KASPR Datahaus

Research Report | Backtesting

Predicting Major Indices

Executive Summary

- In this report we show how KASPR Datahaus PTY LTD (KDH) **alternative data products** can be used to build leading market intelligence for major market indices.
- With a very simple trading rule, and a one-day look-ahead model, we show that KDH alternative data generates $\sim$ **7%**+ **excess return margin**, relative to a baseline model excluding our data product.

The KDH Alternative Data edge

Some time ago, through our academic work, we saw the potential for leveraging granular internet activity and quality patterns as aggregators of human and technical information, at global scale (Fig **A**).

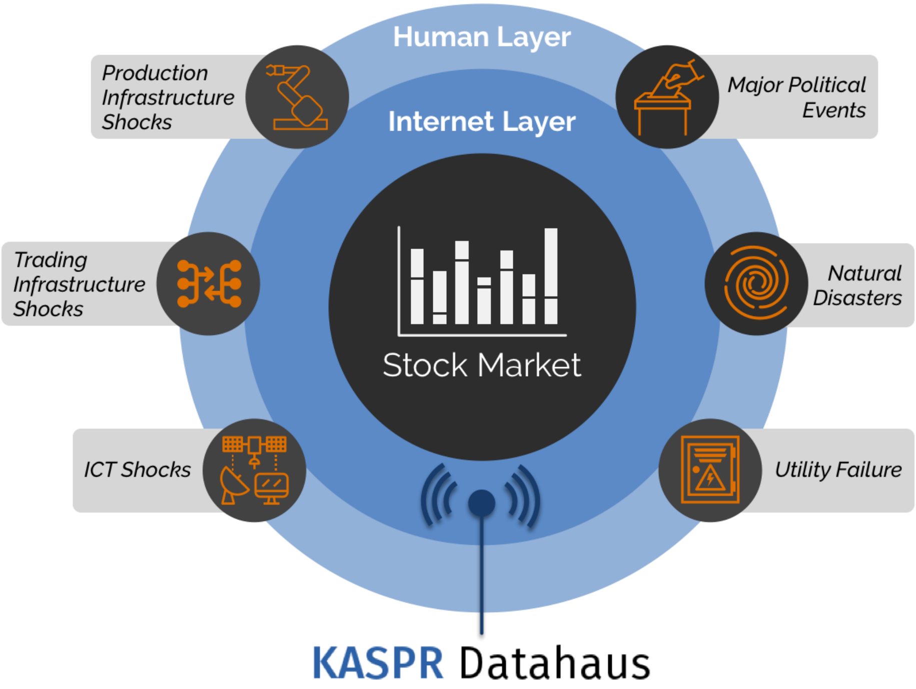

Figure **A**: Billions of events every day can directly and indirectly inform internet signals observed by our technology.

Major events across the world have direct (through online connectivity) or indirect (through changed human–internet patterns) impacts on internet activity and quality, signals that our global technology network picks up in real-time.

Here, we focus our attention on the **KDH Global ICT Intel Data** product which provides several measures of internet activity and quality, at national level. This product is delivered within the first hours of the day concluding, equipping our clients with a unique informational advantage.

The case of Reliance Communications Ltd (INDIA)

As an example, we demonstrate relationships between major ICT activity shocks as measured by KDH Alternative Data technology and stock-market movements.

Specifically, in Fig. **B** we trace the stock returns (closing price in INR) of three large Indian companies over Jan–Apr 2019. Grey vertical areas indicate days in the period of substantially low (< 25th percentile) ICT Activity measurements from our sensing infrastructure, with red line plots the respective stock's closing price.

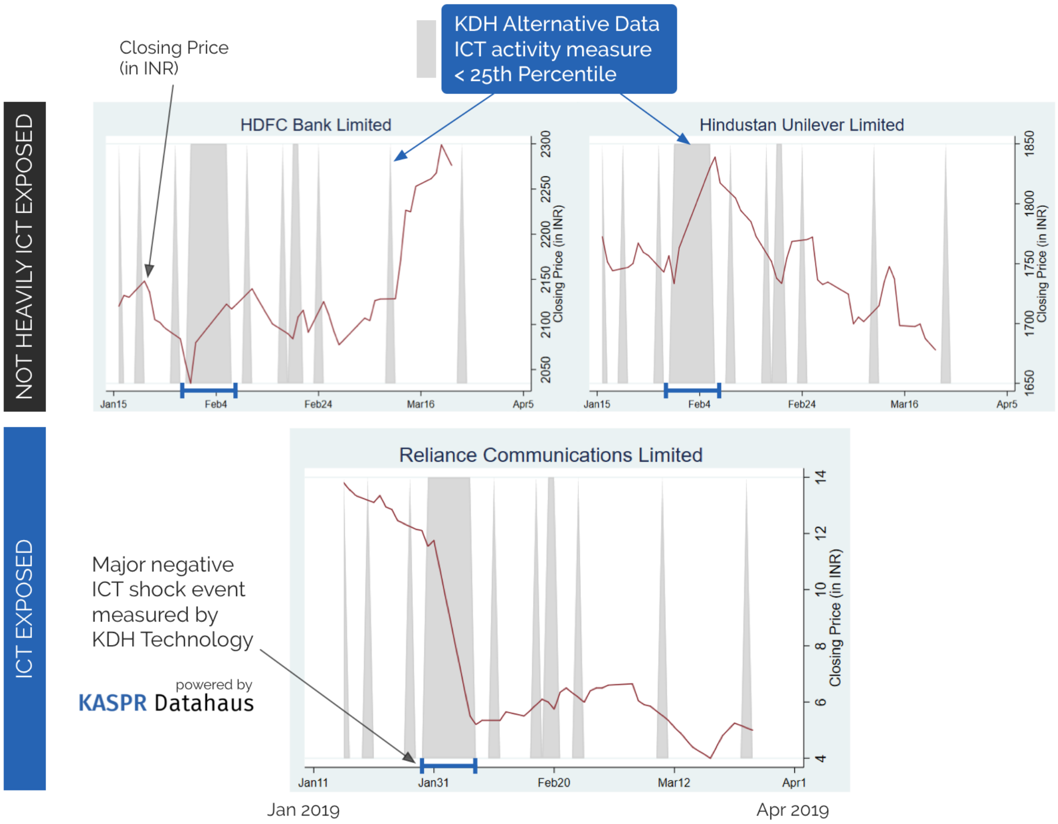

Figure **B**: Comparison of ICT exposed and not exposed company stock returns around the 30 Jan 2019 major ICT activity event observed by KDH P/L technology.

We contrast the experience of two non-ICT exposed companies in the top panel – HDFC Bank Ltd (HDFCBANK), a major banking and financial services company, and Hindustan Unilever Ltd (HINDUNILVR), the well-known and very large manufacturing company – with the major troubled mobile network provider, Reliance Communications Ltd (RCOM.NS) in the lower panel.

Reliance Communications made market announcements on Friday, 1 Feb 2019, indicating that it would seek resolution of its debt position through bankruptcy. By Monday, the stocks lost almost half their value and plunged further in the ensuing days. This tumultuous period coincided with KDH Alternative data measurements of anomalously low internet activity across India just prior to, and during this period.

With HDFC Bank Ltd and Hindustan Unilever Ltd serving as non-ICT exposed controls, the exercise points to the potential of consistent, comprehensive, and remotely measured internet activity and quality as a leading indicator of market outcomes.

Backtesting: details

Prediction Strategy | Overview

In this report we take both a conservative and simplified approach to the modelling and prediction problem in order to demonstrate most clearly the additional predictive signal arising from our data products.

We demonstrate leveraging our data for prediction of the S&P 500 Volatility index (`VIXD`) and UK FTSE All-Share Return Index (`TFTASD`). Training and testing is conducted using a moving window approach such that we train on features from the day before a given day, and use the given day as the true outcome (see Fig.C). For stability, we ensure that the window is at least 100 days in length.

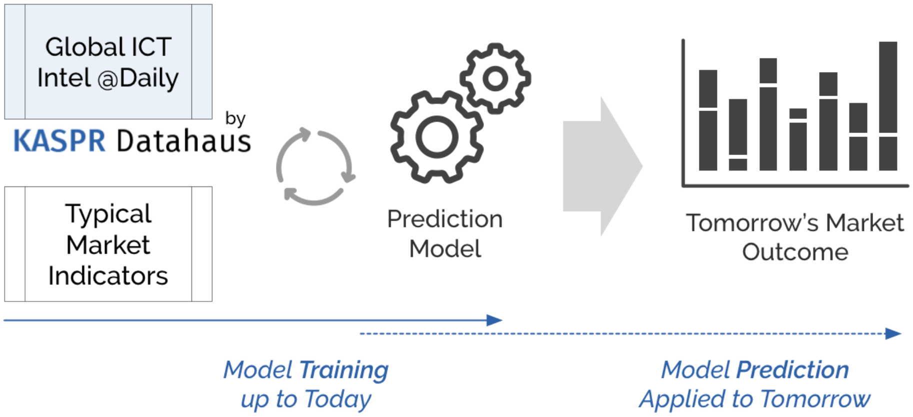

Figure **C**: A moving window train–test prediction analysis demonstrates a strong informational advantage of KDH P/L Alternative Data Products.

For prediction, we use a regression tree ensemble (random forest), an off-the-shelf machine-learning tool[1], with 200 individual trees creating the ensemble.

To measure the informational advantage that our **Global ICT Intel Data** product yields, we apply the train–test windowing approach to all days in May 2019 either with both traditional market indicator data *and* our data product, or with traditional market indicator data only as a benchmark.

After dropping non-trading (weekend) days, we have 21 (FTSE) or 22 days (VIX) of testing days in our sample.

[1] See accompanying codes.

Features from KDH Alternative Data products

The **KDH Global ICT Intel Data** product provides three novel signals of ICT activity and quality: **count_unique_ips**, **rtt_mean_norm**, and **rtt_var_norm**.[2] Together, these measures arise from direct and indirect impacts of major global events on internet signals in near real-time (see Fig.**A**).

For this exercise we generate a series of standard features from our data product measures including lags (1 to 7 days), 1st differences, and for the **rtt** measures a version of the RSI index, with basis at lags 1 to 7.[3]

Features from Typical Market Indicators

Standard features are created in a similar way from the day closing ticker data in question, with lags, first differences and the RSI like index, over 1st to 3rd day lags.

Outcome variable

For either market indicator, the outcome variable for training and testing is the closing indicator value, 1-day in advance (`ln_close_lead_1day`). In other words, the trained random forest prediction model will generate an expected value of the natural log of tomorrow's closing value of the given index.

To translate these predictions into a salient measure of predictive accuracy, we apply the most simplistic of trading rules: go long when the prediction is above today's close, and go short otherwise. It is then trivial to calculate the accuracy of this strategy based on the model prediction.

To provide statistical support to any differences yielded between outcomes derived from traditional versus traditional plus KDH alternative data products, we run 100 independent trials of the prediction exercise in each case to generate a distribution of accuracy scores for comparison, displayed graphically.

It is important to note that the predictions given in this report arise from a basic trading rule, with an off-the-shelf statistical learning prediction tool with no parameter tuning. Analysts would no doubt be able to strengthen these predictions further, extracting the most value from our alternative data products.

Note on Bias

Financial backtesting can be subject to various sources of well-known bias. In this exercise, we seek to avoid the most significant sources of such bias. By working with large composite indices as the outcome variable we side step *survivorship* and *selection* bias that can affect testing at the company, or bag of company, level due to unrepresentative sub-sampling or unintentionally omitting failed companies. By using train–test windowing with strict temporal demarcation between features and outcome, we avoid *look-ahead* bias, as no information from the future is able to leak back to the features used for prediction.

Finally, by comparing outcomes using statistical quantities of modelling exercises, we avoid *small-n* bias that can arise when providing only a single example of a processes which has a randomly varying component. Since the prediction framework employed in this report is statistical machine learning which by design depends on the random number stream of the implementation software, this is an especially important feature of our outcome comparison.

[2] Please see KASPR Datahaus PTY LTD Data Descriptor, 'Global ICT Intel Products', available from https://kasprdata.com for details.

[3] See 'transform_inputs.sql' for details.

Backtesting Results

Predicting the S&P 500 Volatility index VIX

For the VIX we have a total of 22 trading days in May for testing, using the rolling window train–test methodology, summarised above.

Per the method outlined above, we first train–test with only the traditional indicator data, running 100 independent repeats of the exercise to generate a distribution of outcome accuracy. Then, we follow the identical approach but add the KDH Global ICT Intel data features to the exercise, allowing us to compare accuracy differences against the traditional benchmark.

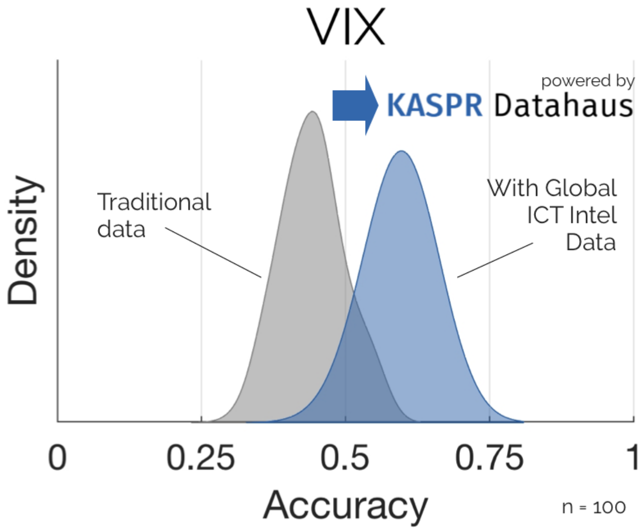

Figure **D**: Accuracy shift with KDH Global ICT Intel data relative to the traditional benchmark features only when applied to day-ahead prediction of the VIX.[5]

Results are summarised in Fig. **D** above, showing that when KDH Global ICT Intel data are added to the feature set (blue) a substantial lift in day-ahead prediction accuracy is achieved over the traditional benchmark (grey). Average accuracy for the traditional features only is 44.3%, whilst with KDH Global ICT Intel data added, average accuracy rises to 59.5%.

Predicting the UK FTSE All-Share Return Index

Next, we apply an identical approach to the FTSE All-Share Return Index, with 21 available days of train–test data available using the rolling window design outlined above.

For the FTSE, we find that average benchmark accuracy of 52.4% is obtained without the KDH Data product measures, whilst average accuracy performance lifts strongly as before with KDH Global ICT Intel data added, to 60.7% (see Fig. **E**).

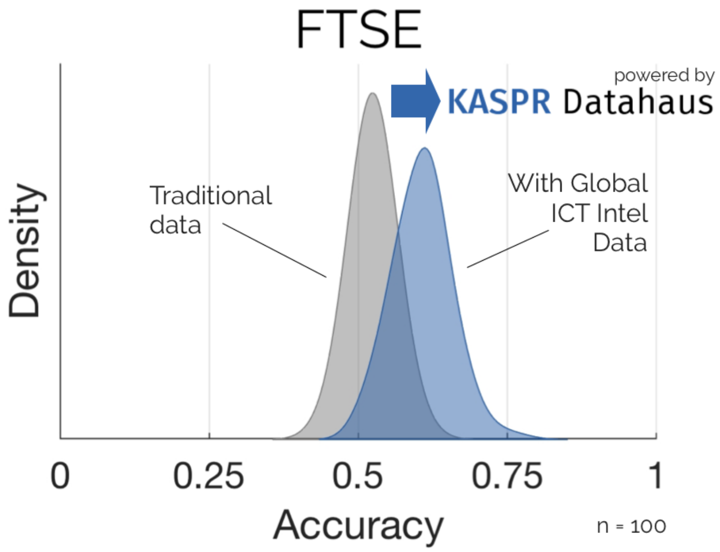

Figure **E**: Accuracy shift with KDH Global ICT Intel data relative to the traditional benchmark features only when applied to day-ahead prediction of the FTSE.

Long-run Historical Performance of ICT Data

In support of the backtesting activities reported here, we have also applied the identical methodology to a historical dataset of ICT activity and quality obtained by our team for academic purposes.[6] The historical scanning technology differed in substantial ways to KASPR Datahaus' proprietary technology and consequently, the historical dataset does not provide anything like the equivalent granularity and depth of geo-spatial coverage. Nevertheless, the historical data's coverage from late 2005 to mid 2013 provides a longer time-series for indicative internet activity signal.

Applying the same 'with' and 'without' comparison as above, in a train–test windowing framework, with at least 100 days of on-ramp, we obtain a steady accuracy differential of at least 7% using the historical data, under the simple short/long trading rule applied above, and with several random seeds (relevant for statistical machine learning).

Conclusion

Taken together, the exercises reported here demonstrate the potency of incorporating KDH Global ICT Intel Data products in day-ahead financial prediction modelling.

[6] Since the purpose was academic, KDH is not in a position to share these data more widely.

Technical Appendix

All codes to regenerate these results available upon request. In addition, main backtesting results reported here are based on the standard KDH Global ICT Intel Data sample.

Files included in the technical package include:

1. **transform_inputs.sql**: SQL code to prepare, transform and generate features for model training and testing;
2. ***.csv**: four `csv` files included in the package include:
   (a) `index_daily.csv` daily indices used in the report (2005-01 to 2019-05),
   (b) `kaspr_daily.csv` daily KDH Global ICT Intel Data sample dataset (2018-10 to 2019-05),
   (c) `comb_kaspr_index.csv` combined and matched dataset for training and testing respectively,
   (d) `accuracy_ntrials.csv` accuracy outcomes for 100 independent trials of the prediction exercise for two indices with and without KDH alternative data features ('IP' in the headings).
3. **test_predictions.m**: main MATLAB model code to prepare windowed test and train data sets, run the prediction exercise, and generate quantitative results;
4. **run_report.m**: MATLAB run file to produce a single prediction series with one random seed (3805), including switches for running model training and predictions with, or without, KDH ICT Daily data products;
5. **run_distribution.m**: MATLAB function to run N inndependent replicates of the prediction exercise over both indices with, and without, KDH alternative data;
6. **plot_distributions**: MATLAB plotting function to process the output file produced by **run_distribution**, `accuracy_ntrials.csv` and create the plots in this report.