities, smallest and largest statistics, and other fields. The Indhead data set contains index header information such as unique index identifier (INDNO) and index descriptive information of all indices available in daily indices database.

Monthly index series are organized the same way as daily series although availability of particular indices is not the same. Some of the monthly indices are monthly version of daily indices and some, such as cap-based portfolios, are available in monthly frequency only.

Data item based data sets are common distinctive feature of both index series and groups available through SASECRSP engine comparing to the stand-alone index product. The difference between index series and index groups is in that that all indices in index series are located consecutively and are sorted by index identifier whereas index groups data sets present index groups consecutively but indices in groups in separate fields across data set.

Index group libraries are set up identically to index series libraries with appropriate SETID used. The data sets are organized the same way as shown in Figure 2 for index series, although internal structure of the data sets is different reflecting the contents of index groups. Examples of index groups include market capitalization deciles, cap-based portfolios, and beta or standard deviation deciles.

SASECRSP allows restricting index database to specific universe of indices and date ranges. This is a valuable efficiency feature particularly when multiple iterative processes are executed. For example, the following statement sets up library containing one index series only – CRSP monthly NYSE Value-Weighted Market Index – restricted to period from January 2001 to December 2006:

```
LIBNAME  mind  SASECRSP  'c:\mstk\'  setid=420
          range='20010101-20061229' indno=1000000;
```

ANALYZING INDICES

SAS represents efficient environment to work with CRSP indices. Availability of both CRSP stand-alone files in SAS format and accessibility of the indices in CRSPAccess format via the SASECRSP engine provides enhanced flexibility in using the data.

A popular distinction in evaluating stock market performance is in using equal- and value-weighted indices. CRSP market indices are calculated in both aforementioned versions. Plotting the indices provides first insight about them. Figure 3 shows CRSP monthly equal- and value-weighted total return indices based on the universe of common stocks listed on the New York Stock Exchange from December 1925 to December 2006.

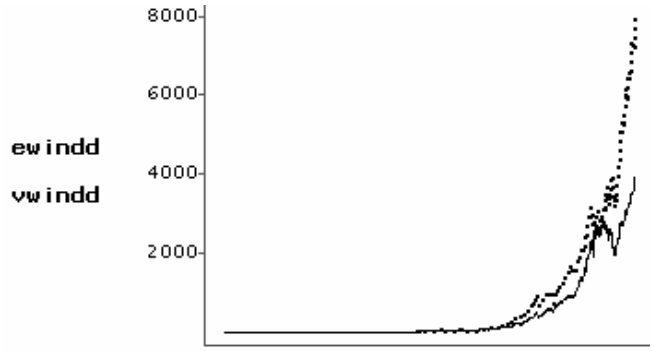

Figure 3. CRSP total return NYSE market indices, 1925-2006.

The bottom line in Figure 3 represents the value-weighted NYSE index whereas the top line is the equal-weighted index. The long eighty one year history of the indices shown in the chart hides an important detail that CRSP market indices are rebased to 100 on December 29, 1972. Therefore ending values of indices in Figure 3 are reflective of that date rather than the beginning date of indices history. This is clearly visible by plotting the earlier indices period separately. Figure 4 shows the indices over about first one third period of their history currently available on CRSP.

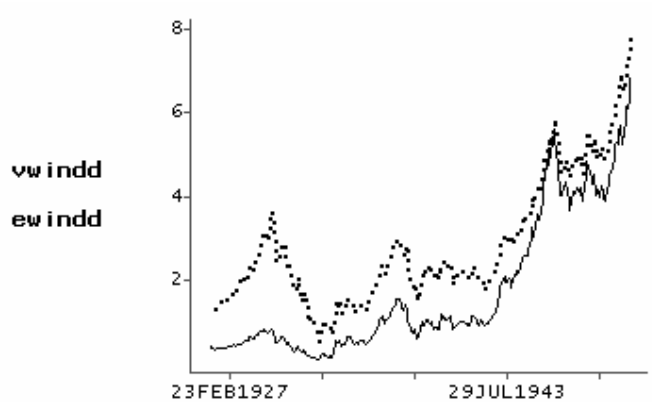

Figure 4. CRSP total return NYSE market indices, 1925-1950.

As seen in Figure 4, indices do not start from the same level. Top dotted line in Figure 4 is the value-weighted index. The chart seems to indicate that contrary to the overall 1925-2006 period the value-weighted index is performing better than the equal-weighted index during the years 1925-1950.

To make comparison of indices over specific date range convenient and reliable the indices need to be set at the same level at the beginning of the period of interest. Once the performance of indices is evaluated from December 31, 1925 as is the case of charts presented in Figure 3 and Figure 4 then indices are rebased to same value on that date. Following is an example of SAS code rebasing the aforementioned NYSE indices to 100 on their starting date in CRSP:

```
data msia_rebased(keep=caldt vwindd ewindd vw ew);
   set ind.msia;
      retain vwfac ewfac;
      if _n_=1
             then do;vw=100;ew=100;vwfac=vw/vwindd;ewfac=ew/ewindd;end;
             else do;vw=vwindd*vwfac;ew=ewindd*ewfac;end;
```

```
                        format caldt date9.;
run;
```

Rebased values of equal-weighted and value-weighted CRSP NYSE market indices during the 1925-1950 period are plotted in Figure 5 below as ew and vw lines respectively. The top dotted line represents the equal-weighted index. Being related to the same base value at the start of the period, performance of both indices is directly comparable.

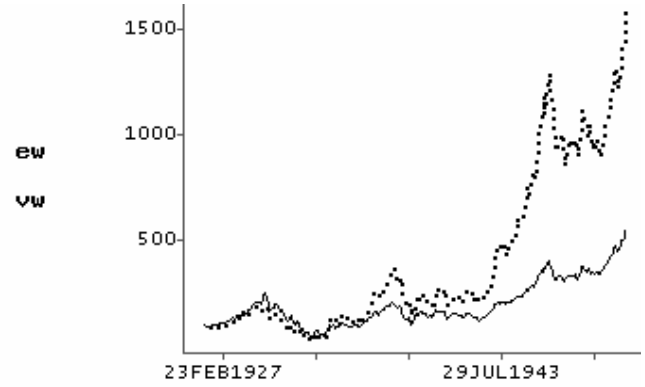

Figure 5. CRSP total return NYSE market indices, 1925-1950, rebased.

As seen in Figure 5, the value-weighted index did not outperform its equal-weighted counterpart during the period under consideration. On the contrary, the equal-weighted index posted much higher gains than the value-weighted index overall and particularly in the second half of the 1925-1950 period.

SAS code above can be used to rebase the equal-weighted and value-weighted CRSP NYSE market indices over the entire period from December 1925 to December 2006. The rebased indices are plotted in Figure 6 were the dotted line represents the equal-weighted index.

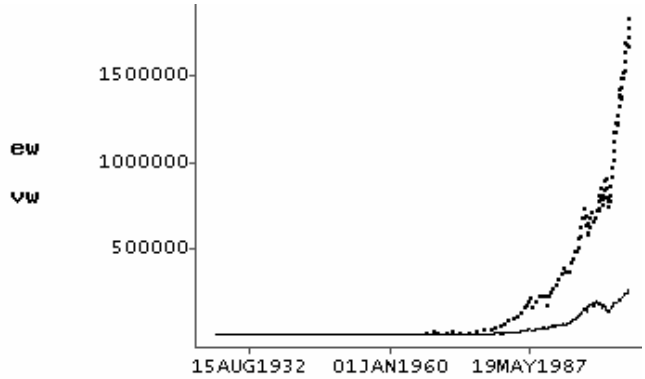

Figure 6. CRSP total return NYSE market indices, 1925-2006, rebased.

While differences between the indices in earlier years in Figure 6 are dwarfed and indistinguishable due to the long eighty one year history of the indices, the change in scale of the y-axis indicating index levels is easily noticeable. Since both indices started from 100 base values in 1925, levels of rebased indices are much higher throughout their history and at the end of 2006. Same rebased starting values of indices make comparisons to the beginning of the period and between the indices more consistent.

Another comparison of interest to investors is between capital appreciation and total returns portfolios. Relevant series are available from CRSP. Figure 7 gives an example of this comparison by showing performance of CRSP NYSE value-weighted total return and price only indices from December 29, 1972 when the series are set to 100 by CRSP.

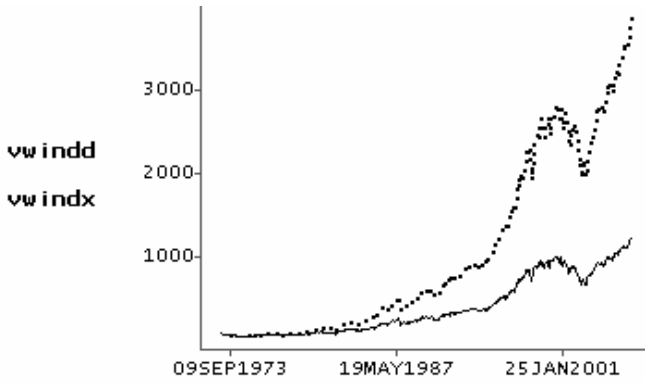

Figure 7. CRSP NYSE value-weighted total return and price only return market indices, 1972-2006.

Vwindd and vwindx series in Figure 7 are total return and price only indices respectively. These are the naming conventions adopted by CRSP in its stand-alone index product. Total return index is the top dotted line. The difference between the lines evidences the cumulative impact of dividends on returns.

While dividends are the obvious driver of total return indices comparing to price only index, the reason for difference in performance of equal-weighted and value-weighted indices is in the composition of the indices. Small capitalization stocks are made equal in weight to large capitalization stocks in the equal-weighted index. The value-weighted index weights the stocks in the index according to their market capitalization. Therefore performance of equal-weighted index is largely driven by smaller stocks whereas large capitalization stocks are the primary performance driver of the value-weighted index.

CRSP Stock File Capitalization Decile Indices make it possible to conduct more refined analysis of performance of stock portfolios based on market capitalization criteria than by using broad equal- and value-weighted indices. In the decile portfolios securities are ranked according to their market capitalization and then divided into ten equal parts each annual rebalancing period. Portfolio one includes the smallest stocks and portfolio ten – the largest.

All decile indices pertaining to a specific universe of securities and type of portfolio can be accessed as one group. This is done by using the INDNO option and relevant group SETID in the SASECRSP LIBNAME statement. For example, CRSP monthly indices database can be restricted to the NYSE Market Capitalization Deciles only and over the period from December 29, 1972 to December 29, 2006 in the following manner:

```
LIBNAME  mindgny  SASECRSP  'c:\mstk\'  setid=400          (1)
             range='19721229-20061229' indno=1000012;
```

Total return indices of deciles one and ten of the CRSP NYSE Market Capitalization Deciles are shown in Figure 8. Starting index values are set at 100 by CRSP.

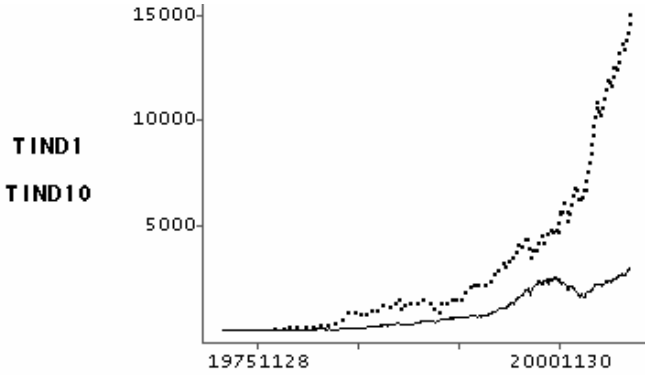

Figure 8. CRSP NYSE Market Capitalization Deciles one and ten, 1972-2006.

Tind1 and Tind10 series in Figure 8 are the total return indices of deciles one and ten respectively with top dotted line representing decile one. The chart confirms that small capitalization stocks show better long term gains than large stocks.

Further insight into behavior of CRSP stock market indices can be gained from exploring underlying returns. Figure 9 and Figure 10 show monthly returns of CRSP NYSE Market Capitalization Deciles one and ten respectively over the period 1972-2006. As mentioned earlier, return series are available from CRSP along with the index levels data.

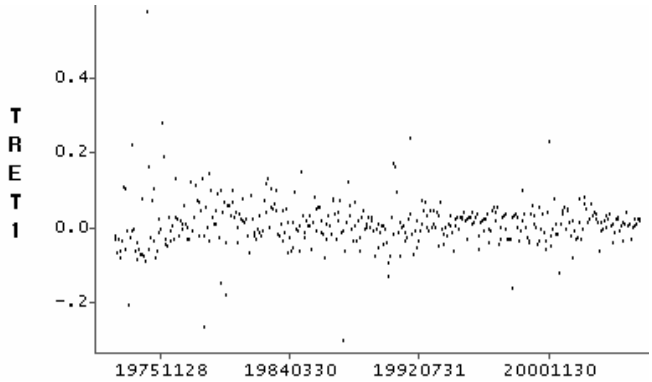

Figure 9. Returns of CRSP NYSE Market Capitalization Decile one, 1972-2006.

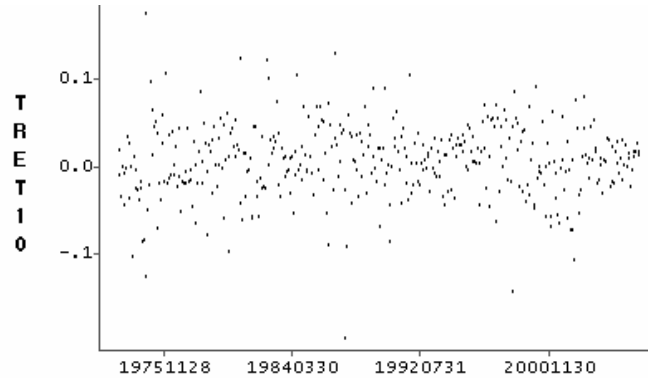

Figure 10. Returns of CRSP NYSE Market Capitalization Decile ten, 1972-2006.

Dispersion of decile ten returns representing large capitalization stocks (Tret10) is smaller than decile one returns (Tret1). Therefore, although small cap decile outperforms large caps as shown in Figure 8, small cap index may be perceived as more volatile and risky. Standard deviation can be used to quantify the dispersion of index returns. SAS code below employs the PROC MEANS to calculate mean and standard deviation of returns.

```
proc means data=mindgny.tret noprint;                          (2)
      var tret1 tret10;
            output out=tretstatg (drop=_type_)
                  mean= std= /autoname;
run;
```

Means of small- and large-capitalization portfolio monthly returns are 1.44% and 0.93% and standard deviations of monthly returns are 0.066 and 0.042 respectively. Therefore, although the mean return of small stock portfolio is higher, so is the risk of that portfolio.

Portfolio risk/return characteristics can be further evaluated using the Sharpe ratio. Assuming annual risk-free rate at 5%, following SAS code can be used to generate the Sharpe ratio for decile one and ten of CRSP NYSE Market Capitalization Decile indices:

```
data tretstatg;set tretstatg;                                    (3)
       expret1=tret1_mean*12;  risk1=tret1_stddev*sqrt(12);
       expret10=tret10_mean*12;risk10=tret10_stddev*sqrt(12);
              shrp1=(expret1-0.05)/risk1;
              shrp10=(expret10-0.05)/risk10;
run;
```

Mean and standard deviation of portfolio returns above are annualized to make these statistics comparable to the units of the risk free rate. The annualized values are assigned to new variables expret (expected return) and risk (standard deviation) for illustrative purpose. The Sharpe ratio is 0.54 for smallest decile and 0.42 for largest decile. Annualized expected total returns are 17.3% and 11.2% and the standard deviations are 0.23 and 0.146 respectively.

The code (1)-(3) above provides example of accessing the largest and smallest CRSP NYSE Market Capitalization Decile indices and calculating some index performance statistics using the index group option available in SASECRSP. The structure of the process in this case will be in many ways close to working with indices in the stand-alone CRSP product. This is reflective of the fact that both SASECRSP engine groups and the stand-alone index product represent deciles data as separate fields in SAS data set. Using index series modality of data access may in many instances be preferable. The code below replicates results of (1)-(3) in the index series access version.

```
LIBNAME  mindny  SASECRSP  'c:\mstk\'    setid=420                    (4)
             range='19721229-20061229' indno=1000002 indno=1000011;

proc means data=mindny.tret noprint;                                 (5)
             var tret; by indno;
                    output out=tretstat(drop=_type_)
                           mean= std= /autoname;

data tretstat;set tretstat;                                          (6)
             expret=tret_mean*12;risk=tret_stddev*sqrt(12);
                    shrp=(expret-0.05)/risk;
run;
```

Once the restriction on the universe of indices has been placed by the LIBNAME statement as shown in (4), further index selection in (5) and (6) is not necessary. The output of (4)-(6) presents each statistic in one column with index identifiers in rows. This may be particularly convenient when analyzing larger number of indices or indices not available in groups. For example, steps (5) and (6) will not require any modification when the number of indices changes. Mean, standard deviation and Sharpe ratio can be calculated for all CRSP monthly indices by simply removing the indno option in (4) and retain the rest of the code intact. Risk-return profile of all CRSP stock monthly indices from 1972 is shown in Figure 11.

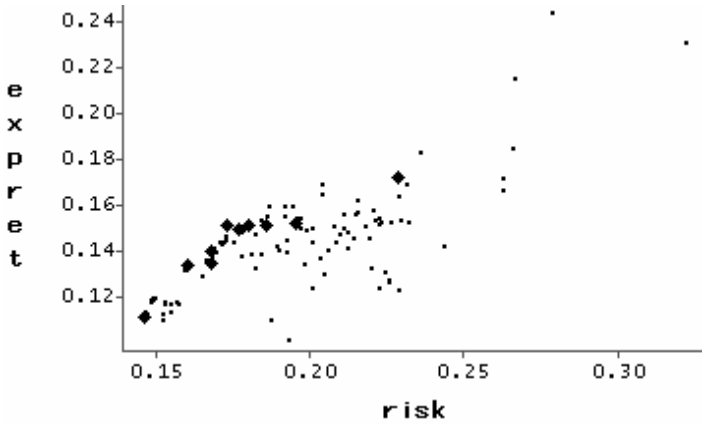

Figure 11. Risk and return of CRSP monthly stock market indices, 1972-2006.

Bold diamond markers in Figure 11 indicate ten CRSP NYSE Market Capitalization Deciles mentioned earlier. Expectedly, decile ten (largest securities) marker is the diamond on foremost left with low mean total return and standard deviation. Decile one is the foremost right bold diamond marker. Markers on far right represent different CRSP small stock portfolios. Overall, the markers appear nearly forming the familiar mean-variance-efficient frontier.

The above confirms that preferences in index series or index groups modalities of data access are dependant upon particulars of analysis and overall design of data processing. While index series access may work better to produce statistics for consecutive series, index groups may be advantageous when analyzing relationship among series within groups. For example, calculating correlation between smallest and largest deciles is more convenient using index groups.

The code below illustrates this for deciles one and ten of CRSP NYSE Market Capitalization Decile Indices. ODS graphics is employed with the CORR procedure to produce the scatter plots for variables as well as prediction ellipses.

```
ods pdf;                                                                (7)
ods graphics on;

proc corr data=mindgny.tret noprint plots=scatter (alpha=.05 .1);
       var tret1 tret10;
run;

ods graphics off;
ods pdf close;
```

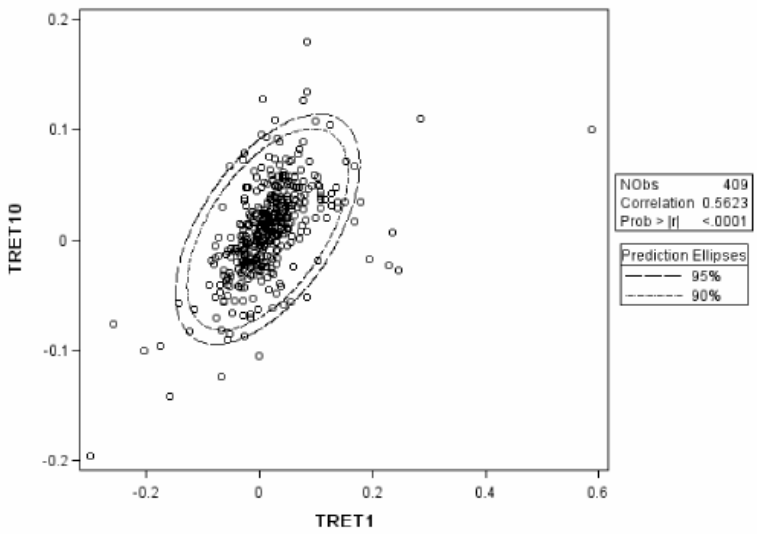

Figure 12. Returns and prediction ellipses of deciles one and ten.

The plot above shows monthly total returns of deciles one and ten of CRSP NYSE Market Capitalization Decile Indices from 1972. Pearson correlation between the two indices is 0.5623 with p-value less than 0.0001. The correlation value as well as prediction ellipses indicate moderate positive correlation. The confidence limits for prediction ellipses are set at 95% and 90%.

Relationship of different segments of financial markets can be further understood by looking into dynamics of changes of relevant statistics over time. For example, mean and standard deviation of returns can be calculated on a rolling basis. These can be moving five year period statistics recalculated every month. SAS implementation of the above can use the index group library set up in (1) and incorporate the previously developed application of PROC MEANS (2) into macro as follows:

```
%macro stats;                                                           (8)
         data tretstat_all;
                _freq_=.;
```

```
                tret1_mean=.;  tret10_mean=.;
                tret1_stddev=.;tret10_stddev=.;
    %do i=1 %to 349;
          data tret;set mindgny.tret;
                if %eval(&i)<_n_<=%eval(&i+60);
          proc means data=tret noprint;
                var tret1 tret10;
                      output out=tretstatg%eval(&i)(drop=_type_)
                             mean= std= /autoname;
          data tretstat_all;set tretstat_all tretstatg%eval(&i);
          proc datasets;delete tretstatg%eval(&i);
          run;
    %end;
  %mend;
  %stats;
```

Rolling means and standard deviations of returns use sixty monthly observations each time the statistics are calculated over the period 1972-2006. The values are plotted in Figure 13 and Figure 14.

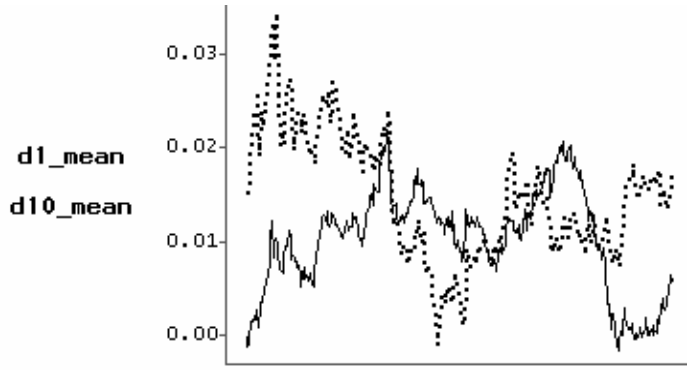

Figure 13. Rolling five year means of monthly returns of CRSP NYSE deciles one and ten.

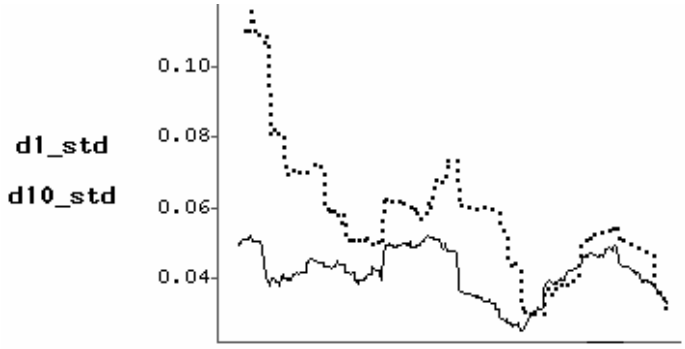

Figure 14. Rolling five year standard deviations of monthly returns of CRSP NYSE deciles one and ten.

Dotted line in Figures 13 and 14 represents CRSP NYSE market capitalization decile one (small stocks). Apparent is the decrease of standard deviation of small capitalization stock decile over the period under consideration to the levels of standard deviation exhibited by large capitalization portfolio.

CONCLUSIONS

A variety of financial indices is available from CRSP allowing multitude of options for studying US financial markets. CRSP offers some of the longest series of financial securities data with histories going back to 1925. The indices relate to different universes of stocks. Some well defined approaches have been employed in developing CRSP index methodologies such as equal- and market capitalization weighting, risk and size splits, different rebalancing periods and data frequency. The index values series have been

supplemented with pre-calculated total and price returns, portfolio counts and weights and other related data items.

CRSP financial indices are directly accessible in SAS. Specific goals of index analysis will determine most appropriate access and processing options and SAS tools used. A number of traditional analytical tools available from SAS can be employed without further modification to explore and analyze CRSP indices. SAS also allows for flexible implementation of user defined processes. This makes SAS powerful and efficient environment to work with CRSP indices.

REFERENCES

Center for Research in Security Prices. 2007. *Data Description Guide. CRSP US Stock and CRSP US Indices Databases. Version CA295.200701.* Chicago: The University of Chicago Graduate School of Business. http://www.crsp.chicagogsb.edu/support/documentation/pdfs/stock_indices_data_descriptions.pdf.

SAS Institute Inc. 2004. *SAS/ETS® 9.1, User's Guide*. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_91/ets_ug_7314.pdf

Zlupko, Taras. 2006. "Compass Rose in Stock Markets: The Power of SASECRSP LIBNAME." *Proceedings of the Thirty First Annual SAS Users Group International Conference*, San Francisco, CA, 183-31.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Taras Zlupko
CRSP, University of Chicago
105 West Adams Street, Suite 1700
Chicago, IL 60603
Work Phone: (312)263-6654
E-mail: Taras.Zlupko@crsp.ChicagoGSB.edu