Note on "Do Industries Lead Stock Markets" [1]

Harrison Hong
Princeton University

Walter Torous
MIT

Rossen Valkanov
UCSD

October 7, 2014

There has been increasing interest in our paper, "Do Industries Lead Stock Markets" (Hong, Torous, and Valkanov (2007), HTV (2007)). We have therefore decided to post data, code, and output files that replicate the main empirical result, namely, that some industry returns lead the market index return at monthly horizons. These results can be found in Table 3, column 1 of HTV (2007). The posted code also runs various robustness checks that are not part of the original paper. The replication and various specifications confirm the HTV (2007) finding that a significant number of industries lead the market. Perhaps the ultimate robustness check is whether this predictability holds out-of-sample. We therefore extend the 1946.01-2002.12 sample in HTV (2007) to 1946.01-2013.12 and present full-sample and rolling-regression results. Overall, we continue to see industry predictability. A subset of the same industries that predicted the market in HTV (2007) continue to predict in the extended sample. The subset of industries is smaller, due to some time-variation in the predictive relations, as suggested by the rolling-regressions. Once time-variation is taken into account (with rolling regressions), we note that a robust core of industries lead the market throughout the sample while another set of industries predict the market in various subsamples.

The main empirical question is whether certain industry returns $R_{i,t-1}$ lead the market index RM_t, after controlling for well-known predictors Z_{t-1} of market returns and lagged market returns, or:

$$RM_t = \alpha_i + \lambda_i R_{i,t-1} + A_i Z_{t-1} + e_{i,t} \quad (1)$$

We run regression (1) industry-by-industry for 34 industries (described below) with the following Z_{t-1} controls: lagged excess market return, inflation, default spread, dividend yield, and market volatility. The focus is on the λ_i coefficients in the regressions at one month horizon. Given that we run 34 regressions, we expect 3.4 of the 34 λ_i estimates to be significant under the null hypothesis of no industry predictability. Similarly, at the 5% level, we expect 1.7 of the 34 λ_i estimates to be significant. The question is whether we observe a larger number of industries that lead the market.

The results referenced in this Note are broken down into two parts: (i) in-sample replication and robustness with the 1946.01-2002.12 sample and (ii) extended-sample results with data spanning 1946.01 to 2013.12.

[1] We would like to thank Weikai Li (HKUST) and Yue Cao (UCSD) for independently replicating our results with Stata.

I. Replication and Robustness

I.1. Data: us_data_variables_used.xls

In our original paper, we used a 2002 WRDS download of CRSP, which we did not keep. We have a more recent 2005 WRDS download, which we use to recreate the dataset us_data_variables_used.xls for our paper. It is not entirely innocuous to use any download since each new download of CRSP makes potentially significant changes to the entire CRSP history. The 2005 download and the 2002 download were very similar, so there is little difference in the results. We also consider a 2014 download of WRDS and find similar results. As all the other variables in HTV (2007) are available from public sources, we had no trouble reconstructing them. Here are the specifics of the data:

Market returns proxies: We use three proxies for market returns.

1. CRSP value-weighted (NYSE/AMEX/NASDAQ) return, which was downloaded from WRDS in 2005. The sample used is 1946.01-2002.12.
2. The S&P500 return, downloaded from Amit Goyal's website, his original 2005 dataset. The sample used here is 1946.01-2002.12.
3. The CRSP value-weighted (NYSE/AMEX/NASDAQ) return, downloaded from WRDS in September 2014. The samples used here is 1946-2002.
   The correlation between the first and second, first and third, and second and third market proxies is 0.988, 0.986, 0.986, respectively. All proxies of the market returns are in excess of the risk-free rate and are denoted by RM_t.

Industry returns proxies:

1. Industry returns, downloaded from Kenneth French's website, old specification, available until 2004. The sample used here is 1946-2002. We eliminate five industries (Garbg, Steam, Water, Govt, Other) due to insufficient data.
2. The real estate return is NAREIT's total REIT return, obtained from NAREIT's website.

   The 33 French portfolio returns and the REIT return comprise our set of 34 industry returns. All industry returns $R_{i,t-1}$ are in excess of the risk free rate.

Controls:

1. Lagged inflation, default spread, dividend yield, and market volatility, as defined in HTV (2007). The data was downloaded in 2005.
2. From Amit Goyal's website, we download the Goyal and Welch (2007) predictors of market returns (original 2005 dataset). These predictors include inflation, default spread, dividend yield, and market volatility, term spread, book-to-market, and net equity expansion. We control for these predictors in the robustness section, as they have been widely used in the predictability literature.

There are small differences with respect to the original HTV (2007) data. Namely, the value-weighted index return, the industry returns, and the controls in us_data_variables_used.xls were downloaded in

2005, while the original data was downloaded in 2003. However, the differences in the portfolio returns and controls ought to be small.

I.2. Specifications and Results: Matlab file table3_various_gosp.m

Running the Matlab code produces the Excel output file with results from five specifications which differ with respect to market proxies and controls. We discuss all specifications below. For the purposes of this Note, we have highlighted in red industries that are significant at the 5% or 10% level. Industries that are significant in some specifications and close to significant in others (t-statistics above 1.3) are highlighted in yellow. The results are in file Output_2002_Posted.xls.To facilitate comparison with HTV (2007), we reproduce their exact estimates, t-statistic, and R2s from Table 3, column 1 in the file, columns B-D.

Specification 1: The first specification, which uses the CRSP value-weighted return as a proxy for market return, the old French industry returns, and lagged inflation, default spread, dividend yield, and market volatility as controls, is identical to the one in HTV (2007). The λ_i estimates, t-statistics, and R2s, are displayed in columns G-I of the output file. Given the persistence in the predictors and the heteroskedasticity in returns, all standard errors are Newey-West with one year's worth of monthly lags.[2]

The results from this specification are very similar to the ones in HTV (2007). The coefficient estimates in columns B and G are comparable in sign and magnitude. The correlation between the λ_i coefficients in HTV (2007) (column B) and the estimates in this replication (column G) is 0.93. More specifically:

- Out of the 34 industries, 14 are significant at the 10% level and 9 at the 5% level.
- In HTV (2007), 14 are significant at the 10% level and 12 at the 5% level.
- Out of the significant industries, 11 are the same as in HTV (2007): RLEST, STONE, APPRL, PRINT, PTRLM, LETHR, METAL, TV, RTAIL, MONEY, SRVC.
- Two other industries, MINES and TRANS, have high t-stats, but are not significant, while they are significant in HTV (2007). The industries TXTLS and INSTR are significant while they have high t-stats but are insignificant in HTV (2007).
- The biggest differences are in INSTR and UTIL. INSTR was positive and insignificant in HTV (2007) and is positive and significant in this data. UTIL is positive and significant in HTV (2007) and is positive but insignificant in this data.

Overall, we replicate very closely the results in HTV (2007). The small differences as we noted are due to the different versions (2002 versus 2005) of the WRDS downloads.

Specification 2: As a proxy for market return, we use the S&P500 return, obtained from Amit Goyal's website (original 2005 data). This proxy for market returns has been extensively used in the predictability literature in the last couple of decades and is of the same vintage as the CRSP value-weighted return in specification 1. We replicate the industry predictability with this market return, using our controls for inflation, default spread, market dividend yield, and market volatility. In other words, the only thing that changes with respect to specification 1 is the proxy for market return.

[2] There is a type in HTV (2007). They state that their Newey-West statistics are produced with three monthly lags, when in fact they are produced with 12 monthly lags.

Eleven industries are significant at the 10% level, and they are mostly industries that are significant in HTV (2007) and specification 1. Three industries, STONE and APPRL, and LETHR, are barely insignificant.

Specification 3: As in specification 2, we use the S&P500 return from Amit Goyal's website (original 2005 data) as a proxy for market return. However, we use the Goyal and Welch (2007) controls for inflation, default spread, market dividend yield, and market volatility, found in their (2005) dataset.

Twelve industries are significant at the 10% level, and they are mostly industries that are significant in HTV (2007) and specification 1. PTRLM, LETHR, and INSTR have high but insignificant t-statistics.

Specification 4: As in specification 2, we use the S&P500 return from Amit Goyal's website (original 2005 data) as a proxy for market return. However, we use an extended set of the Goyal and Welch (2007) controls. We add the term spread, book to market, and net equity expansion as additional predictors.

The results are very similar to those in the previous specifications.

Specification 5: We use the CRSP value-weighted return, downloaded in September 2014, as a proxy for market returns. Over the years, there have been changes in the construction of this index. We run the predictive regressions with our controls. Hence, the only thing that has changed with respect to specification 1 is the vintage of the proxy for market returns.

Nine industries are significant at the 10% level which is higher than the 3.4 expected under the null of no predictability, but five fewer than in HTV (2007) and specification 1. However, the industries that do predict the market are a subset of the industries that are significant in the previous specifications. Moreover, three more industries have high but insignificant t-statistics.

II. Extended-Sample and Rolling-Regressions Results: 1946-2013

II.1. Data: us_data_1946_2013_goyal_french_wrds.xls

For the extended sample, we use the following datasets.

Market returns proxies: We use two proxies for market returns.

1. The S&P500 return, downloaded from Amit Goyal's website, his updated 2013 dataset. The sample is 1946.01-20113.12.
2. The CRSP value-weighted (NYSE/AMEX/NASDAQ) return, downloaded from WRDS in September 2014. The samples is 1946.01-2013.12.

   The correlation between the two proxies of market returns in this sample is 0.987.

Industry returns proxies:

1. Industry returns are downloaded from Kenneth French's website, new specification, available until 2013.12. We eliminate the same five industries (Garbg, Steam, Water, Govt, Other) due to insufficient data.
2. As above, the real estate return is NAREIT's total REIT return.

The 33 industry portfolio returns (33 from French plus one from NAREIT) comprise our set of 34 industry returns.

Controls:

From Amit Goyal's website, we download the updated Goyal and Welch (2007) predictors until 2013. The predictors include inflation, default spread, dividend yield, and market volatility, term spread, book-to-market, and net equity expansion.

II.2. Robustness Results: Matlab file table_3_2013_gosp.m

We run three specifications with the extended 1946-2013 sample, the results of which are displayed in file Output_2013_Posted.xls. As in section I.2, we highlight in red coefficients that are significant at the 10% level, and in yellow, coefficients that have a t-stat of 1.3 or higher (p-value of 0.2).

Specification 1: We use the S&P500 return as a proxy for market return. The controls are inflation, default spread, market dividend yield, and market volatility. This corresponds to specification 2 in the previous section.

Specification 2. Same proxy for market return, the S&P 500 returns, but we extend the number of controls to include the term spread, book to market, and net equity expansion. This is specification 3 in the previous section.

Specification 3: The CRSP value-weighted return is proxying for market return. The controls are inflation, default spread, market dividend yield, and market volatility from Goyal's website. This is specification 5 from the previous section.

In all three specifications, 7 out of 34 industries predict the market at the 10% level. The significant industries are STONE, SMOKE, PRINT, PTRLM, TV, RTAIL, MONEY (Specification 1). Three more industries, LETHR, METAL, SRVC have high t-statistics, but are not significant at the 10%. These industries are almost a perfect subset of the ones that significantly predict the market in the shorter sample and HTV (2007). From that perspective, there is a robust core of industries, whose returns lead the market return.

The number of industries that lead the market is higher than the expected 3.4 industries under the null but is lower than the results in the 1946-2002 samples. In the next subsection, we explore time-variation and instability in the regressions.

II.3. Stability of Coefficient and Rolling Regressions: Matlab file rolling_estimates_2013_gosp.m

We estimate regressions (1) over rolling windows of 40 years. The rolling regression estimates $\lambda_{i,t}$ allow us to explore the stability of the predictive regressions. Instability of the coefficients might be due to purely statistical issues (sample representativeness, outliers, etc.) or changes in the underlying economic structure.

In the rolling regressions, we first use the 1946.1 to 1986.1 sample to estimate the first set of coefficients. Then, we roll the sample by one month and estimate another set of coefficients over the sample 1946.2 to 1986.2, and so on. In this fashion, we obtain a time series of the estimates $\lambda_{i,t}$ for each industry, and corresponding time-series of Newey-West t-statistics for the period 1986.1 to 2013.12. All regressions are run with the controls inflation, default spread, dividend yield, and market volatility. The data file is us_data_1946_2013_goyal_french_wrds.xls. The program produces five figures.

In Figures 1-3, we plot the time series estimates of $\lambda_{i,t}$ for the 14 industries that were significant in HTV (2007).[3] Results for the first five industries (RLEST, MINES, STONE, APPRL, and PRINT) are shown in Figure 1, for the second five (PTRLM, LETHR, METAL, TRANS, TV) in Figure 2, and for the last four (UTILS, RTAIL, MONEY, and SRVC) in Figure 3. The top plot in each figure graphs the estimates while the bottom one displays the Newey-West t-statistics.

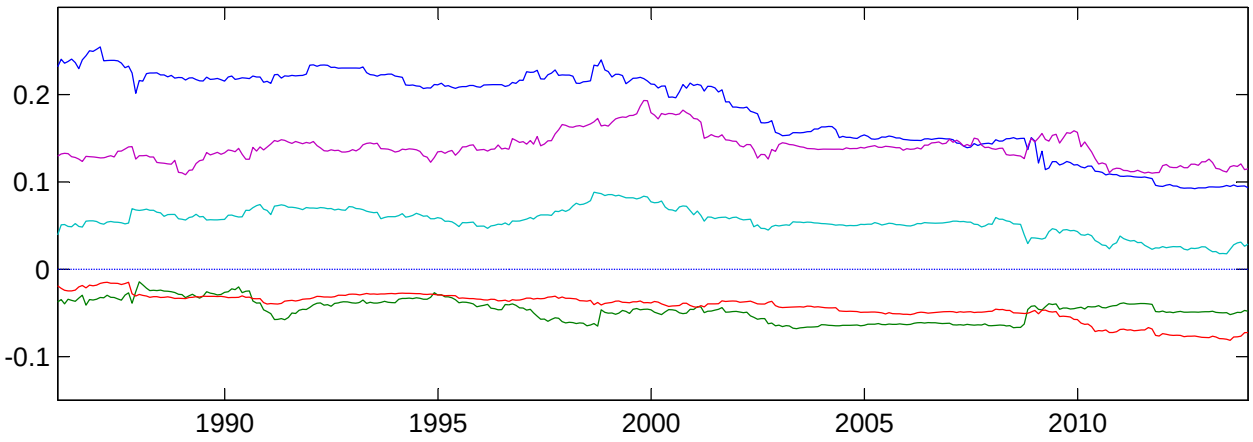
In all three figures, we observe stability in the sign of the estimates of $\lambda_{i,t}$. None of the coefficients change sign over the period. We do not observe severe instability or variations, although for several industries, the magnitude of the coefficients changes over the sample. For several industries, the coefficients are always significant. For others, the coefficients are significant over shorter subsamples.

In Figure 4, we plot the number of industries in the rolling regressions that are significant at the 10% level. In the 1980s and early 1990s, about 7 to 8 industries forecast the market. The number increases dramatically in the late 1990s and early 2000s (the HTV (2007) sample) to 14. It then decreases to about 10-11 industries. These numbers are higher than the 7 significant industries that we observe in the entire 1946-2013 sample. This is mainly due to the fact that while there is a core of about 7 industries that predict the market over the entire sample, a few others are significant over shorter horizons. The rolling regressions capture, in a sample fashion, the time-variation in these additional industries.

In Figure 5, we display the number of period that an industry has significant $\lambda_{i,t}$ coefficient in the rolling regressions. As pointed out before, some industries, such as RLEST, PRINT, PTRLM, TV, and MONEY have a significant $\lambda_{i,t}$ over almost the entire rolling sample. Others, such as MINES, STONE, APPRL, LETHR, RTAIL, and SRVC are significant in large number of periods.

[3] The program can be used to plot $\lambda_{i,t}$ for all 34 industries.

Figure 1: Rolling λ_i Estimates for RLEST(blue), MINES(green), STONE(red), APPRL(cyan), PRINT(magenta)



t(NW) statistics for RLEST(blue), MINES(green), STONE(red), APPRL(cyan), PRINT(magenta)

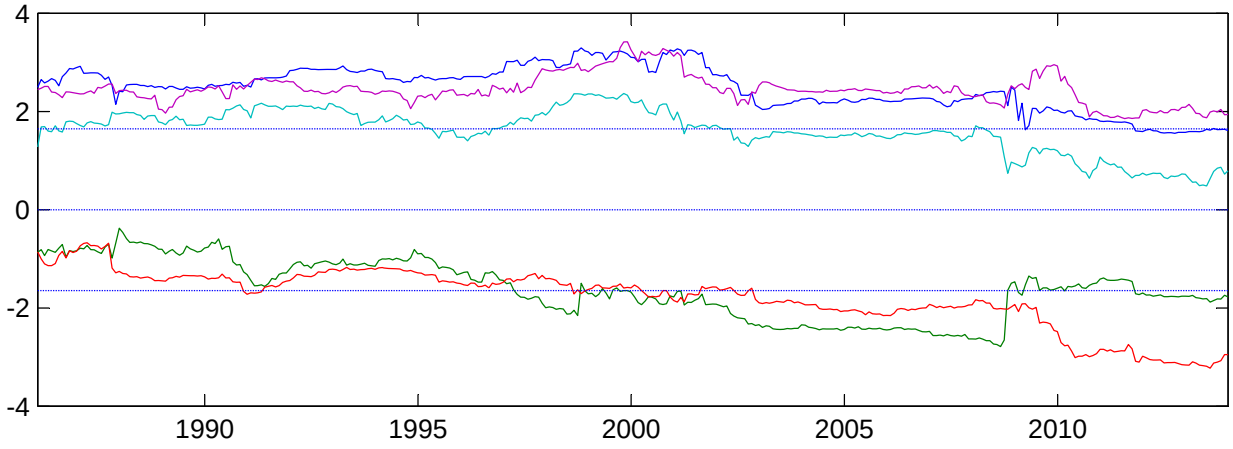
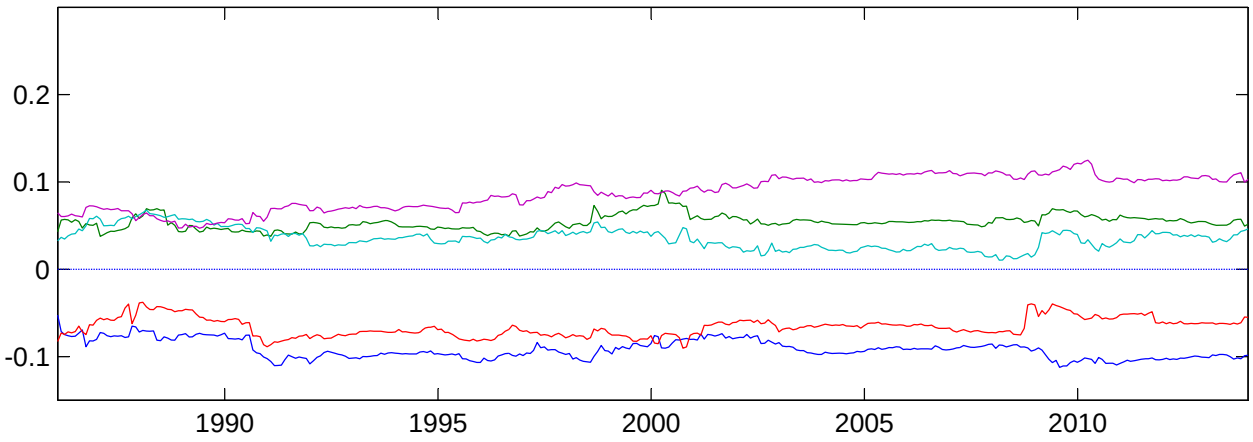


Figure 2: Rolling λ_i Estimates for PTRLM(blue), LETHR(green), METAL(red), TRANS(cyan), TV(magenta)



t(NW) statistics for PTRLM(blue), LETHR(green), METAL(red), TRANS(cyan), TV(magenta)

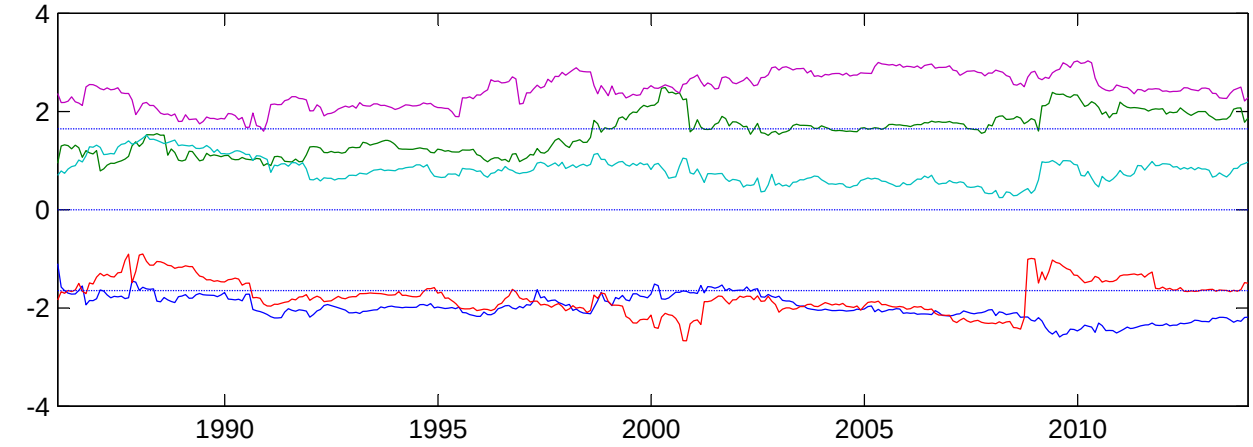
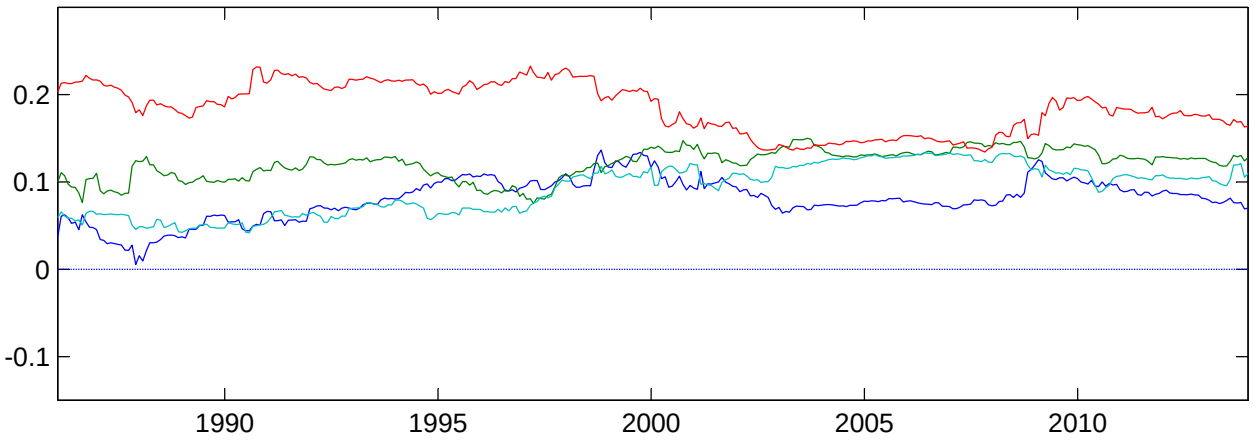


Figure 3: Rolling λ_i Estimates for UTILS(blue), RTAIL(green), MONEY(red), SRVC(cyan)



t(NW) statistics for UTILS(blue), RTAIL(green), MONEY(red), SRVC(cyen)

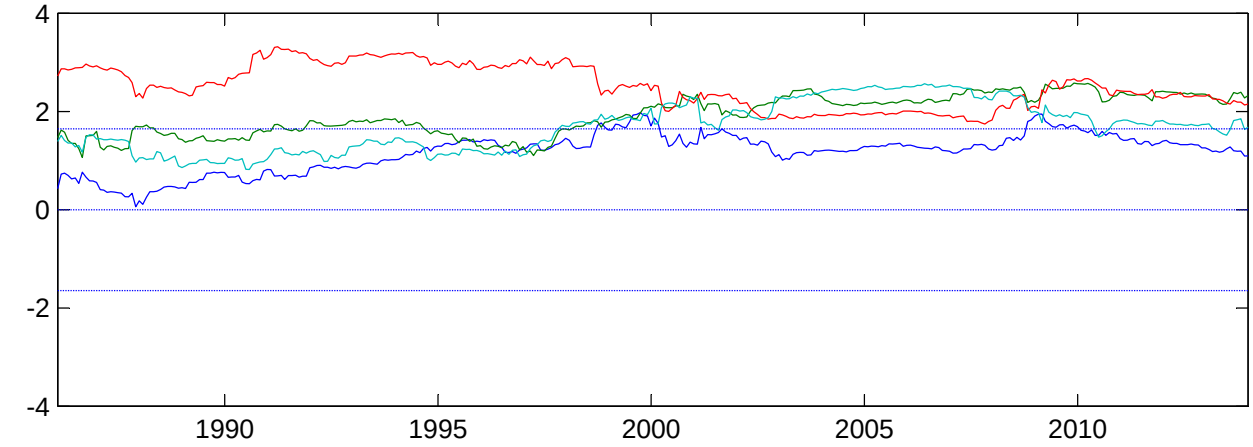


Figure 4: Number of Significant Industries (Rolling Regressions)

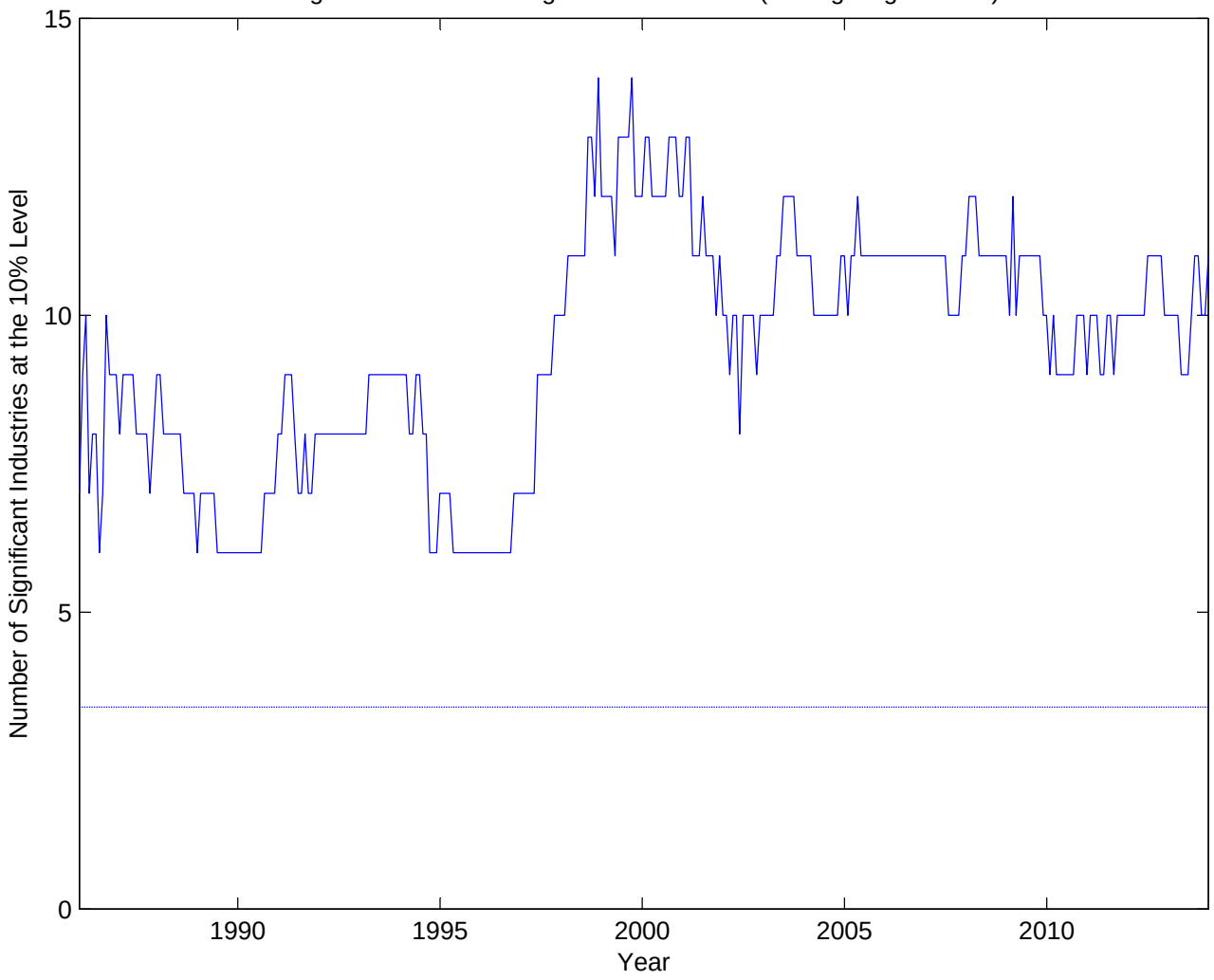


Figure 5: Number of Periods with Significant t(NW) (Rolling Regressions)