Alternative data for investors

Saeed Amen, Quantitative Strategist

Founder of Cuemacro

www.cuemacro.com

saeed@cuemacro.com

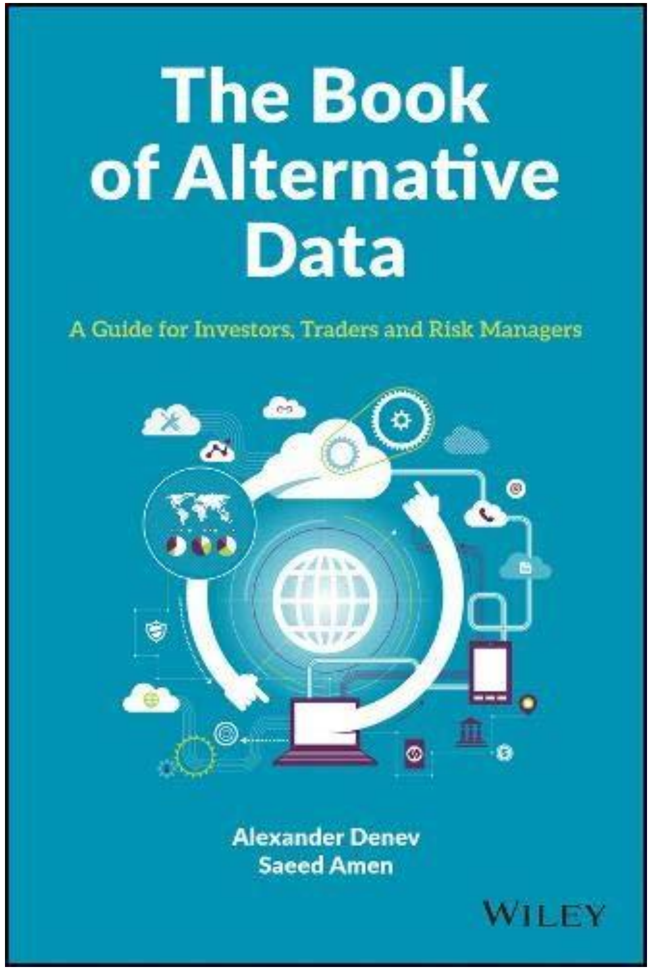
@saeedamenfx

Founder of Cuemacro – Saeed Amen

- Over decade in currency markets starting at Lehman Brothers and latter at Nomura as an Executive Director developing systematic trading strategies
- One of team who created Lehman Brother's MarQCuS FX factor model, which had **2bn USD AUM**
- Created finmarketpy, findatapy and chartpy open source **Python financial analysis** libraries (grew out of pythalesians library) – finmarketpy is number 2 Python trading library on GitHub
- Co-founded **the Thalesians** a quant think tank, with finance events in London, New York & Budapest
- Now established **Cuemacro**, focused on quant consulting in **macro markets** and creating innovative datasets to model macro economic sentiment
- Projects for companies including **Investopedia** (financial news website) and **Freepoint** (commodities trading) other clients include several **large UK quant funds**.
- Presented my research at **IMF, ECB, Federal Reserve Board** and **Bank of England** and major quant conferences
- Author of **Trading Thalesians**: What the ancient world can teach us around about trading today (on Palgrave Macmillan)
- Co-Author (& Alexander Denev) of **The Book of Alternative Data** (on Wiley in early 2020)

The Book of Alternative Data

- Co-authored by Alexander Denev and Saeed Amen
- **The Book of Alternative Data** (on Wiley in early 2020)
- On Amazon already for pre-order
- Presentation is based on the book!



Alternative data primer

What is alternative data?

- Common properties
  - Less commonly used by market participants
  - Tends to be more expensive
  - Often outside financial markets (is tick data “alternative”?)
  - Shorter history
  - More challenging to use
- “Exhaust data” a byproduct of other processes
  - Digital footprint from individual and corporate activity
  - Resulted in a rapid rise in the number of alternative datasets
  - Can provide an additional revenue stream for those who collect “exhaust data”

Types of alternative data

- Satellite/aerial photography
- Location data
  - mobile phones
  - apps
- Text
  - Web
  - Social media
  - News
  - Internal data
- Consumer transactions
  - Credit card transactions
  - E-mail receipts
- Corporate
  - Supply chain
  - Internal metrics
- Market
  - High frequency tick
  - Flow data
- Crowdsourced data
  - Alpha capture
  - Analyst estimates
- And much more!
- Have some case studies later and in our book!

The V's of Big Data

- Volume (increasing) – lots of data
- Variety (increasing) – not just numerical data, can be text, image, video etc.
- Velocity (increasing) – speed that data is being generated
- Variability (increasing) – inconsistencies in the data
- Veracity (decreasing) – difficult to tell if accurate (e.g. social media)
- Value (increasing) – business value of the data

Value of alternative data

- Decay of investment value
  - Signal from less common data may decay less quickly
- Monetary value of data
  - Market value
  - Economic value
- Predictive value of data
  - Does it add value for investors? Also depends on the type of investor
  - Unusual data is not necessarily always of value for investors

Legal questions

- Before buying data, we need to be aware of the legal aspects
  - Can the data be sold? (e.g. GDPR issues and consent)
  - Have the personal details been properly scrubbed?
  - Does the data need to be aggregated before being sold to “blur” it?
  - Are there issues for “exclusive” datasets?
  - Very important for sellers to be aware of the legal aspects (as well as buyers), must investigate beforehand
  - Issues will vary between datasets

Data challenges

- Entity matching
  - Matching to traded assets (e.g. iPhone to Apple)
- Missing data
  - Data can be sparse, how can we fill (averages?)
- Structuring the data
  - Converting unstructured data, often images and text into a more structured form, often ultimately into a time series of numerical data
- Anomalies
  - Data which deviates substantially from what is expected, e.g. outliers in tick data

Finding the right dataset

- Identify the right dataset
  - Hypothesis approach: often need to consider what the question and hypothesis
  - Data driven approach: start with data and then identify the “rational” for the market tends to be more challenging and easier to have data mining issues
- Do the analysis to verify the hypothesis
  - Plotting early on in the process
  - Potentially trying regressions and correlations, with appropriate market data or economic forecasts
  - Create a market model
- Clearly, not every alternative dataset will be useful for your purposes

Searching for alternative data

How to find alternative data?

- Web directories
  - Can find datasets listed on web (free!)
  - Approach data firms directly
  - Eg. www.alternativedata.org
- Data firms which aggregate alternative data include
  - Typically take revenue share from underlying supplier
  - Make it simpler to interact with many data firms (one billing etc.)
  - E.g. Open:FactSet, Bloomberg, Quandl, Eagle Alpha etc.
- Directly to raw data source
  - Corporate firms – but can be challenging
  - Or can collect yourself – time consuming

Data strategists/scouts

- Within funds, there are data strategists, who
  - search for datasets
  - act as bridge between external data firms, and internal portfolio managers and data scientists
- External data scouts
  - Also see external firms in this space to help internal data strategists/scouts
  - Act as intermediary between data firms and data users
  - Paid by data user (ie. buy side), not by data firms
  - E.g. Neudata

… and don't forget about your data!

Cuemacro

- Every organization has internal data, financial organisations are no different, in particular sell side
- The difficulty is that it isn't often well catalogued
  - Does every team know about every dataset? Unlikely!
- Create a web directory of datasets as a start, to allow browsing
  - Some datasets cannot be made available to everyone (e.g. compliance reasons, licensing costs etc.)
- Benefits of centralization of data sourcing
  - Can negotiate better deals with data vendors vs. team by team
  - Can keep track of data subscriptions better, reducing unnecessary duplication

Costs of data

- Depends on several factors
  - Asset coverage
  - Frequency
  - Uniqueness
  - Trials are mixture of paid/free
- Can reduce cost by
  - Accessing dataset by company (ie. only those companies you are interested in)
  - Getting lagged data (which is fine for long term investing)
- Most datasets are under 100k USD annually (some can be a lot more, but a rarer)

Data delivery

- Delivery via
  - Flat files (CSV/XML) – for example downloadable from Amazon S3 buckets
  - API (historical and realtime feeds)
  - Web GUI

Structuring data: focus on NLP

Text is everywhere!

- A large amount of data available is in text form
  - Web
  - Social media
  - Newswire
  - Internal only data (e.g. e-mails, memos etc.)
- Need access to the text to start
- How can investors make sense of this text?
- Slides taken from The Book of Alternative Data (Alexander Denev/Saeed Amen), which is due out in 2020 on Wiley, more info at https://www.cuemacro.com/altdata/

Natural language processing (NLP)

- Convert various texts (unstructured) into an easier to use format (structured)
- In a structured form, we can more easily use it within the investment process
- Natural language processing encompasses many of the tasks we can use to do this
- We'll introduce the topic here

Various stages of NLP

Higher level	Pragmatics
	Semantics
	Syntax
	Morphology
	Phonology
Lower level	Phonetics

Phonetics, phonology & morphology

- Phonetics
  - Specific sounds generated by humans
- Phonology
  - Sounds of a specific language
- Morphology
  - How words are constructed and their decomposition
    - e.g. burgers can be broken down into burger (root) and 's' is the suffix
    - e.g. different verbal forms of eat (verb), eating (adjective) and eating (noun)
  - Can be very important for certain languages
    - e.g. Arabic, where verbs usually consist of three root letters

Syntax, semantics & pragmatics

- Syntax
  - How words are combined to make a sentence
    - Grammar dictates how words can be combined together
    - E.g. Word order “Alex consumes burgers” and “Burgers consume Alex” are both grammatically correct but have different meanings
- Semantics
  - Involves meaning in a language
    - Asking questions such as who, what, why, where, when?
- Pragmatics
  - Understanding the text with context, often requires additional information not within the text

Normalization

- Breaking down text into a more common form so we can do higher level NLP tasks
- Word segmentation or tokenization to identify what are words
- Using a space? Need to be aware of exceptions
  - E.g. Burger King is a single entity despite having a space
- Chinese has different word segmentation algorithms
- Removal of "stop words" like "the" and "a" which do not aid meaning, but still need to be careful:
  - E.g. The 1975 won a Brit award (referring to "The 1975" band)
  - E.g. The 1975 United Kingdom European Communities membership referendum resulted in entry to Europe (referring to the year 1975)

Word embeddings: bag-of-words

- Word embeddings are a vectorized representation of our text
- Bag-of-words is a simple form, ignores grammar and word order
- Words are represented as a “bag”, with their frequency
- Can also give positive/negative scores for words and combine with their frequency
- Take an average to get a score.. but ignores word order, which impacts meaning

Extending to n-grams

- Can also look at n-grams, which take multiple items (like words) together
- Google's Ngram Viewer - https://books.google.com/ngrams - search engine for n-grams with printed books between 1500-2008
- But n-grams still struggle to capture the negative meaning in a sentence like "it was not at all good"
- What about counting the number of co-occurances of words in sentence, extending the vector to a matrix? But this results in a very sparse matrix (many words will not co-occur with others)
- These are handcrafted features, involving a rules based approach
- Difficult for a rules based approach to be absolutely exhaustive
- Machine learning to create a dense word embedding?

word2vec and BERT

- As the name suggests converts words to vectors!
- Computes the probability that words are likely to be written near each other ie. a probabilistic classifier
- Creates a dense representation
  - CBOW (continuous bag of words) tries to predict the target word from the context of other words around it
  - Skip gram works in the opposite direction
- “context” here means words near it with a specific sized window
- Also have a similar method GloVe (ratio of co-occurances)
- Newer techniques like BERT (Bidirectional Encoder Representations from Transformers) give different vector representations of the same word depending on context (e.g. bank for “river bank” and “bank deposit”)

Topic modelling

- So far mostly discussed words and documents
- What about topics, which sits between words and documents?
- Document is a number of topics and each topics consist of a group of words
- LDA (latent Dirchlet allocation) is a technique for extracting groups of words
- It’s “latent” because we can’t observe the topics, whereas we can observe words and documents
- LDA helps us find the distribution of topics in a document, the number of topics and how those words are distributed

Tools for NLP and text

Collecting text and cleaning

- Regular expressions in Python (and in many programming languages) are a good start
- Extracting text
  - BeautifulSoup – extracts text from webpages, stripping unnecessary tags
  - selenium – web browser emulator
  - scrapy – web scraping crawler
  - Twython – Python wrapper for Twitter's API to read tweets
  - search-tweets-python – Python wrapper for enterprise Twitter
  - tabula-py – Python wrapper for Tabula (Java), to extract tables from PDF
  - PDFMiner.six – extract text from PDF
  - newspaper – extract newspaper articles from web

NLP tools

- NLP tasks
  - NLTK –most well known NLP library for Python
  - spaCy – many NLP tasks like extracting entities from text
  - textblob – easy to use wrapper for NLTK
  - gensim – topic modelling (includes LDA & word2vec)
  - Stanford OpenNLP – natural language library
  - BERT – TensorFlow code and pre-trained models
- Commercial solutions for NLP available from Refinitiv

Challenges in NLP

- Entity matching
  - Translating brands to traded assets
    - E.g. an article might mention Audi A6, but Audi is not a tradable asset (its parent company Volkswagen is)
  - Matching people to roles
    - E.g. Barack Obama as President of USA during office, but as a former president after his office
  - And much more!
- Sentiment analysis
  - Training against a specific domain (e.g. finance) vs. a generalized model
- Can be easier to use text data which has already been structured rather than attempting to structure the dataset yourself

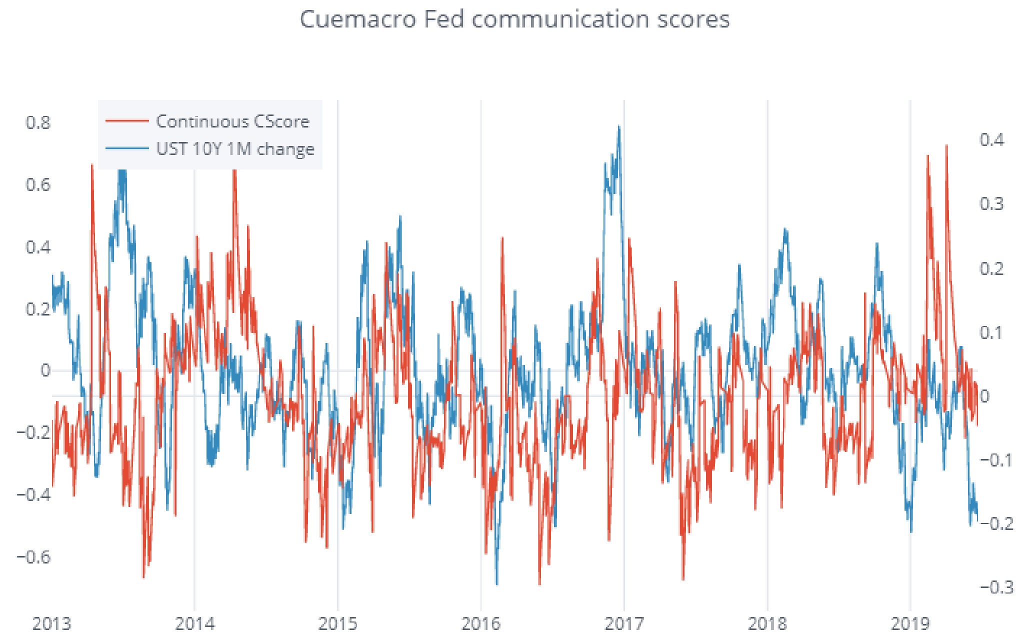
Case study: Federal Reserve Communications

Federal Reserve data

- Federal Reserve regularly communicates with markets
- Through speeches, statements, minutes etc.
- Market reacts to this!
- Can read publicly available communications from the web
- Create a dataset of web communications
- Apply NLP to determine the sentiment of individual texts
- Construct an index to give an overall view of FOMC sentiment
- Positive sentiment is hawkish whilst negative sentiment is dovish

Fed sentiment vs. UST10Y yield changes

- Can see a relationship between them, as we would expect



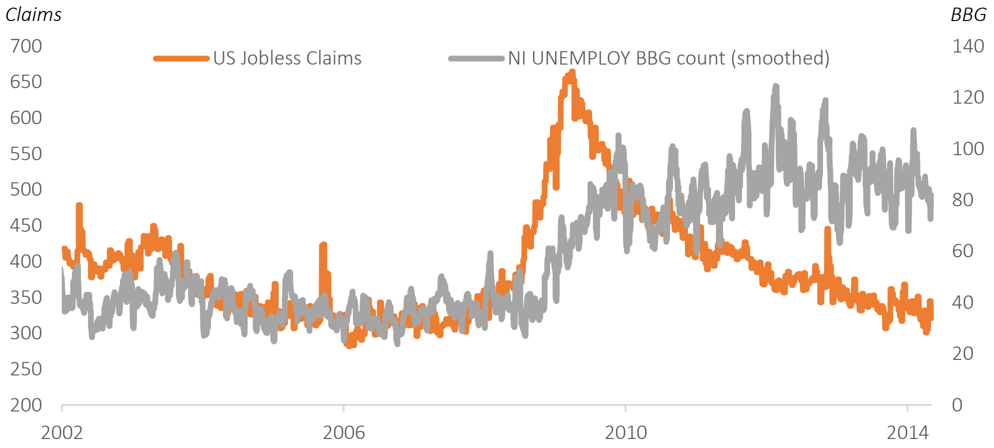
Case study: Bloomberg News to trade FX spot

Unstructured & structured news data

- Unstructured news data
  - Read news articles, blogs etc. in their raw text form, then clean and then directly apply text based analysis to add tagging and other fields
  - Very time consuming as we need to handle large amounts of data and also need to do natural language processing, which is non trivial
- Structured news data
  - Vendors processes a large amount of news from numerous sources into a more manageable dataset for us to explore
  - Data more easily accessible with additional fields (eg. tagging topics)
  - Traders can concentrate on creating effective trading rules and running risk, rather than spending that time dealing with cleaning up massive quantities of unstructured news

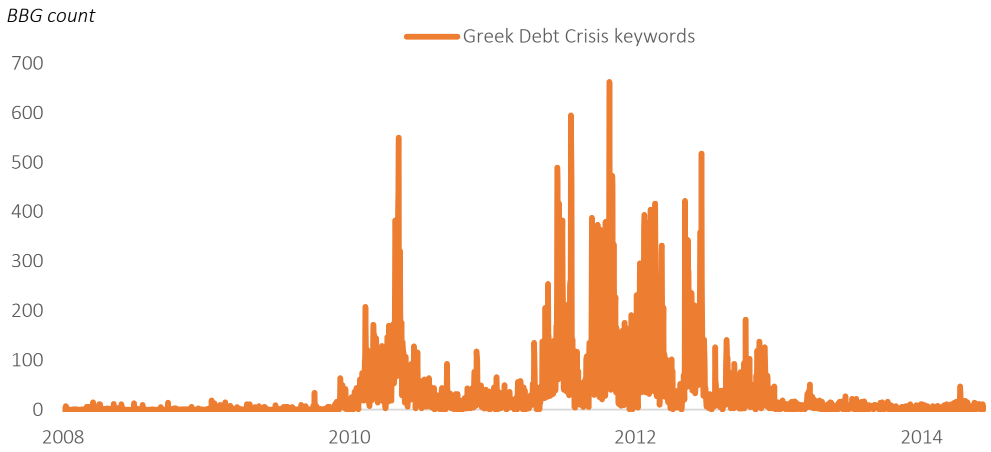
Automating news filtering

- Using news to trade markets is not new idea
- A trader essentially “filters” news into the “signal and the noise”
- But there is simply too much news for humans to read!
- How can we read news in automated fashion?
- Easier to use structured news datasets
- However, what news filters do we use?
- News related to unemployment?
- Buy/sell signals?



General approach to news filtering

- Several approaches
  - Pick words or sectors which are relatively generic (and also intuitive) like “job cuts”
  - The approach to this “picking” depends on our data source, each one is different
  - Fit the best words according to a backtest!
- “Fitting” words which are not obviously related is data mining
- Resulting model will likely be unstable when run live
- Also caution when using hindsight to pick words
- For example, “Greek debt crisis” was obvious
- But only after the event!
- NT<GO> is nice way to visualise news
- Bloomberg has machine readable news
- Use natural language processing



Specific steps for text datasets

- We can formulate a few generic steps that are used when dealing with a text based dataset for trading purposes
  - Raw data collection – web scraping and accessing internal databases
  - Cleaning dataset – removing HTML tags and invalid observations
  - Structuring dataset – adding tags (eg. sentiment) and compress into single database record
  - Filtering dataset – choose most relevant entities/topics to prune search space
  - Create an indicator – aggregate records to create indicators
  - Apply a trading rule to the indicator – how to convert into buy/sell signals directly or added to other trading factors (eg. carry)

Using Bloomberg News dataset

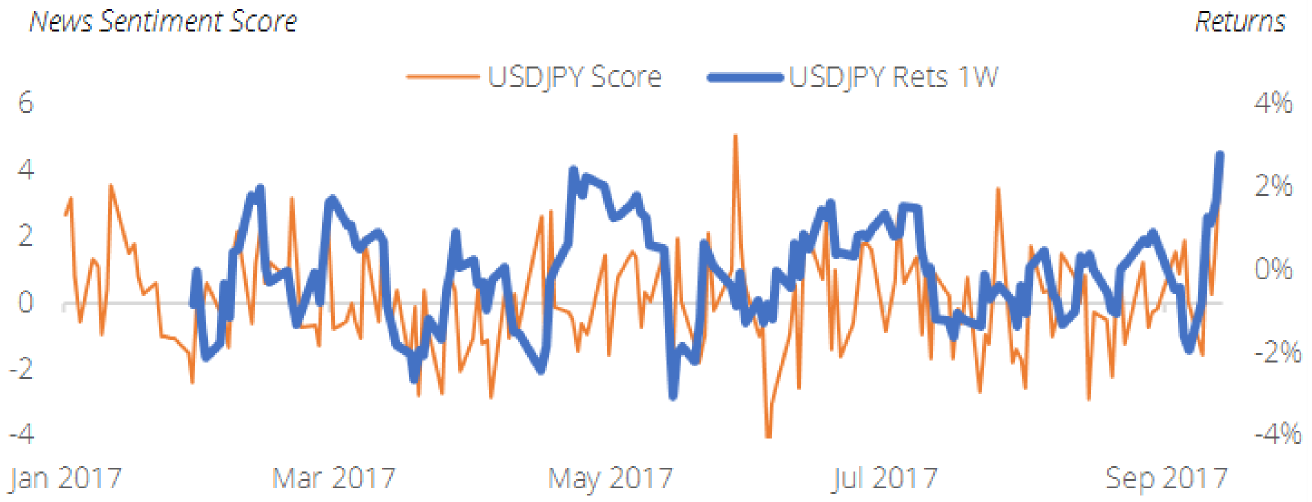
- We shall use a dataset consisting of Bloomberg News articles from 2009-2017
- It is a structured dataset, which saves time (eg. we avoid the time consuming raw data collection step)
- Bloomberg News is written in a consistent style, so easier to process than general web content
- Each news article has a number of fields tagged including:
    - Timestamp of news article
    - Title of news article
    - Text body of the news article
    - Tagging for tradable tickers related to the news (eg. %EUR for EURUSD)
    - Tagging for the topic related to the news (eg. FED for articles related to Federal Reserve)
- Company specific news also has additional news analytics fields such as sentiment, readership statistics etc.
- Topics we choose will depend on underlying dataset

Generate news signals for FX

- We want to use news to inform FX trading strategies
- Want to develop longer term strategies (ie. not high frequency headline trading)
- Hence, focus will be on macro specific news to trade FX in particular
    - Tickers: %EUR, %GBP, %AUD, %NZD, %USD, %CAD, %NOK, %SEK and %JPY
    - Topics: FED and ECB
    - Could have chosen many other relevant macro topics
- Helps us prune the search space to most relevant news
- Steps we shall do
    - Clean body text slightly (eg. remove start of article)
    - Ignore very short articles as difficult to gauge sentiment
    - Apply sentiment analysis for each article (shall use open source Python based libraries)
    - Aggregate data into daily observations (careful about holidays!)
    - Create indices for each currency/topic (Z scores for comparability)
    - Also generate a news volume score (Z score for comparability)

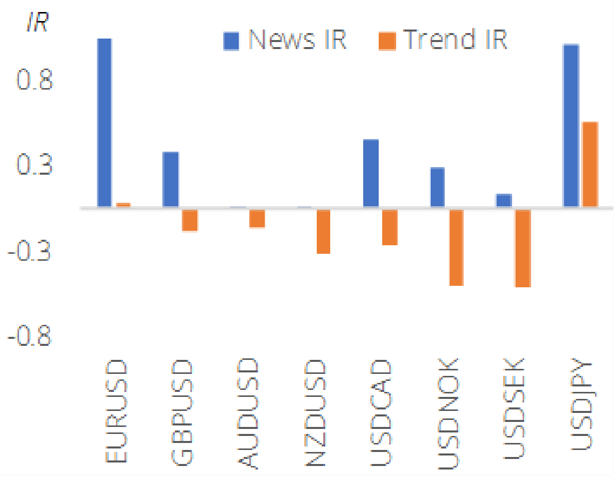
Currency pair sentiment score

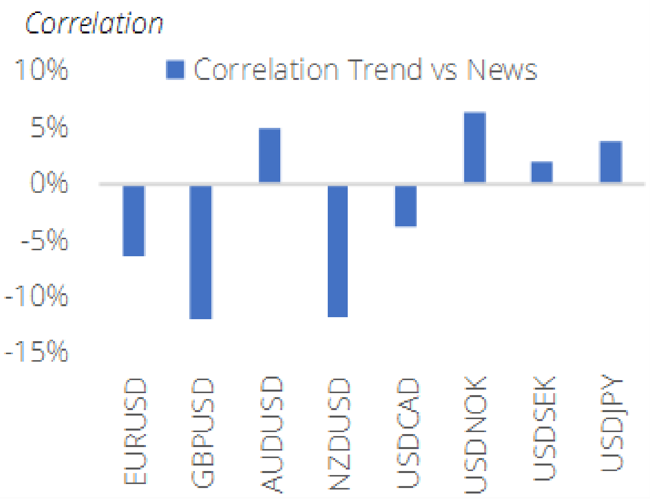
- Currency pair score = base score – terms score
- When eg. USD/JPY score is positive buy, otherwise sell



News trading rule by currency pair

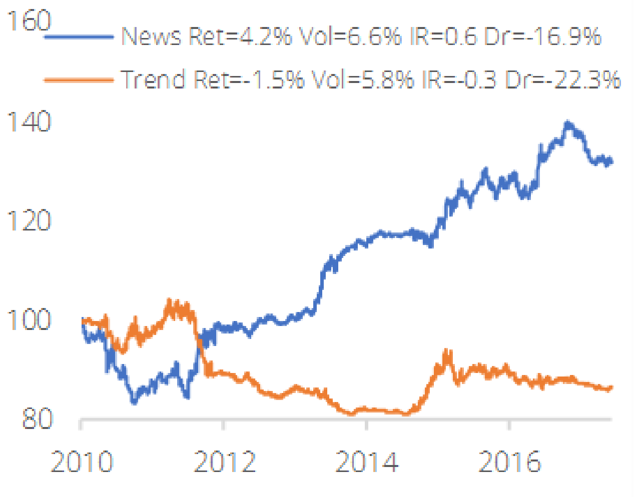
- Present risk adjusted returns and compare to a generic trend following strategy
- Apply vol targeting in each instance
- News based trading role outperforms trend significantly in our sample

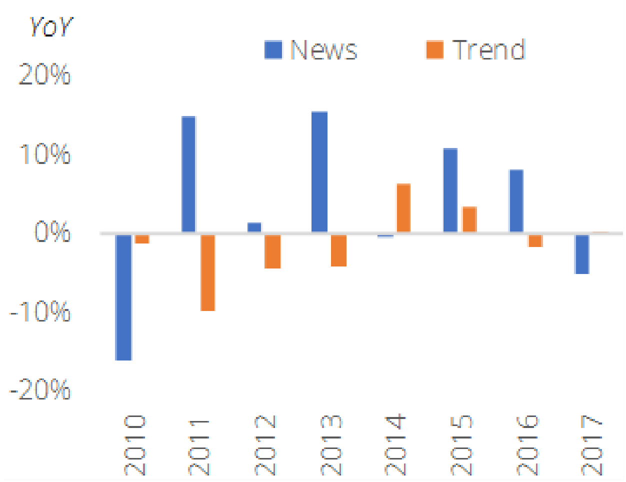




News trading rule as basket

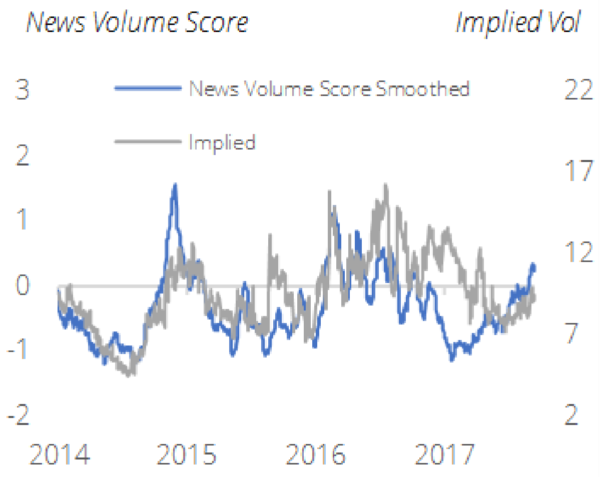
- Create news and trend baskets
- News basket heavily outperforms trend basket

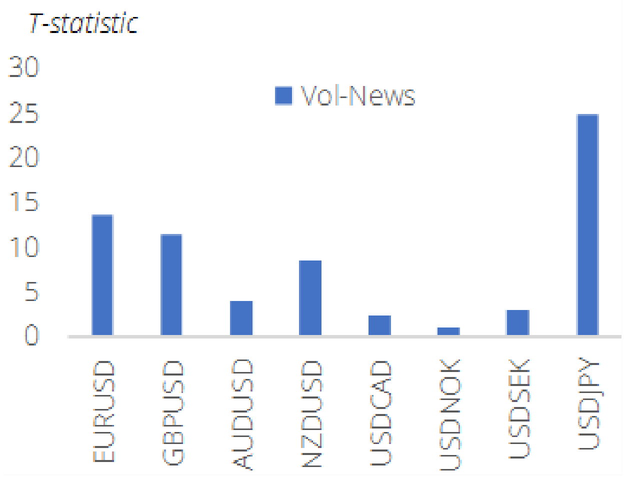




What about news volume?

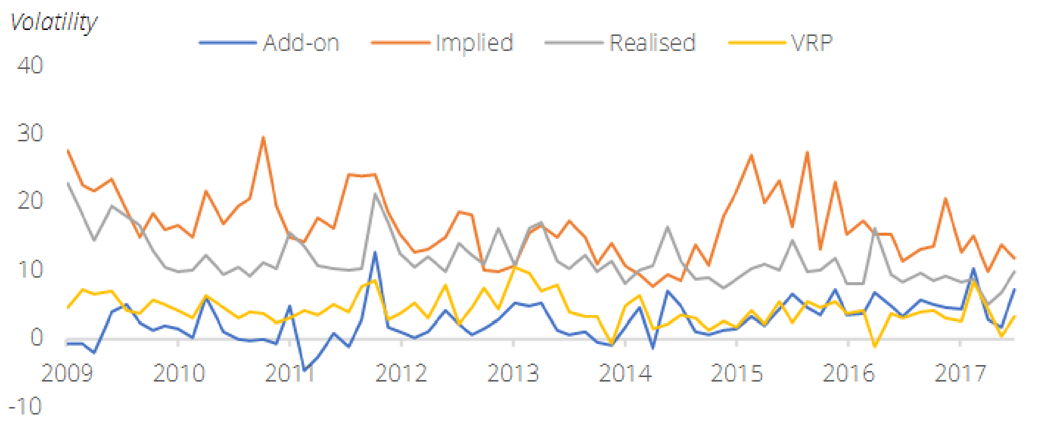
- News volume on a currency pair is heavily correlated with its implied volatility, which seems intuitive!
- T statistics show a statistically significant relationship in nearly every currency pair in our sample
- News volume can be used to help us model FX volatility – is FX volatility in line with what we could expect based on newsflow?

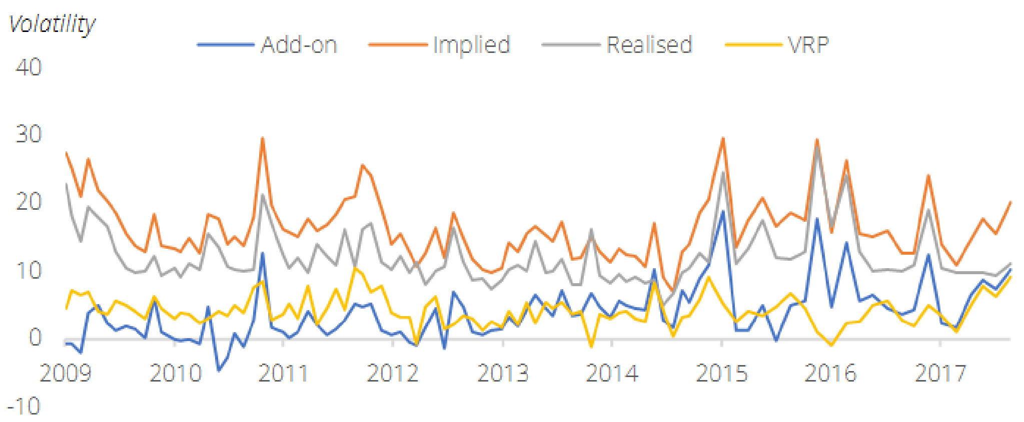




Scheduled events

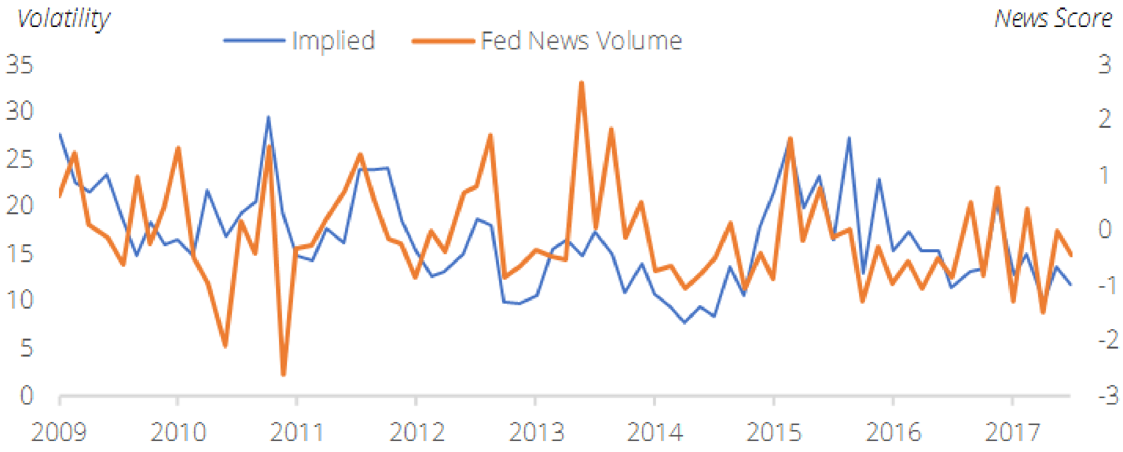
- Before scheduled events, FX vol market makers will mark up vol curve
- Known as event volatility add-on
- LHS show EUR/USD ON vol on Fed days, and RHS for ECB days (ignores all other days)
- Have model for estimating add-on (assumes only one big event per day)
- Typically, realized underperforms on these days.. Sell vol*!
- *within reason…

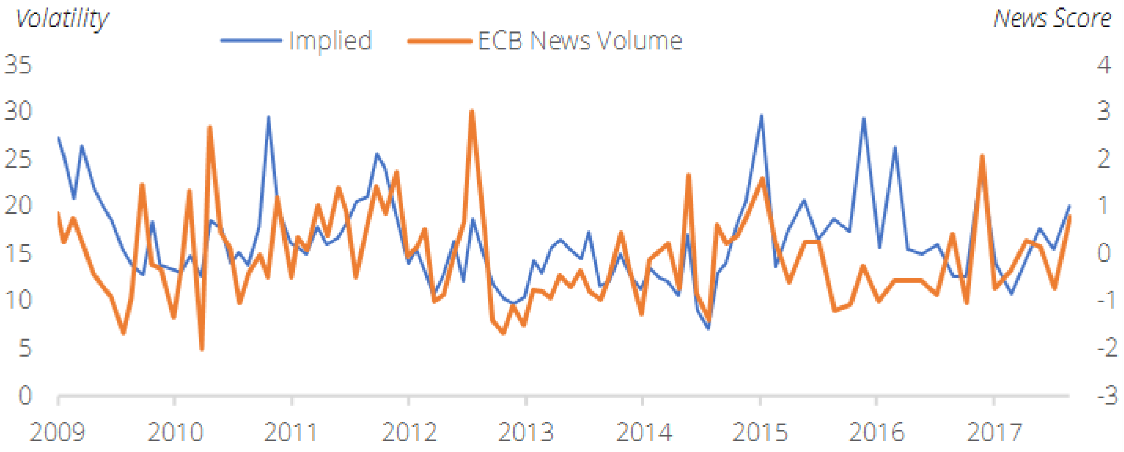




News for scheduled events

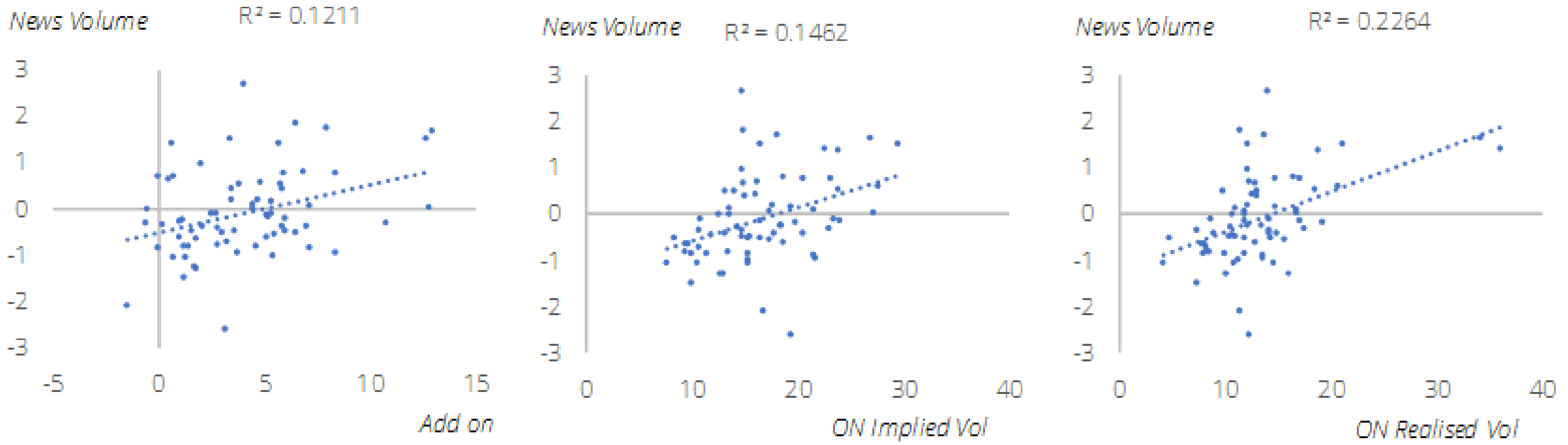
- Can we use news around scheduled events, eg. FED and ECB topics in our case to inform where the add-on is
- And also to give us an idea of where realized vol would be subsequently? Gamma traders are taking a view on where implied – realized will be
- There does seem to be a relationship between EUR/USD vol and news before FOMC and ECB meetings





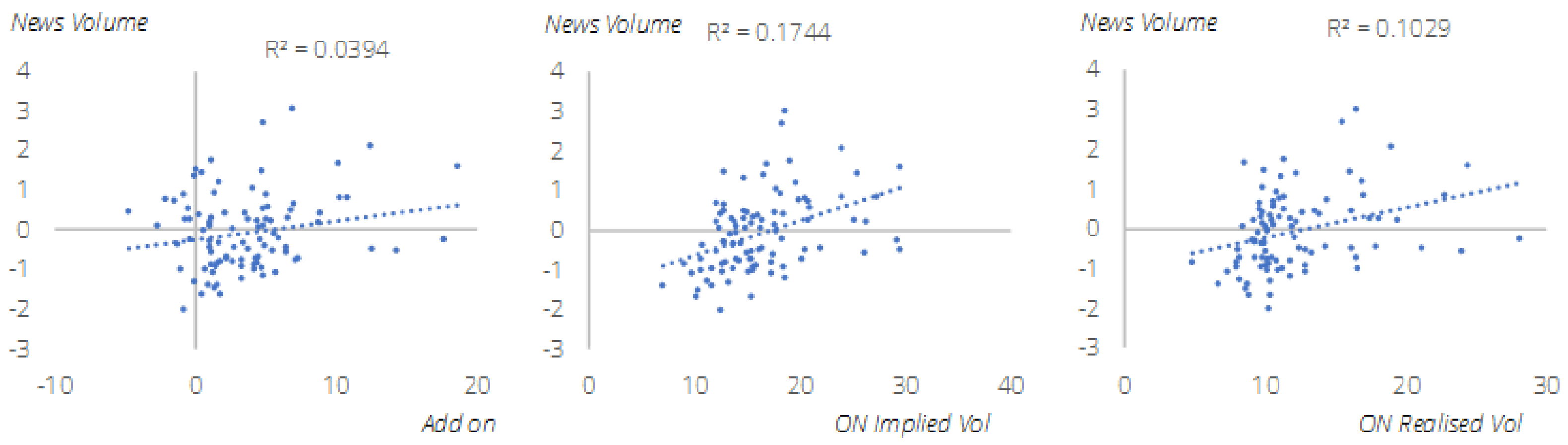
EUR/USD vol and news on FOMC days

- Showing news volume versus add-on, implied and realized ON in EUR/USD on FOMC days



EUR/USD vol and news on ECB days

- Showing news volume versus add-on, implied and realized ON in EUR/USD on ECB days



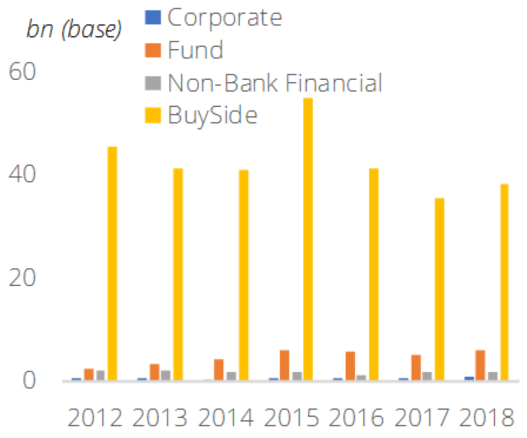
Case study: CLS FX flow data to trade FX spot

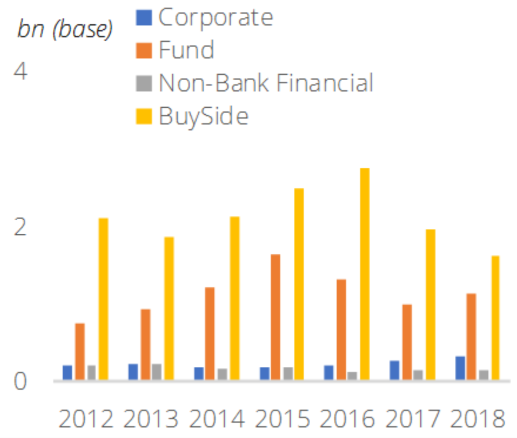
CLS data

- FX is a more fragmented market than other asset classes
  - Vast majority is OTC
  - Many different trading venues
  - Bilateral trading
- Difficult too find comprehensive FX volume & flow data
- CLS settle most OTC deliverable FX – coverage over 50% of market
- They collect and distribute
  - Hourly FX volume data
  - Hourly FX flow data for price takers
  - 30 minute lag – historical data since later 2012

EUR/USD volume vs. abs net

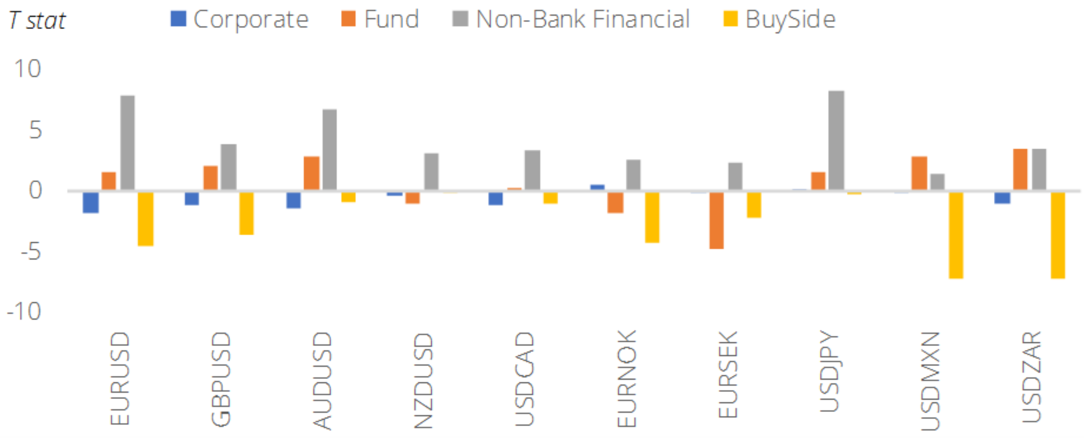
- Buy side encompasses fund, non-bank financial and fund
- Buy side as a whole is relatively two-way
- Fund tends to be more directional





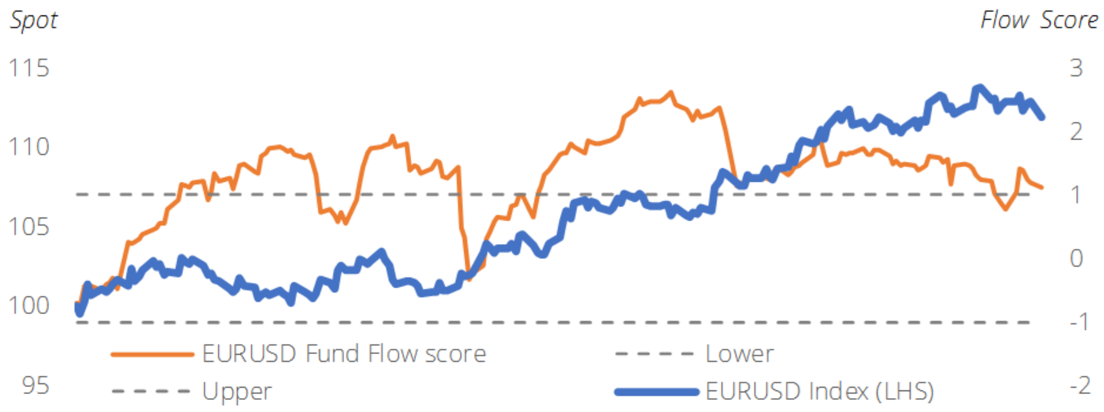
Flow vs FX spot regression

- Report T-statistics of multiple regressions for each FX pair
- Positive coefficients for fund and non-bank financials
- Negative coefficients for buy side and corporate



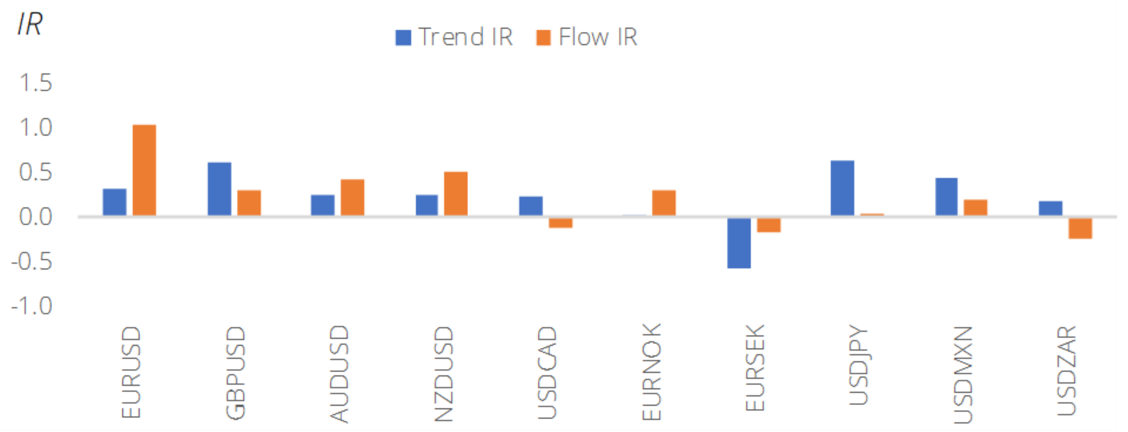
Create fund FX flow index

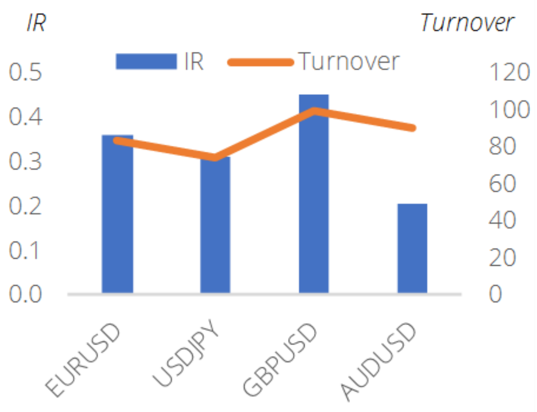
- Use fund FX flow data – tends to be more directional and positive correlation with spot
- Create fund FX flow index
  - Buy spot when very positive
  - Sell spot when very negative



Risk adjusted returns by cross

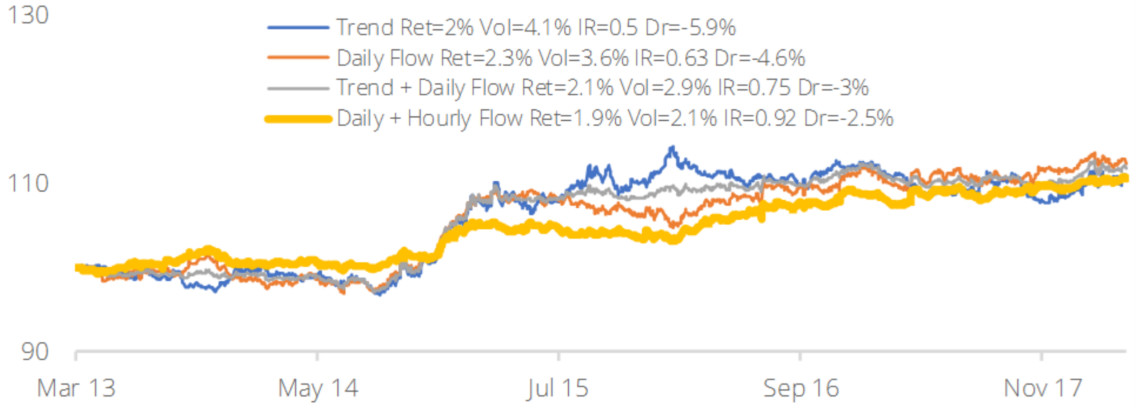
- Present historical trading returns
- Daily strategy (left) and hourly trading strategies (right)
- Stick to more liquid pairs for hourly strategy

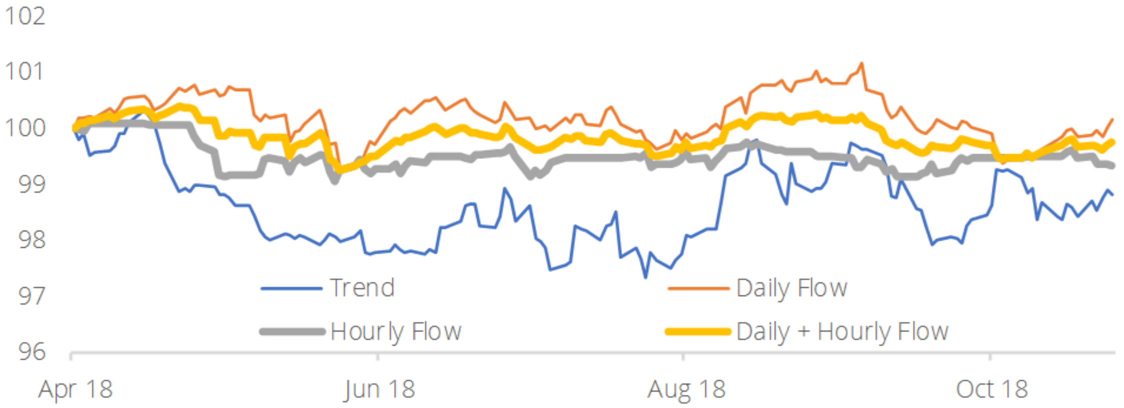




Creating daily & hourly flow baskets

- Create trading baskets for daily and hourly flow strategies
- Historically, improves risk adjusted returns vs trend alone
- In-sample (left) and out-of-sample (right)
- Flow outperforms trend out-of-sample





Case study: Geospatial Insights satellite data to estimate EPS

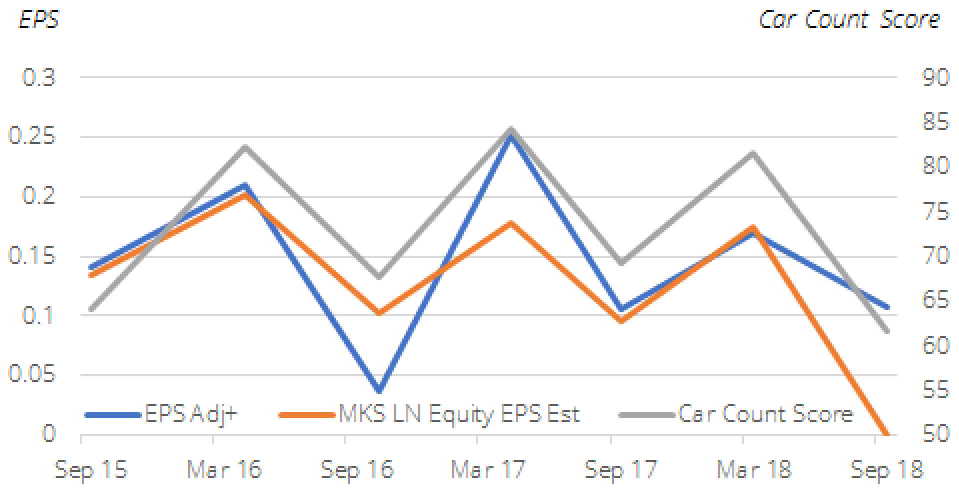
Geospatial Insights: RetailWatch

- It is well known that satellite photography can be used to help forecast earning per share for retail stocks
- Has been used extensively in US markets (Orbital Insight), but not as much for European firms
- Uses car counts as a proxy for retail activity
- RetailWatch covers a number of European retailers (both publicly traded and private companies)
- Relatively new dataset

Using car counts to estimate EPS

Cuemacro

- Created a car count score based upon the 6 months of activity related to the earnings period
- Compare against Bloomberg's consensus and actual EPS
- Present results for Marks & Spencers



Preliminary Results from The Book of Alternative Data (Wiley) est 2020

Case Study: Alternative data for private investing/risk management

Data for private investing

- Private companies are not required to disclose as much information
- Hence, more opaque market
- Also challenging because there's no "comparison" to benchmark against, like EPS for public firms
- So what can venture capital and private equity firms do?
- What are the solutions?
  - Trying to proxy the private company by publicly traded competitors? Can capture the sector, but not idiosyncratic factors
  - Also think about using proxies for performance such as hiring, consumer activity etc. which can be tracked with alternative data

Newswires

Cuemacro

- Traditionally use newswire datasets to trade publicly tradable assets
  - Bloomberg News
  - Refinitiv (Thomson Reuters)
  - RavenPack (Dow Jones and web sources)
- Use news volume as a risk management tool (tends to be correlated to volatility) and detecting abnormal newsflow
- Use newswire datasets to track sentiment regarding larger private firms, before they IPO:
  - Uber pre-IPO

Social Media

- Twitter offer full feed access
  - Can query for specific keywords to download/count
  - History goes back 10 years
  - Can be useful for tracking sentiment for brands
- BrandWatch builds their product on top of social media

Web scraped data

- ThinkNum uses web scraping to gather statistics about firms (in a structured manner)
- Cover 400,000 companies including many private ones
- Can track company specific information such as:
  - Job hirings
  - Store locations
  - LinkedIn details
  - Web traffic for firm

Case Study: Saving “alpha” with transaction cost analysis

tcapy

- Big Data and alternative data isn't just for generating alpha
- It can also be used to "save" alpha, to reduce our transaction costs
- tcapy is a Python based library by Cuemacro which does transaction cost analysis to identify how much traders are paying for their liquidity
- Needs high frequency market tick data and also trade data from the client
- Will do a quick demo if there's time

Conclusion

- Alternative data primer, introducing the topic
- Talked about where to find data
- Dived into structuring text data
- Showed examples of how to generate (and save!) alpha using alternative data examining
  - CLS FX flow data to generate FX trading signals
  - Text based datasets for Fed communications and Bloomberg News
  - Geospatial Insights satellite imagery to estimate EPS
  - Alternative data for private investing/risk management
  - tcapy to reduce trading costs for FX

Any questions?

- Drop me an e-mail at saeed@cuemacro.com, ring me or tweet to @saeedamenfx (or even talk to me now, the old school way!)