Performance Tests of Insight, ESG Momentum, and Volume Signals


Stephen Malinak, Ph.D.
Chief Data and Analytics Officer
TruValue Labs

Greg Bala
Lead Data Scientist
TruValue Labs

November 2017
# Table of Contents

3. Executive Summary

4. Quantifying Intangible Risk Factors

7. Signal Testing Methodology

10. Portfolio Strategy Development

11. Performance Summary

20. Plans for Future Work

21. References
Executive Summary

This whitepaper tests the effectiveness of timely Environmental, Social, and Governance (ESG) signals as screening tools and quantitative “alpha” factors for large-cap U.S. stocks included in the S&P 500 benchmark over the past five years. TruValue Labs (TVL)’s timely signals scored by artificial intelligence constitute a new tool for the investment industry.

Companies with strong information flow, good overall ESG scores (the TVL Insight Score), and positive ESG momentum (the TVL Momentum Score) outperform the S&P 500 benchmark over a five-year period, adding about 3% to 5% (300-500 bps) of alpha annually. The signals also show low enough correlation to traditional quant signals, such as value, momentum, quality, and low volatility to be additive in a multi-factor combination. These results show the signals are effective tools for stock picking purposes and portfolio construction. ESG alpha findings also serve to suggest longer-term, fundamental value adds for ESG factors.
Quantifying Intangible Risk Factors

TruValue Labs’ mission is to build a data platform for investment professionals that harnesses artificial intelligence technology to overcome major hurdles that confront the industry in analyzing intangible risk and extra-financial factors such as environmental, social, and governance (ESG) issues.

The Shortcomings of Traditional ESG Signals

The Traditional ESG signals suffer from numerous hurdles.

1. ESG data is largely sourced from company materials for many raters, and scores are too dependent on disclosure levels for some.
2. Scores and underlying data are time-lagging, often by a year or more.
3. Different providers look at different measures, and have differing definitions of many overlapping measures—as a result, there is a low correlation between scores from different data providers for the same companies over the same period.

Dr. Jim Hawley investigates this lack of correlation and suggests what might be driving it [1].

Minimal Performance Improvements from Data Driving Traditional ESG ETFs

Figure 1 looks at the performance of the index underlying the largest (by assets under management) Exchange Traded Fund in the “ESG Universe” as reported by ETF.com. Note that this ETF, based on traditional ESG data, has consistently underperformed its non-ESG parent index over the past five years.
**Outside-in: Sourcing Unstructured Data and Extracting Insights**

TruValue Labs leverages natural language processing (NLP) and artificial intelligence (AI) to implement a different approach to ESG data and other intangible risk factors.

Instead of providing an inside-out view of company reports based on levels of disclosure and disclosed data that is time-lagging, TruValue Labs delivers an outside-in perspective. The Insight360 platform includes information from a wide variety of sources, including reports by analysts, various media, advocacy groups, and government regulators. Insight360 aggregates unstructured data (text) from more than 75,000 sources into a continuous stream of relevant ESG data for monitored companies and sectors. Insight360 documentation describes the scoring methodology in more detail and provides more background on the machine learning approach [2].

**SASB Taxonomy for ESG and Intangible Risk**

The TruValue Labs SASB Lens reflects topics defined by the Sustainability Accounting Standards Board™ (SASB™). This includes 30 different topics covering Environment, Social Capital, Human Capital, Business Model & Innovation, and Leadership & Governance. Documentation of these categories plus their underlying rationales can be found at [3].

**Explanation of TVL Signals**

The cognitive computing system behind Insight360 uses natural language processing, or NLP, to interpret semantic content and generate analytics. It does so by applying criteria that are consistent with established sustainability and ESG frameworks, scoring data points on performance using a 0 to 100 scale. A score of 50 represents a neutral impact. Scores above 50 indicate positive performance, and scores below reflect negative performance.

*Figure 2* shows how TVL uses this stream of semantic content to generate a series of scores.
The **Pulse Score** is a measure of near-term performance changes that highlights opportunities and controversies, enabling real-time monitoring of companies. It focuses on events of the day and provides a responsive signal to alert investors to dynamic moves. Each company covered gets a pulse score summarizing the overall sentiment of documents on that topic for that company on any given day, in addition to sentiment of documents in categorical areas particular to established ESG frameworks.

The **Insight Score** is a measure of a company's longer-term ESG track record, similar to a ratings system. Scores are less sensitive to daily events and reflect the enduring performance record of a company over time. Scores are derived using an exponentially-weighted moving average of the Pulse, and the half-life of an event's influence on the Insight score is 6 months.

The **Momentum Score** measures the trend of a company's Insight score. It is a unique ESG metric in the industry that gives investors a high-definition view of the trajectory of a company's ESG performance. It does so by revealing a metric correlated with the slope of upward or downward movement, making it a measure that enhances quantitative workflows.

The **Volume Score** aggregates the running total of articles tagged to SASB or TVL topics over the past 12 months. Each day the volume score gets updated as new articles arrive.

This whitepaper focuses on the testing of the Insight and ESG Momentum Scores in stock selection and portfolio construction. Because the Pulse Score is much faster moving, it fits with a higher turnover trading strategy that will be examined in a future research note. Each of these metrics have different uses and are appropriate in different situations. For example, momentum is not sensitive to rapidly changing events, including those which may be critical to the future of a company; e.g. Equifax's hacking dramatically and suddenly changed how it was valued. This is captured accurately by pulse, while takes a longer time for momentum to reflect the change.
Performance Tests of Insight, ESG Momentum, and Volume Signals

Signal Testing Methodology

Current coverage for 2015, 2016 and 2017 is around 8,000 stocks, roughly aligned with the Russell 3000 in the US and the MSCI All Cap World Index (ACWI). TruValue Labs is in the process of building out deeper histories for these signals, prioritized by the most widely followed stocks. The first build-out of history covers 5+ years for the S&P 500: 2012 – 2017. Because the ESG Momentum score looks at change over the past 12 months, the test period for this study begins in 2013.

Universe Mapping to S&P 500 Benchmark

To eliminate survival bias in the test set, this study uses historical constituents for the S&P 500 as it has existed month-by-month for the past five years. Recent history of major indices such as the S&P 500 is available on public sources such as ETF holdings and various other Internet sites (for instance, https://en.wikipedia.org/wiki/List_of_S%26P_500_companies).

Investors who have not closely followed benchmark changes over time may be surprised to find out how much these benchmark constituents change from month to month. For example, the S&P 500 does not always have exactly 500 constituents, and there are on average two constituent changes per month over the past few years.

Portfolio Construction

Different fund managers have different holding periods, from day-traders to those who hold on to their investments for years. This study uses a typical test period for the fundamental equity quant, with monthly rebalancing. Portfolios are formed at the start of each month, exiting old positions and entering new ones at the opening price on the first trading day of each month. To avoid look-ahead bias, the portfolios are based on information from the end of the previous day. No slippage or commissions are included, nor are any attempts made to mitigate turnover. This initial study is meant to give a first cut at the alpha available in the various signals and is not intended as a precise trading strategy. There are various techniques to address liquidity and turnover issues, and future studies may investigate these topics.

The portfolios are formed with equal-weighted constituents each month. There is no attempt to optimize the portfolio weights based on expected alpha, expected correlations, or expected exposure to various risk factors. Again, the focus is on an estimation of the raw alpha in the signals before any attempts at more advanced portfolio construction techniques.

Large cap benchmarks such as the S&P500 have both cap-weighted and equal-weighted versions available both as indices and as ETFs. Figure 3 compares the equal-weighted full test universe to equal-weighted version of the S&P500 benchmark. Note that performance of the test set is closely in line with the benchmark, suggesting that there is little bias introduced by the small differences in coverage of TVL test data each month vs. the full
Performance Tests of Insight, ESG Momentum, and Volume Signals

S&P 500 constituent set. Because subsequent tests are formed using equal-weighted portfolios, the RSP (equal-weighted S&P 500 ETF) has been chosen as the most appropriate benchmark.

Coverage is complete at the present day, nearly complete at the start of 2015, and falls off to around 400 stocks as of the start of the test period. Figure 3 shows that the coverage gap in the early part of the test period does not lead to any meaningful biases in terms of returns for the covered part of the test set vs. the entire S&P 500.

**Backtesting Software**

As quantitative strategies grow in popularity, backtesting software solutions are proliferating. For applications using traditional financial data and overnight or longer holding periods, python is currently the main tool of choice for data scientists and quants. This study was performed using an open-source python package called Zipline, produced by the FinTech firm, Quantopian. Leveraging a widely-used open-source package helps to standardize measurements and reduce the chance of common backtesting problems having to do with look-ahead bias in its various forms. Figure 3 above and many of the figures that follow were produced using Zipline and its included analysis packages. Note that using the equal-weight RSP as benchmark leads to an alpha of 0.00 and beta of 1.00 for the monthly rebalanced test set.

**A note on in-sample vs. out-of-sample results**

Most quant studies carefully delineate separate in-sample and out-of-sample test periods. Many backtested quant algorithms are developed by examining their backtest results. This study, however, differs from the typical quant algorithm backtest. None of the TVL signals were developed with any view to their performance based on stock returns. The Pulse signals were trained using natural language processing (NLP) and sentiment analysis with no input whatsoever involving stock price returns. Similarly, the Insight score uses a basic Exponentially Weighted Moving Average with a six-month half-life that was not “optimized” in any way involving quant backtesting of price returns; rather, it was built simply to smooth out the day-to-day volatility in the Pulse score. Similarly, the Momentum score measures change in the Insight score again with no optimization based on stock price returns. Thus, the entire suite of TVL scores were developed with no use of stock price return
data. That means that studies of signal performance using stock price returns are effectively all “out-of-sample” with respect to the construction of the signals themselves.

That said, customers who choose to customize or recombine the various TVL signals based on this study or their own work would be well-advised to establish clear in-sample and out-of-sample sets for their own work to ensure no biases or over-optimization result from constructing those custom signals.
Portfolio Strategy Development

Article volume is an important contributor to the meaningfulness and accuracy of the TVL Insight score. If only one or two articles have been seen for a given company in the past 12 months, it is hard to draw reliable conclusions from those articles. Hence, TruValue Labs Insight360 product assigns categories of volume for each company based on the past 12 months of article flow. Low coverage is 1-4 articles (one per year to one per quarter). Medium coverage is 5-25 articles (above one per quarter but less than one per two weeks). High coverage is 26+ articles in the past 12 months (at least one article every two weeks on average). These categorizations were developed by looking at the distribution of articles over the full TruValue Labs dataset.

Insight + Volume Portfolio

The most basic test of the TVL Insight score is to use it as a negative screen to develop a broad portfolio of companies that excludes the lowest scores. Applying a volume filter to ensure good information flow, the Insight + Volume Portfolio includes the top 40% of companies by article volume in the S&P 500, then excludes the bottom 1/6 of those with the lowest Insight scores. $\frac{4}{10} \times \frac{5}{6} = \frac{1}{3}$ of companies pass this dual filter. For the S&P 500, this would be 167 companies. Toward the early end of the test years, as coverage of the S&P 500 drops off, the strategy continues to pick the top 40% by volume for the most comparable test of data available at the time.

This approach emulates a traditional ESG-focused ETF using a breadth typical for many Smart Beta ETFs. The portfolio should exclude the great majority of companies with ESG controversies assuming that any controversies are known and important enough to generate sufficient information flow.

Momentum + Insight + Volume Portfolio

The second test portfolio takes more of a quant research approach, generating a narrow portfolio with the highest probability of positive returns. The Momentum + Insight + Volume Portfolio begins with the same dual screen as the Insight + Volume Portfolio, then narrows the selections to the top 30% of companies by ESG Momentum. $\frac{3}{10} \times \frac{1}{3} = \frac{1}{10}$ or the top decile according to this triple filter combination. This portfolio targets higher returns while allowing for higher turnover as well.
Performance Tests of Insight, ESG Momentum, and Volume Signals

Performance Summary

Long-Only Results

Figure 4 shows the performance of these two portfolios vs. the S&P 500 benchmark. The Insight + Volume Portfolio outperforms by 3.5% annualized over this 4.5-year period, while the Momentum + Insight + Volume Portfolio outperforms by 5% annualized. Figure 5 shows the ramp-up of excess returns over the test period. Table 1 shows year-by-year performance for each strategy as well as excess returns over the benchmark.
Performance Tests of Insight, ESG Momentum, and Volume Signals

Table 1: Insight and Momentum Strategy Returns vs. S&P 500 by Year

<table>
<thead>
<tr>
<th>Year</th>
<th>Benchmark</th>
<th>Insight</th>
<th>Momentum</th>
<th>Insight</th>
<th>Momentum</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>27.2%</td>
<td>33.4%</td>
<td>32.6%</td>
<td>6.2%</td>
<td>5.4%</td>
</tr>
<tr>
<td>2014</td>
<td>14.0%</td>
<td>16.7%</td>
<td>20.1%</td>
<td>2.8%</td>
<td>6.1%</td>
</tr>
<tr>
<td>2015</td>
<td>-2.7%</td>
<td>3.2%</td>
<td>3.5%</td>
<td>5.9%</td>
<td>6.2%</td>
</tr>
<tr>
<td>2016</td>
<td>14.4%</td>
<td>12.3%</td>
<td>14.7%</td>
<td>-2.1%</td>
<td>0.3%</td>
</tr>
<tr>
<td>2017 YTD</td>
<td>9.4%</td>
<td>12.1%</td>
<td>13.6%</td>
<td>2.7%</td>
<td>4.2%</td>
</tr>
<tr>
<td>Since Inception</td>
<td>77.1%</td>
<td>103.0%</td>
<td>115.2%</td>
<td>25.9%</td>
<td>38.1%</td>
</tr>
<tr>
<td>Annualized</td>
<td>13.5%</td>
<td>17.0%</td>
<td>18.6%</td>
<td>3.5%</td>
<td>5.0%</td>
</tr>
<tr>
<td>Sharpe Ratio</td>
<td>1.00</td>
<td>1.29</td>
<td>1.36</td>
<td>0.29</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Figure 6 looks at a 6-month rolling average of the Sharpe Ratio for the Momentum + Insight + Volume Portfolio. The Sharpe Ratio is defined as (annualized return minus risk-free rate) / annualized volatility. This rolling average Sharpe Ratio is positive throughout the test period except for a 6-month drawdown in late 2015, and it appears to be increasing slightly over time as data information coverage increases subsequently. Putting these numbers in context, a Sharpe ratio of 1.3 represents a 30% improvement in the returns-to-risk ratio vs. the benchmark portfolio which has a Sharpe ratio of 1.0.

Figure 6: Rolling Sharpe Ratio
Performance Tests of Insight, ESG Momentum, and Volume Signals

Sensitivity to Rebalancing frequency

The Insight Strategy shows about 5% turnover per monthly rebalance. The Momentum overlay increases returns by restricting the portfolio to companies that have strong momentum, but at the cost of increasing turnover to 20% per monthly rebalance. But how sensitive are these portfolios to rebalance frequency?

Figure 7 looks at the annualized returns for each strategy for monthly, quarterly, semi-annual, and annual rebalance frequencies. The Momentum strategy is more sensitive, as expected, but surprisingly does slightly better with quarterly rebalancing than with monthly rebalancing, at least over this particular test period. With holding periods of a year, the Momentum strategy loses its incremental alpha generation over the baseline Insight strategy portfolio. The Insight strategy is by design both broader and lower turnover, and it performs best at a 6-month hold, maintaining strong returns out to a full year between rebalances.

![Figure 7: Impact of Rebalancing Frequency on Annualized Returns](image)

This analysis shows that the Insight strategy is applicable for investors who rebalance only quarterly or even annually, while the Momentum strategy is intended for those who rebalance their portfolios more frequently. Note that beyond this study, portfolio turnover could be reduced even further by allowing scores to drop beyond the entry point by some amount before replacing them in the portfolio, thereby lowering portfolio churn.
Performance Tests of Insight, ESG Momentum, and Volume Signals

Performance of SASB Material Categories vs. SASB All Categories

The SASB framework defines 30 categories but also specify which categories are the most material for each of the SASB 79 industries. Typically, 3 to 11 categories are classified as material, with the average number of material categories being 5. SASB provides detailed documentation of their categories including which categories are material for each industry [4]. TruValue Labs runs two aggregate scores for the SASB Lens: SASB All Categories and SASB Material Categories.

Returning to the base case of monthly rebalancing, using the ESG Momentum + Insight + Volume strategy, SASB All Categories generates 5.02% annualized alpha. SASB Material Categories generates 4.46% of that alpha, which is 89% of the alpha. Yet only 5 of the 30 categories are material on average, which is 17%. This result suggests that SASB has done a good job of identifying materials categories, since 17% of the categories create 89% of the value added (a bit beyond the classic 80/20 rule of thumb that 80% of results typically come from 20% of causes or activities). The residual alpha could be explained by categories that are material at the company level more than the industry level. Figure 8 graphically shows the comparison of alpha creation.

![Figure 8: 89% of Alpha comes from just the 17% of categories that SASB identifies as material.](image)

Performance of Individual Categories

Any of the underlying 30 components from the SASB Lens can be used to create portfolios for backtesting. Not surprisingly, some signals work better than others. Data volume also varies significantly from one category to the next, but the backtests all start with taking the top 40% of companies by data volume as the first screen to test each category in a consistent way. Because of the smaller sample size for these components, it would be wise to expect some reversion to the mean going forward, but it is interesting to see which signals did best and worst over the test period. Figure 9 shows the results for the S&P 500 over this test period.
Performance Tests of Insight, ESG Momentum, and Volume Signals

Looking at the list in detail, it is interesting to see customer welfare sort to the top of the performance chart, as it makes sense that shareholders will respond strongly to how businesses treat their customers. Data Security and Privacy shows up as a key factor on both lists. At the other end of the spectrum, Labor Relations and Fair Labor Practice do not appear to be getting the same level of scrutiny from shareholders, at least on this particular sample period. Some measures such as Air Quality may be showing less performance simply because they get tagged in fewer documents and thus have a lower level of overall information flow, as these charts were constructed using the top 40% on category-specific volume as the first filter in applying the Volume + Insight + Momentum strategy. Future work will extend this analysis to larger data sets.

These categories can be combined in many different ways. TruValue Labs current provides two high-level summary measures: SASB All Categories and SASB Material Categories. Intermediate categorizations could be constructed as well, such as scores for Environmental, Social, and Governance as three separate intermediate categories. Furthermore, TruValue Labs can develop new customized lenses in addition to the SASB Lens to measure other kinds of intangible risk factors.

Comparison to Common Quant Factors

Certain quantitative factors have attracted investor interest for decades. Value, for instance, have been used heavily by Warren Buffett and others dating back to Graham and Dodd in the 1930s [5]. Factor investing started attracting attention in academia following Fama and French's seminal work on size and value in 1992 [6], followed quickly by Carhart's addition of price momentum in 1994 [7]. Along with market beta, size, value, and price momentum are among the most commonly considered fundamental quant equity factors. The recent book *Efficiently Inefficient* provides a good introduction to these most common quant equity strategies. [8].
The most common starting point for testing the uniqueness of new factors is to first check their correlations to the best-known academic factors: the Fama-French-Carhart factors. The Quantopian backtesting package provides a module for performing this test.

*Figure 10* shows the betas (factor loadings) over time between the SASB Materiality factor and Size (blue), Value/Growth (green), and Price Momentum (red). The strongest beta over the most recent period is with size, due in part to testing the factor by forming equal-weighted portfolios which overweight companies with smaller market-cap relative to a cap-weighted benchmark. However, there was a period of a year where the beta with size was actually negative. Beta with value/growth is low or negative throughout the test period. There was a strong positive beta with price momentum at the start of the test period, but then that went to zero and then negative in following years. The beta with Price Momentum is possibly due to periods where both factors were working well initially, then Price Momentum did poorly later on. The key conclusion from this chart is that the betas with all of these factors are moving around over time, crossing from positive to negative, suggesting that there is no stable relationship between these factors and the new factors derived from the TVL volume, insight, and momentum signals.

Large cap stocks, the S&P 500 in particular, are difficult grounds for finding alpha, particularly since the Global Financial Crisis of 2008. These are some of the most liquid, most heavily traded, and most widely followed stocks in the world, and one should not expect many market inefficiencies to persist for these kinds of companies. Yet some quant factors still do work on the S&P 500, though the resulting alpha is smaller and less reliable than in more inefficient markets elsewhere.

Smart Beta is growing in popularity, with many ETFs proliferating. Size, value, price momentum, minimum volatility, and quality are some of the most popular factors. Size is less relevant for an S&P universe as all of these companies are large cap. *Figure 11* looks at top decile performance for basic versions of the other common smart beta factors on the same S&P 500 test universe.
Figure 11 shows how Quality has been the best performing traditional quant factor on the S&P 500 over this test period, followed by value. Price Momentum and Minimum volatility factors have moderately underperformed the benchmark. Interestingly, both outperforming factors (Quality and Value) incorporate fundamental information, whereas the two underperforming factors (Price Momentum and Minimum Volatility) only incorporate price information. A combination of all four factors tends to perform most reliably, and that is indeed the case here as well, with the four-factor combination performing about as well as the best individual factor (see the light blue bars). In comparison, the ESG Momentum and Insight signals (the green bars) outperform all of these traditional quant signals, including the 4-factor combination. The new signals are also additive when blended with the 4-factor model, as the dark blue bars show.

![Graph showing factor performance](image)

**Figure 11: Standalone and Combination Factor Performance**

### Long-Short Strategy

Many fundamental quant equity traders look mainly at long-short performance. For these new data sets, the best short strategy is not necessarily to do the opposite of the long strategy. The reason has to do with the interaction of volume with the other scores.

It turns out that there is a significant bias to returns based on volume of stories logged by the TruValue Labs system. At least in the S&P 500, companies with high SASB or TVL category volume tend to strongly outperform companies with low volume, regardless of the pulse, insight, and momentum scores. Recall that for low volume companies, the other scores are less meaningful. A company with low volume may have good, bad, or neutral characteristics in reality, as it is hard to make a judgment from a low volume score.

Volume is related to market cap, as higher market cap companies are more likely to have coverage from news, social media, and other sources. But Figure 12 shows how the effect of the TruValue Labs volume measure is
Performance Tests of Insight, ESG Momentum, and Volume Signals

Figure 12: Performance by buckets of volume and market cap

about 4.7x as strong as the pure market cap effect, comparing the spread between high volume and low volume companies.

This observation about volume suggests a long-short trading strategy to buy the high-volume companies and short the low-volume companies, rebalancing monthly. Figure 13 shows that this strategy returns about 6% pure market-neutral alpha over this test period at 2x leverage (100% long + 100% short, market neutral).

Figure 13: Long-Short strategy based on TVL volume metric
Using the pulse score is likely another interesting way to build a long-short strategy, but likely requires higher turnover. When examined on a monthly rebalance basis, companies with strong negative scores frequently show a strong recovery bounce after the initial crash. Future research will use event studies to investigate the best way to trade around positive and negative events.
Performance Tests of Insight, ESG Momentum, and Volume Signals

Plans for Future Work

As five years of history become available for a broader coverage of global equities, more extensive backtesting studies will be possible. For example, the S&P 500 does not really provide the breadth of companies do perform detailed analyses at the sector level. Future research notes may cover the following topics:

- U.S. performance on the Russell 1000, 2000, and 3000
- Ex-U.S. performance including MSCI World and MSCI ACWI
- Performance results by sector

Different time horizons are of interest to different investors. Shorter-term traders would likely be interested in event studies showing day-by-day price response to the most significant positive and negative sentiment ESG events. At the other end of the duration spectrum, long-term investors are interested in which categories are most material to long-term fundamental measures such as ROE, ROA, and cost of capital. A recent paper summarizes more than 2000 empirical studies and 90% of them find that ESG information is additive to corporate financial performance in a non-negative way [9]. Future research notes and papers may pursue these financial performance topics in more detail using this newly available data.
References


[2] Reference on TVL methodology

[3] SASB Score overview

[4] SASB Score reference explaining material factors


