Data Mining Techniques Applied to Econometric Panel Data for FX Forecasting

Charles Naylor

Introduction

Summary of Currency Markets

Macroeconomic forecasting is generally considered to be one of the hardest challenges in finance. In comparison with forecasts on stocks or bonds, currency (a.k.a "foreign exchange", or FX) price movements react more purely to the flows of global trade and geopolitical risk, because they have no intrinsic value. A currency's worth is quoted only relative to other currencies.

The Carry Trade

That said, there is still a basis for investing in currencies in that any holding will earn interest, just like a savings account at a bank. In the USA, the 3-month interest rate is about 2.3%. In Turkey, it's about 17%. Other things being equal, an investor could borrow money in American Dollars, then lend it out in Turkish Lira, and earn the difference between these rates for a nearly 15% annual return. This is called the **carry trade**.

The carry trade is not riskless because the spot rate, i.e. the number of USD that a Turkish Lira can buy, is not fixed over this period. In fact, undergraduate economics classes still teach that any profit possible from this trade will be neatly ironed out due to spot price movements, a phenomenon known as **Covered Interest Rate Parity**. The persistance of the carry trade in the face of theory is a reminder that economic models elide substantial frictions experienced in the real world.

Drivers of Returns

In FX, that interest rate differential, the "carry", is relatively stable but subject to punctuated equilibrium as new data comes to light. In developed markets, this data primarily consists of central bank rate decisions and the economic news that might affect those decisions. Price movements, however, are the deterministic result of countless iterative, interacting agents. While we may know many of these agents' motives, it is not possible to aggregate their behavior with any accuracy, because the actions of each agent are affected by those of all of the other agents, and small measurement errors compound. For example, it is impossible to tell the periodicity of market data without context. A day's worth of price movements at 5-minute increments looks the same as a year's worth of daily movements. The asset price measures the result of a chaotic, nonlinear dynamical system: a scale-free network`<insert reference to Barabasi>`.

Motivation

Speculators in currencies try to pocket the carry while avoiding the risk of major movements in currency prices. The strategy has been likened to picking up nickels in front of a steamroller. Thus, in spite of the difficulties inherent in forecasting currency movements, a *successful* speculator must have some idea of what those movements will look like in the future.

There are many possible methods available to create these forecasts. Professionally, I have seen econometric panel data (i.e. the same set of economic measurements repeated for multiple countries) plugged into a

Kalman Filter in order to capture the evolving relationship between indicators and their currencies, while also recognizing that one currency's movements will affect all other currencies, as well. The Kalman filter has the advantage of having a closed-form solution and being well-adapted to testing in systems in which one expects to add new data regularly. In its most basic form, the disadvantage is that a Kalman filter requires careful tuning of the relationship between its output variables, and of the covariance of evolution of its factor weightings.

Some months ago, I wrote a case study of generative Bayesian forecasting techniques`<insert GP reference>` that applied a Gaussian Process Regression to panel data for 11 of the most traded world currencies. The result was a well-specified and validated forecasting model whose error term was far larger than its signal. This jibed with the results I had seen professionally using a Kalman Filter, but thanks to posterior predictive checking the flaws are glaringly obvious. I would like to see how this forecast can be improved.

I am particularly interested by the work being done at the University of Maryland into the application of neural network reservoirs to chaotic systems.`<insert Pathak reference>` I plan to apply these techniques to the residuals of the Gaussian Process regression, or in combination with some other regression on the panel data.

Goals

- Run the panel data through a set of standard data mining techniques to provide a baseline.
- Apply reservoir computing techniques to existing or new regressions.

The Data

Data consists of the weekly currency returns, plus a raft of weekly economic data. A full

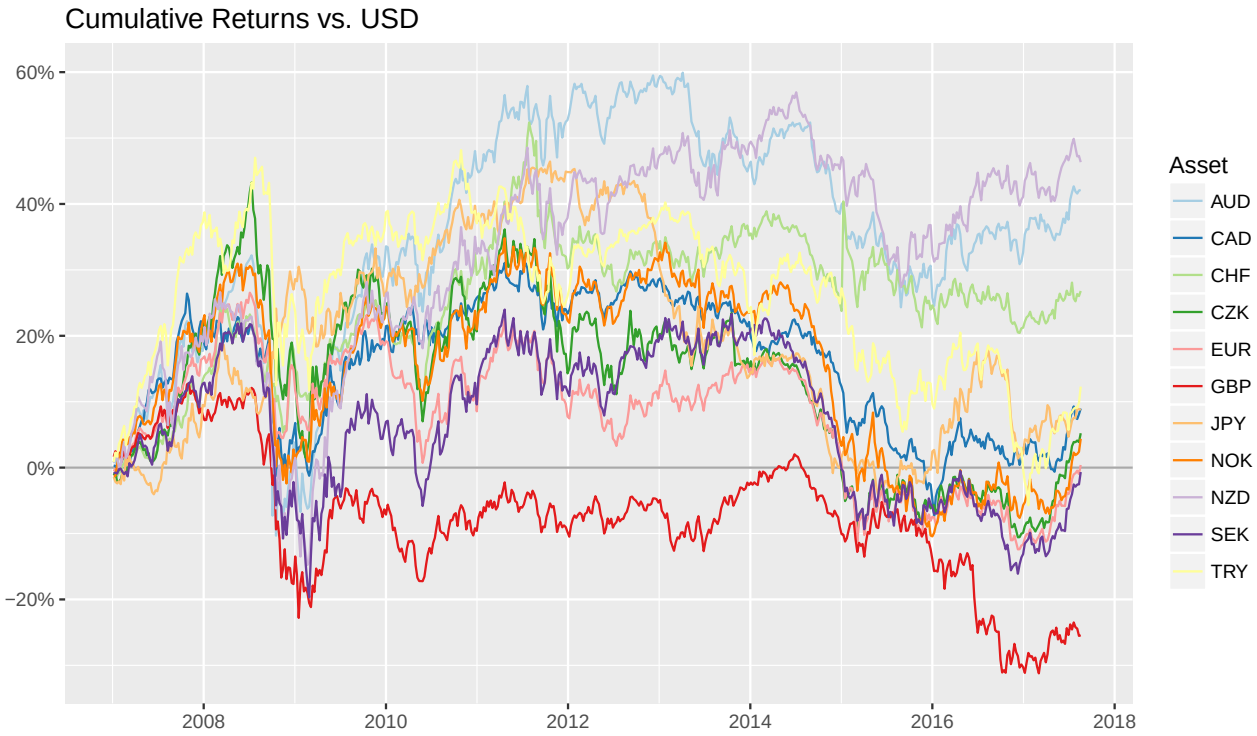
Y variables

The endogenous, 'Y' variable comprises weekly currency returns on 11 currencies for a ten-year period starting in 2007 and ending in 2017.

Weekly currency returns consist of the change in spot rate against the US dollar, plus the carry, defined as the (time-adjusted) difference in 1 month forward rates between the local currency and the USD. Without going into detail, the forward rates are a reasonable proxy for what a professional speculator might expect to earn by placing trades using currency derivatives.

It's important to note that our endogenous variable is *multivariate*. We cannot assume that different currencies have returns that are independent of one another.

```
## Warning: Removed 22 rows containing missing values (geom_path).
```



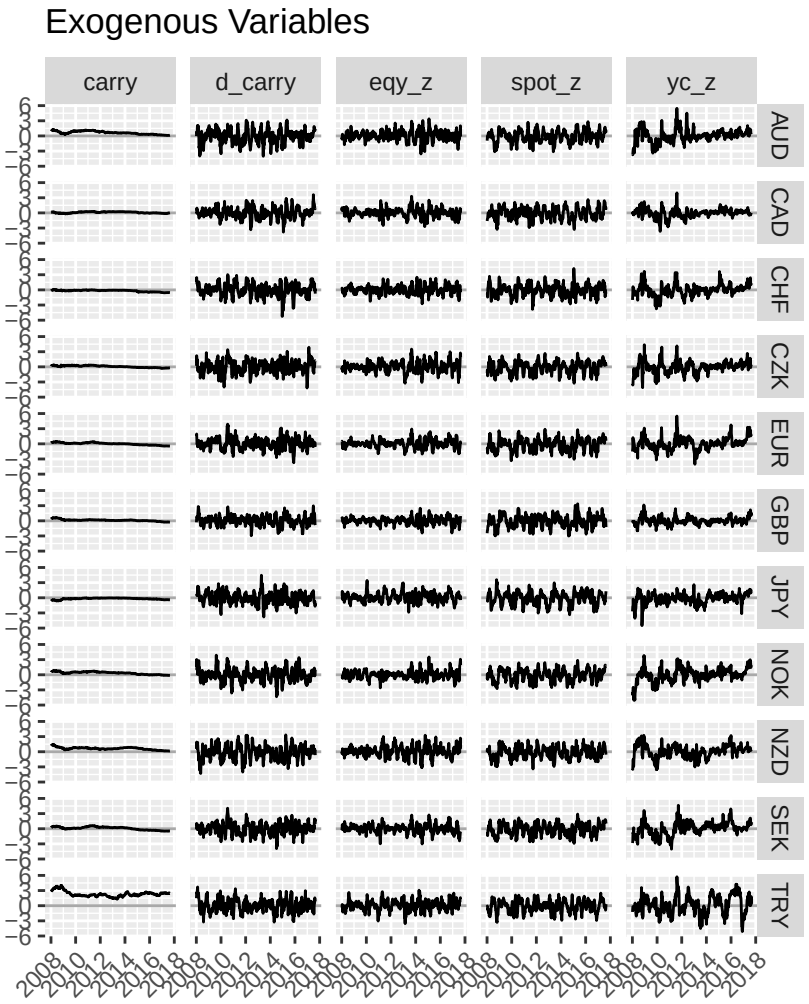
X variables

The exogenous, 'X' variables consist of weekly data for the following:

Factor	Description
Equity_d8W	The 8 week change in the local equity index
Spot_d8W	The 8 week change in spot rates
SwapRate2Y	The 2Y Swap Rate
TwoYear_d8W	The 8 week change in 2Y Swap rate
YieldCurve	The spread between 10Y and 2Y Swaps

There is clear room for improvement if we were to fit a model that permitted multiple time scales, such as a MIDAS regression`<insert midas ref>`. It would also be usual to include some sort of measure of liquidity and credit conditions.

These variables have been scaled and normalized.



Note that Carry is not a stochastic variable. It is, however, a rate, and impacts asset returns consistently at that rate. It's also worth noting the degree to which rate differentials have vanished since the Credit Crisis of 2008.

Models

Classification

Introduction: Regime Switching Models

There has been a substantial amount of work done in economic forecasting in which time periods are split up into *regimes*, periods in which markets are expected to behave similarly given similar data. For example, the markets in the run up to the bursting of the tech bubble in 2000, or the housing credit bubble in 2008, behaved substantially differently to the markets just after those bubbles. The Kalman Filter and Gaussian Process regression models attempt to account for changing reactions to market conditions by fitting a continuously changing regression to those conditions.

One alternative is a **Regime Switching Model**. Split the time periods into several regimes, then run a separate regression for each regime. The process is complicated by the fact what it can be extremely difficult to tell at any given moment in which regime markets were operating at any given time, even in hindsight.

We will attempt determine these regimes by running various classifiers and clustering algorithms against the exogenous and endogenous data.

Hierarchical Clustering

We will need to compute a dissimilarity matrix between periods. This calculates the distance between all of the values we have for each week.

We ought to try fitting both a global set of regimes, and one per country, as we are primarily interested in what happens when there are differences between countries.

Global Regimes

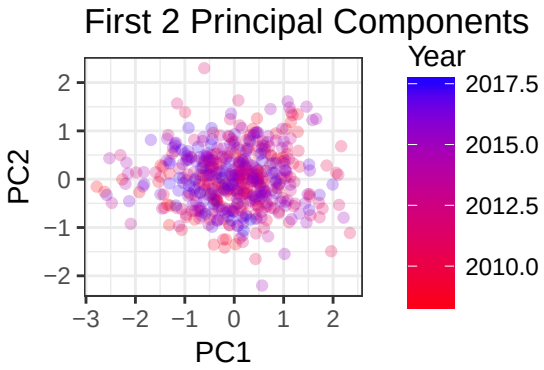
Here we assume the market reactions to economic conditions are the same across all countries.

```
# Organize data into per-row observation matrix
endo %>%
  gather(Asset, value, -Date) %>%
  mutate(Exog="Endo") %>%
  bind_rows(exogs) %>%
  filter(Date >=as.Date("2008-01-04")) %>% #endo has more data than exog at the beginning
  filter(Date <=as.Date("2017-08-18")) %>% #exog has more data than endo at the end
  unite(asset_exog, c("Asset", "Exog")) %>%
  spread(asset_exog, value) ->
  all_obs_matrix

all_obs_matrix %>%
  select(-Date) %>%
  daisy(metric="euclidean") ->
  global_dissimilarity
```

Principal Component Analysis

It's going to be hard to see what any of these clusters looks like, as we are working with a high-dimensional space. First, let's take a look at groupings of weeks using the first two principal components. These can form the basis of later visualizations.

```
all_obs_matrix %>%
  select(-Date) %>%
  prcomp() ->
  global_pca

#Plot
(global_pca$x %*% global_pca$rotation[,1:2]) %>%
  as_tibble() %>%
  bind_cols(all_obs_matrix %>%
              transmute(Year=decimal_date(Date))) %>%
  ggplot(aes(x=PC1,y=PC2, col=Year)) +
  theme_bw() +
  scale_color_gradient(low="red",high="blue")+
  geom_point(alpha=0.25) +
  ggtitle("First 2 Principal Components")
```

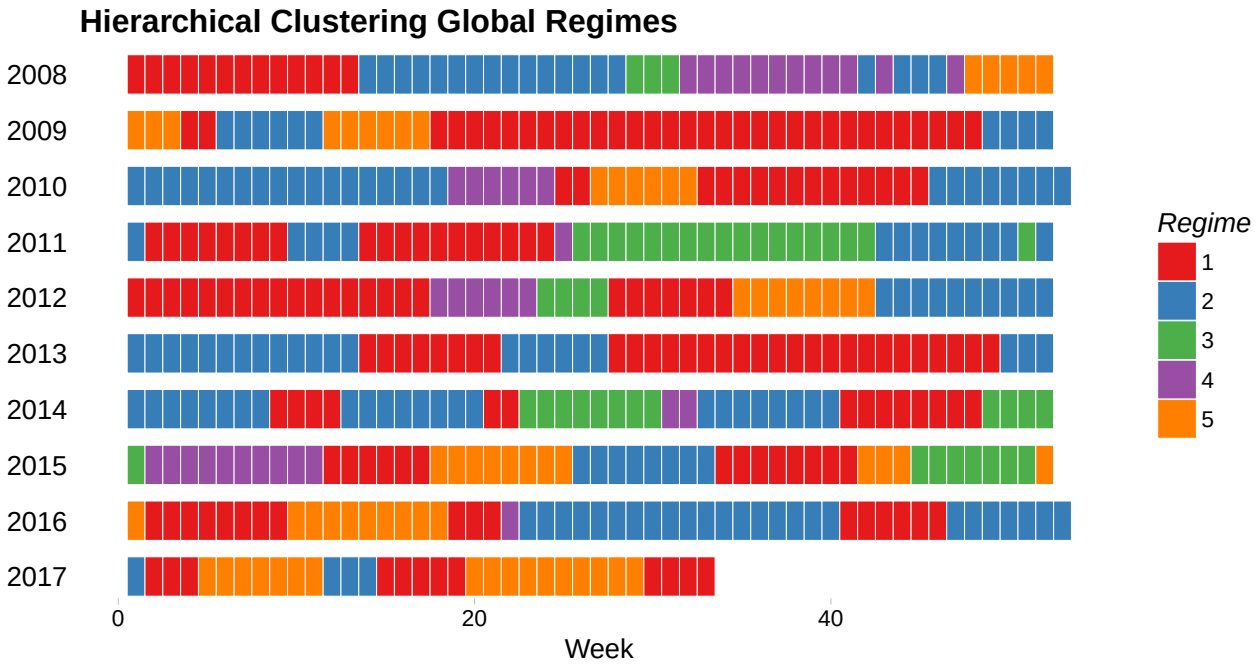


We do not see much in the way of clear regime clusters.

First, try plain hierarchical clustering on the dissimilarity matrix. Ultimately, we have 503 weeks of data, so we will need to cut the cluster tree a bit higher up in order to be able to see the results.

We can visualize the higher clusters with a calendar heatmap.

```
global_hclust <- hclust(global_dissimilarity, method='complete')

all_obs_matrix %>%
  select(Date) %>%
  mutate(Regime=factor(cutree(global_hclust, k=5)),
         Year=factor(year(Date)),
         Week=week(Date)) %>%
  ggplot(aes(x=Week, y=0, fill=Regime)) +
    theme_pander() +
    facet_grid(Year~., switch="y") +
    geom_tile(color="white") +
    scale_fill_brewer(type="qual",palette="Set1") +
    theme(strip.text.y = element_text(angle=180),
          axis.text.y=element_blank(),
          axis.ticks.y = element_blank()) + ylab("") +
    ggtitle("Hierarchical Clustering Global Regimes")
```

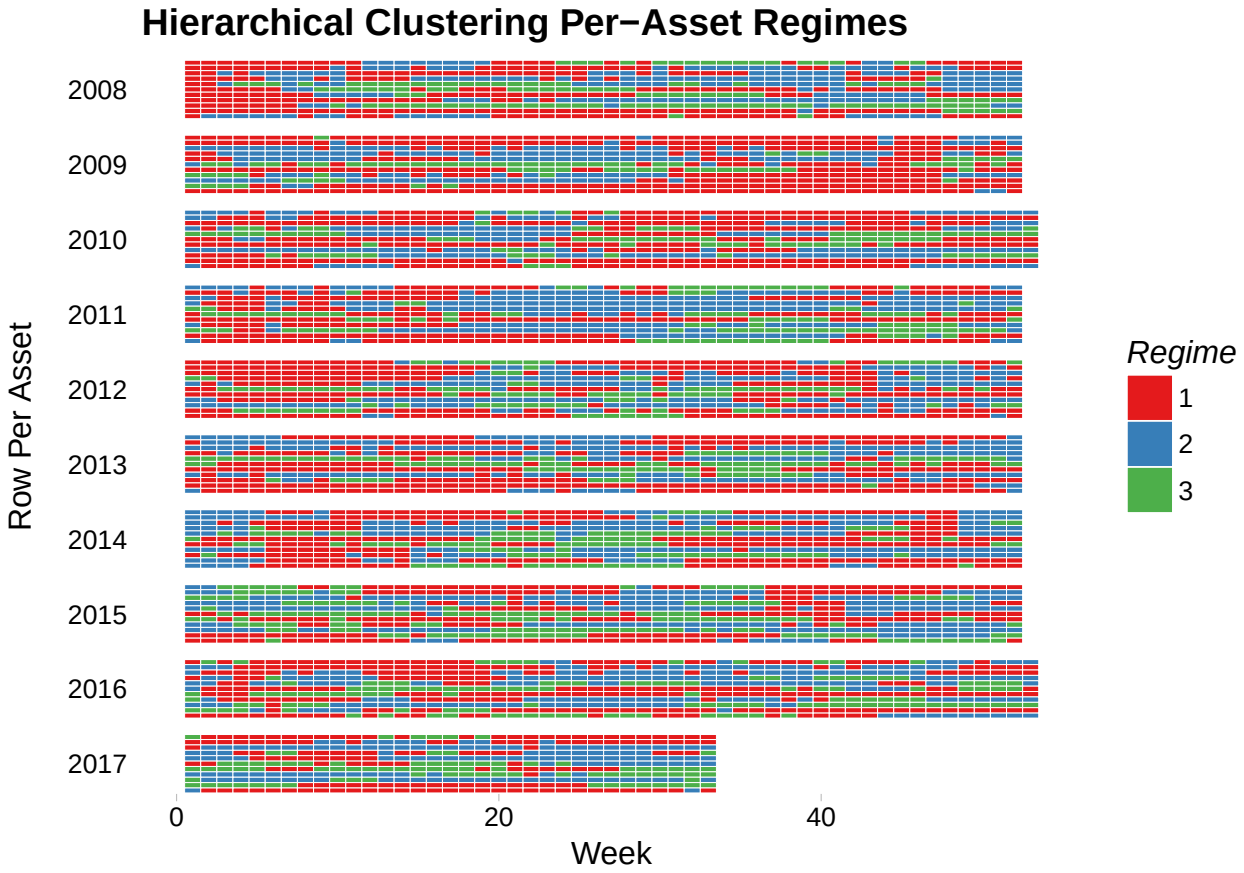


The good news is we have a lot of regimes in contiguous blocks, so we have clearly identified something. 5

regimes is probably too many; 2 or 3 would be more likely, but it would have been difficult to validate if regimes were contiguous if we had plotted so few.

Per-Country Regimes

How similar are the regimes between countries? There may be an identification issue as there's no guarantee that, e.g. regime 1 in one country will be encoded as regime 1 in another, even if they refer to similar underlying clusters.

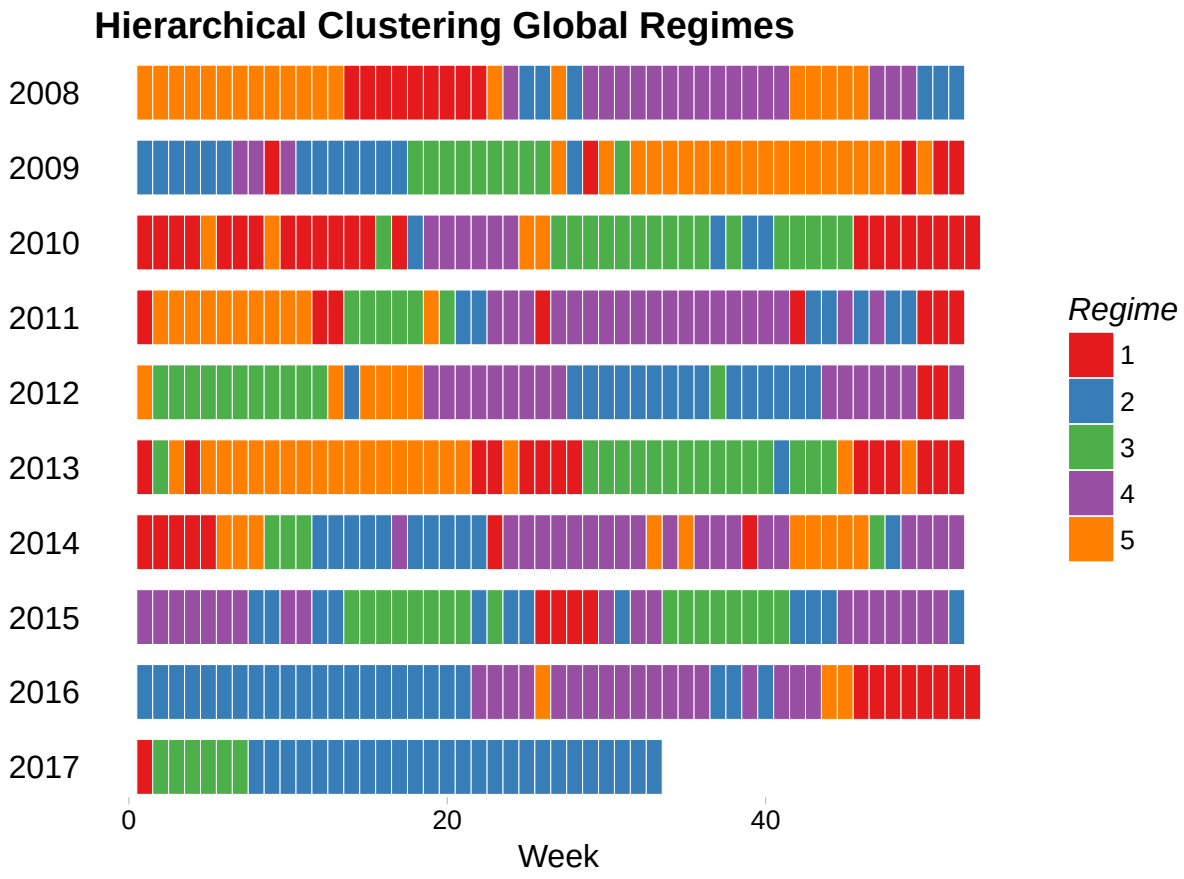


There is still quite a lot of commonality between regimes. The least similar currency is the Turkish Lira, on the bottom row. The British Pound (middle row) enters a new regime 3 weeks *before* the vote on exiting the EU, in week 25 of 2016, which is evidence against the validity of this technique.

K-means Clustering

```
global_kmeans <- kmeans(all_obs_matrix %>% select(-Date),
                        centers=5)

all_obs_matrix %>%
  select(Date) %>%
  mutate(Regime=factor(global_kmeans$cluster),
         Year=factor(year(Date)),
         Week=week(Date)) %>%
  ggplot(aes(x=Week, y=0, fill=Regime)) +
    theme_pander() +
    facet_grid(Year~., switch="y") +
```

```
    geom_tile(color="white") +
    scale_fill_brewer(type="qual",palette="Set1") +
    theme(strip.text.y = element_text(angle=180),
          axis.text.y=element_blank(),
          axis.ticks.y = element_blank()) + ylab("") +
    ggtitle("Hierarchical Clustering Global Regimes")
```



Support Vector Machines

It might be possible to use support vector machines for forecasting. As with most other asset classes, currency returns have fat tails, and the majority of an investor's profit (or loss) is typically made in a small number of periods. Suprise economic news or geopolitical events such as the British vote to leave the EU can cause the markets to reassess appropriate price levels drastically. As a result, when using a classifier to forecast it is important to distinguish between large moves and smaller ones. Categorizing merely by positive or negative returns would throw out some of the most important information.

Random Forests