Bayesian Variable Selection for Nowcasting Economic Time Series

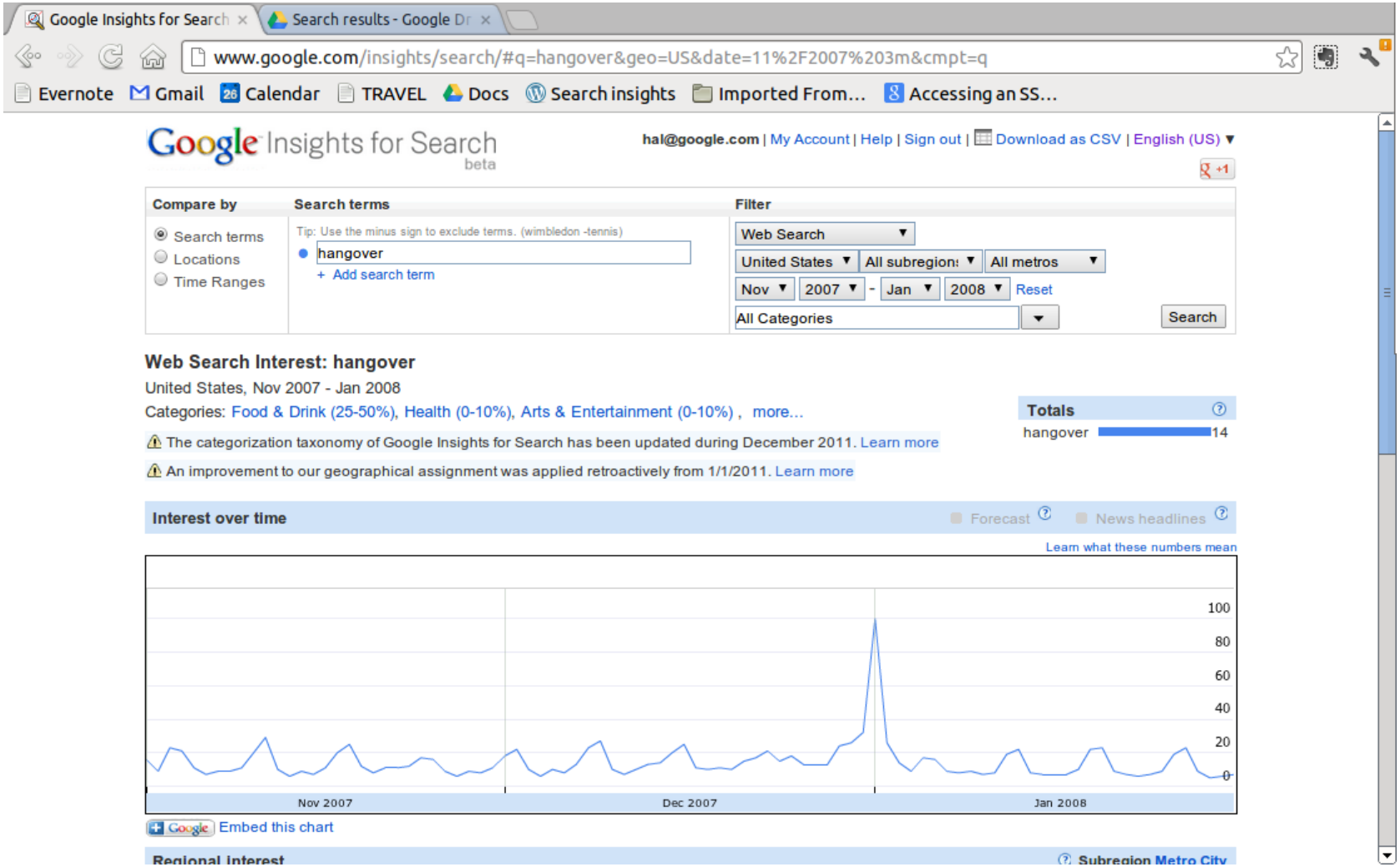
Steve Scott
Hal Varian

August 31, 2012

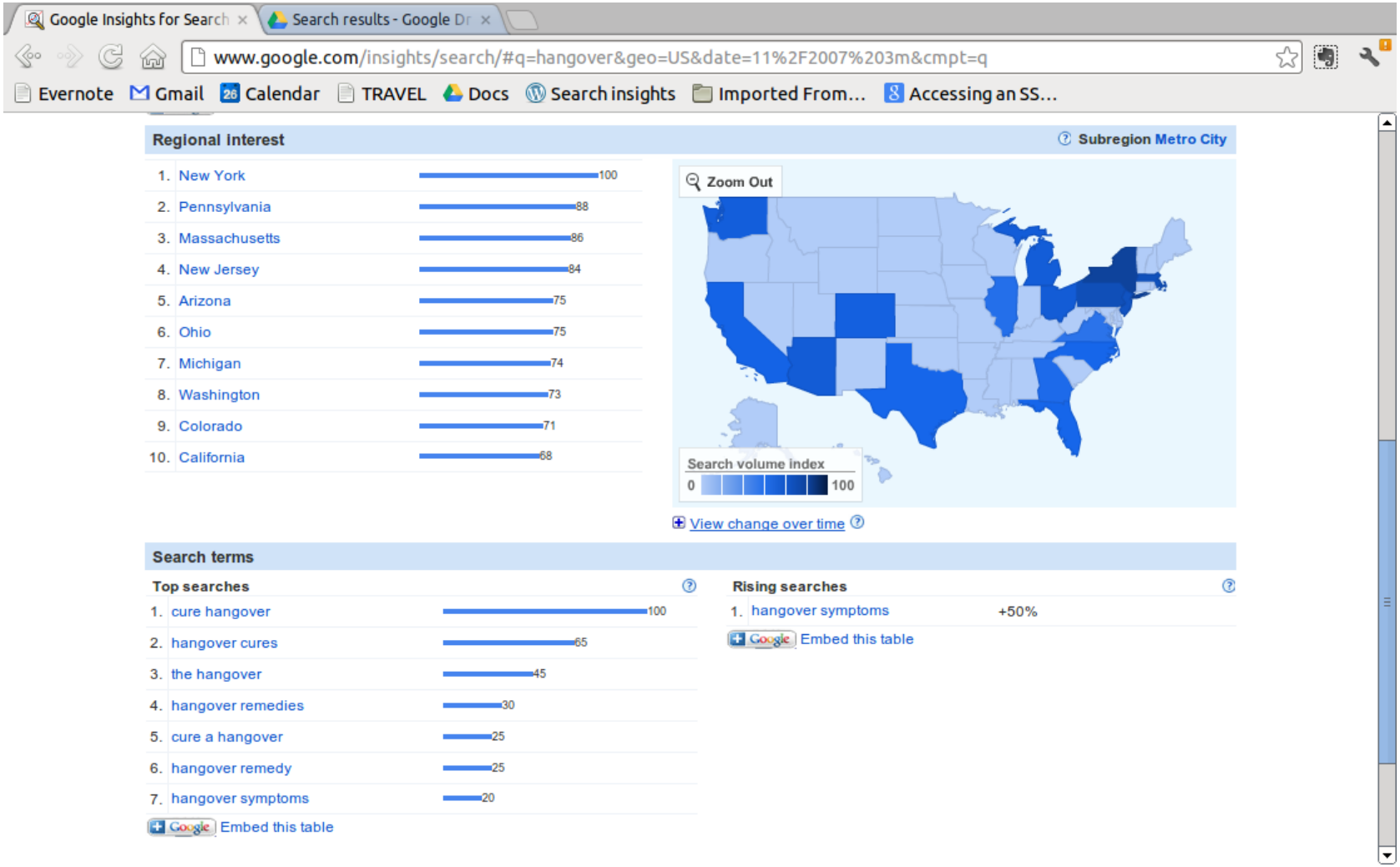
Which day of the week are there the most searches for 'hangover'?

1. Sunday
2. Monday
3. Tuesday
4. Wednesday
5. Thursday
6. Friday
7. Saturday

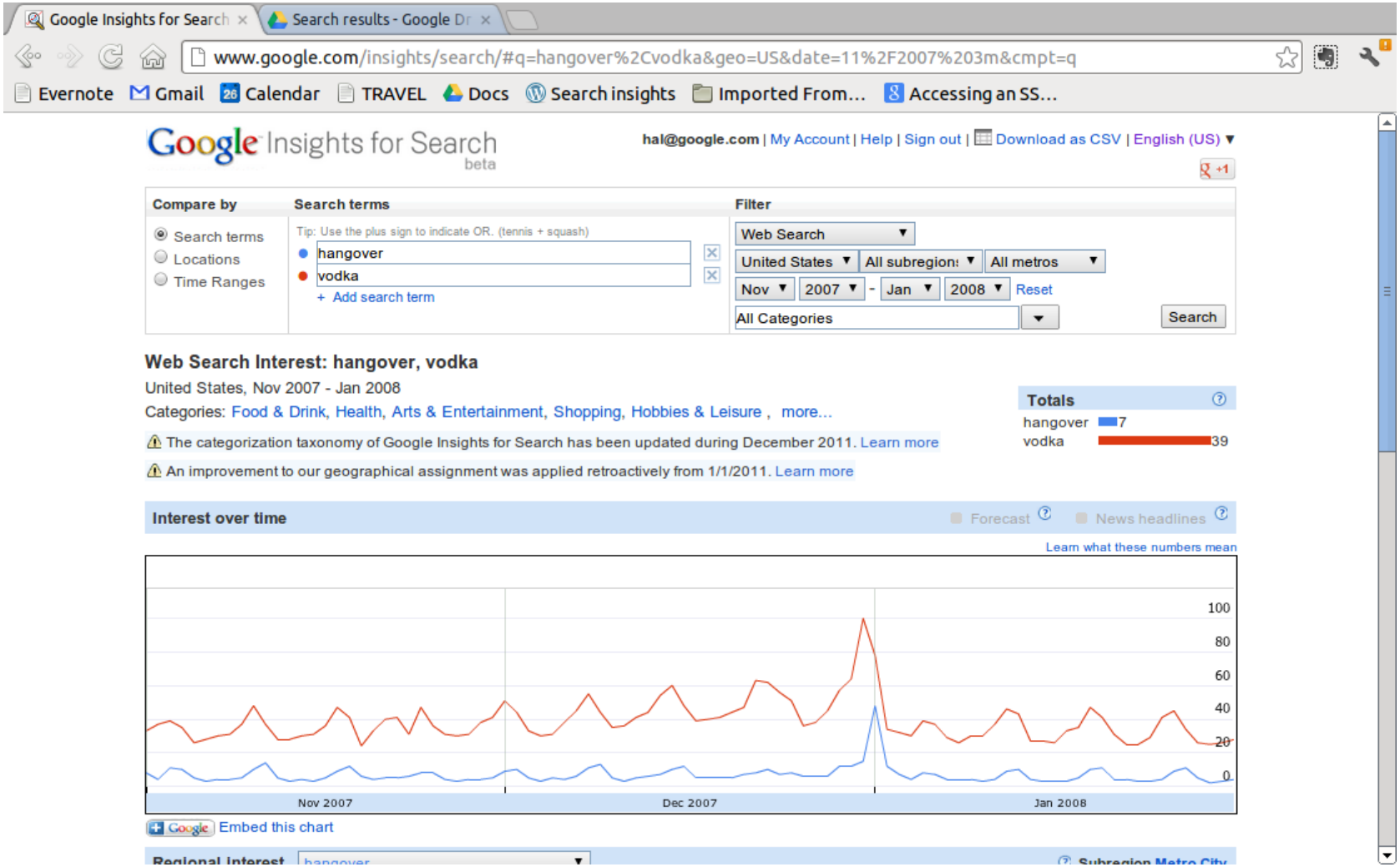
Google Insights for Search



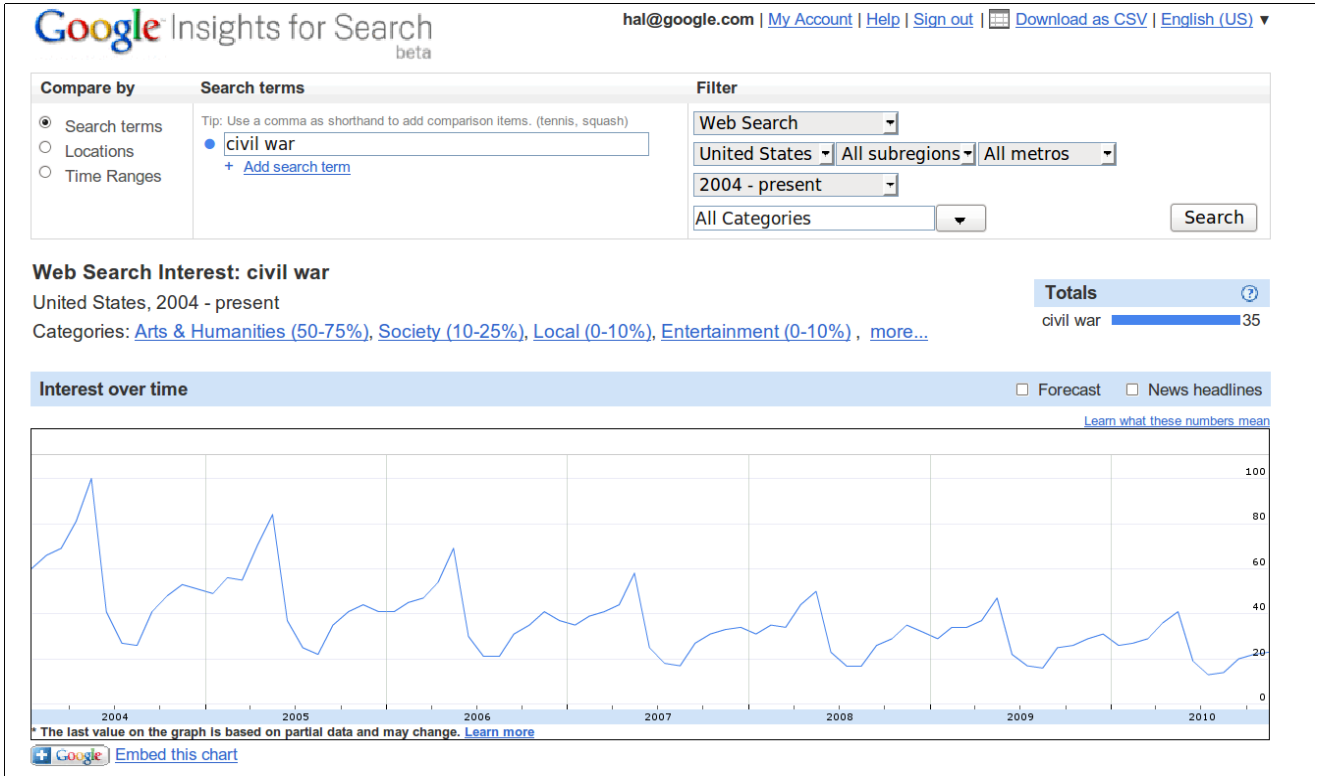
Google Insights for Search



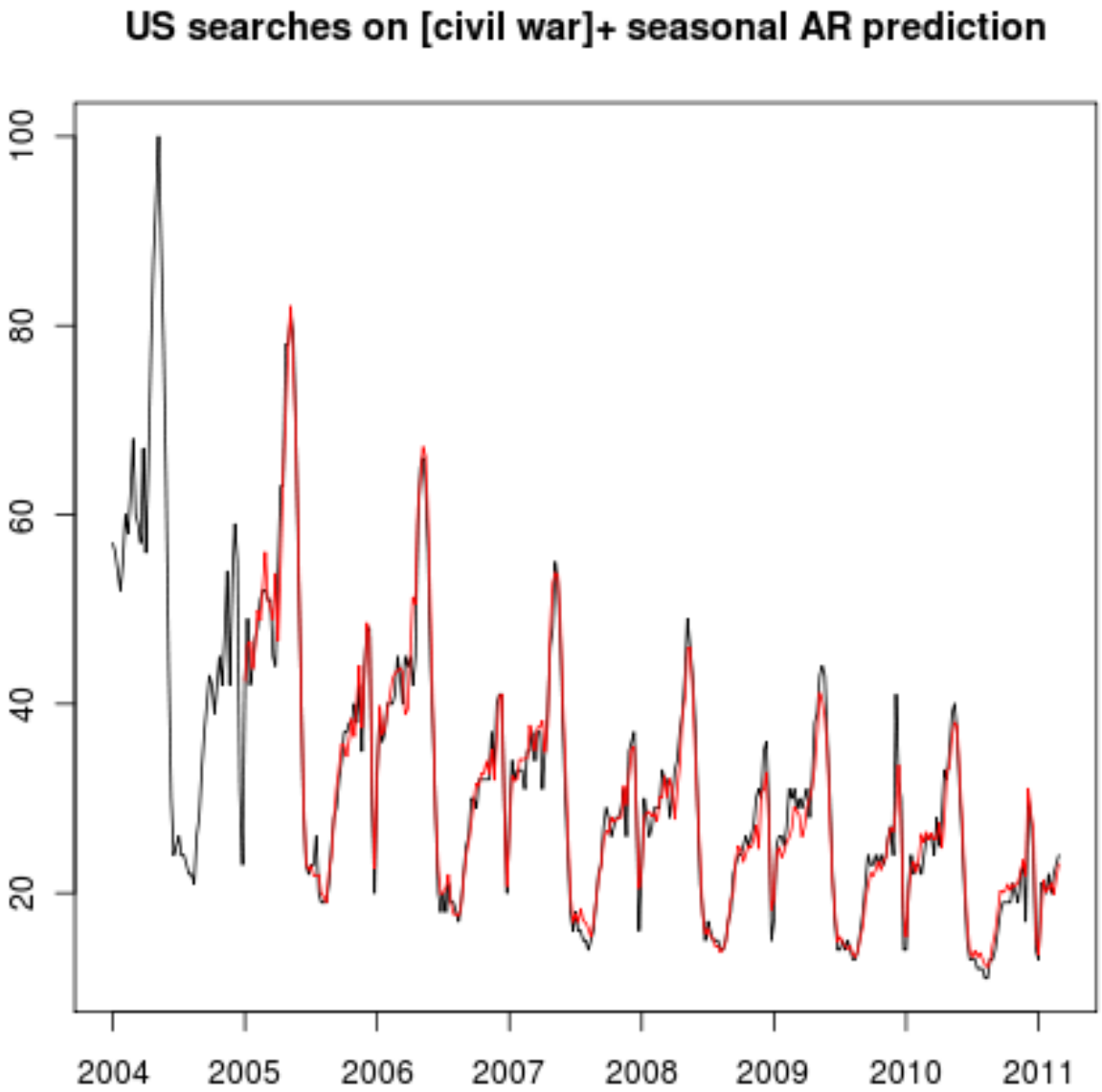
Google Insights for Search



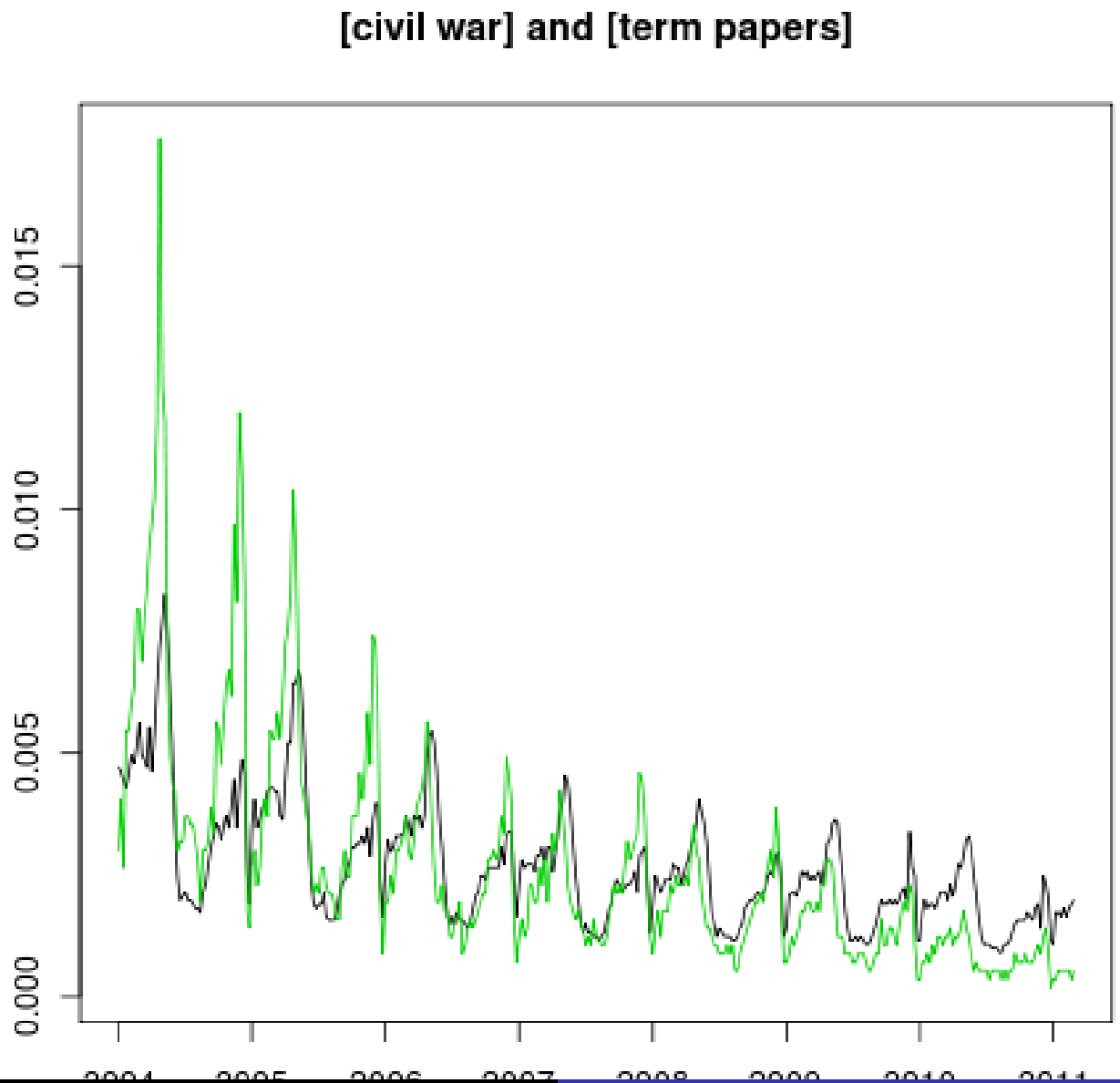
Searches for 'civil war'



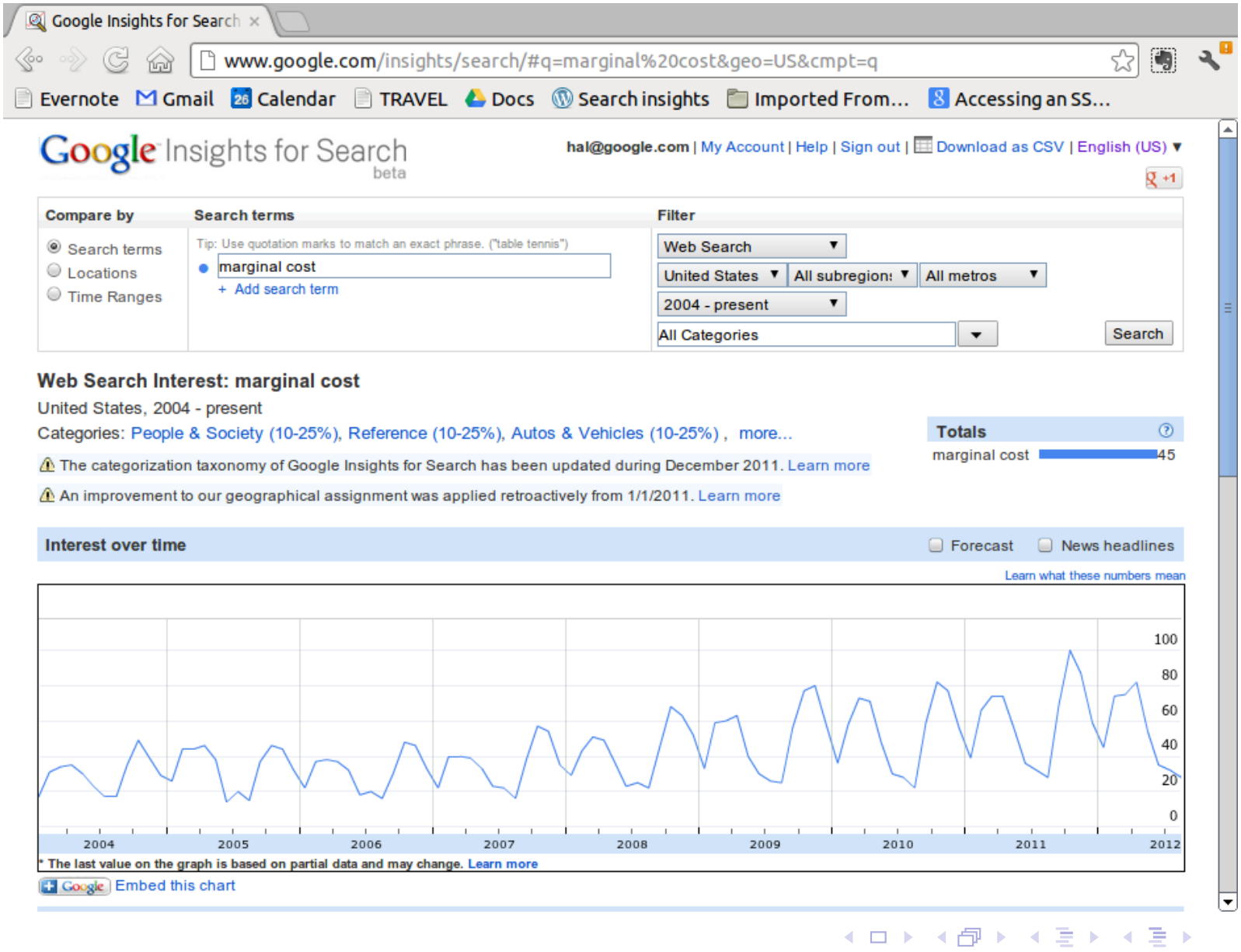
Predicting searches for 'civil war'



Searches for 'civil war' and 'term papers'



Searches for 'marginal cost'

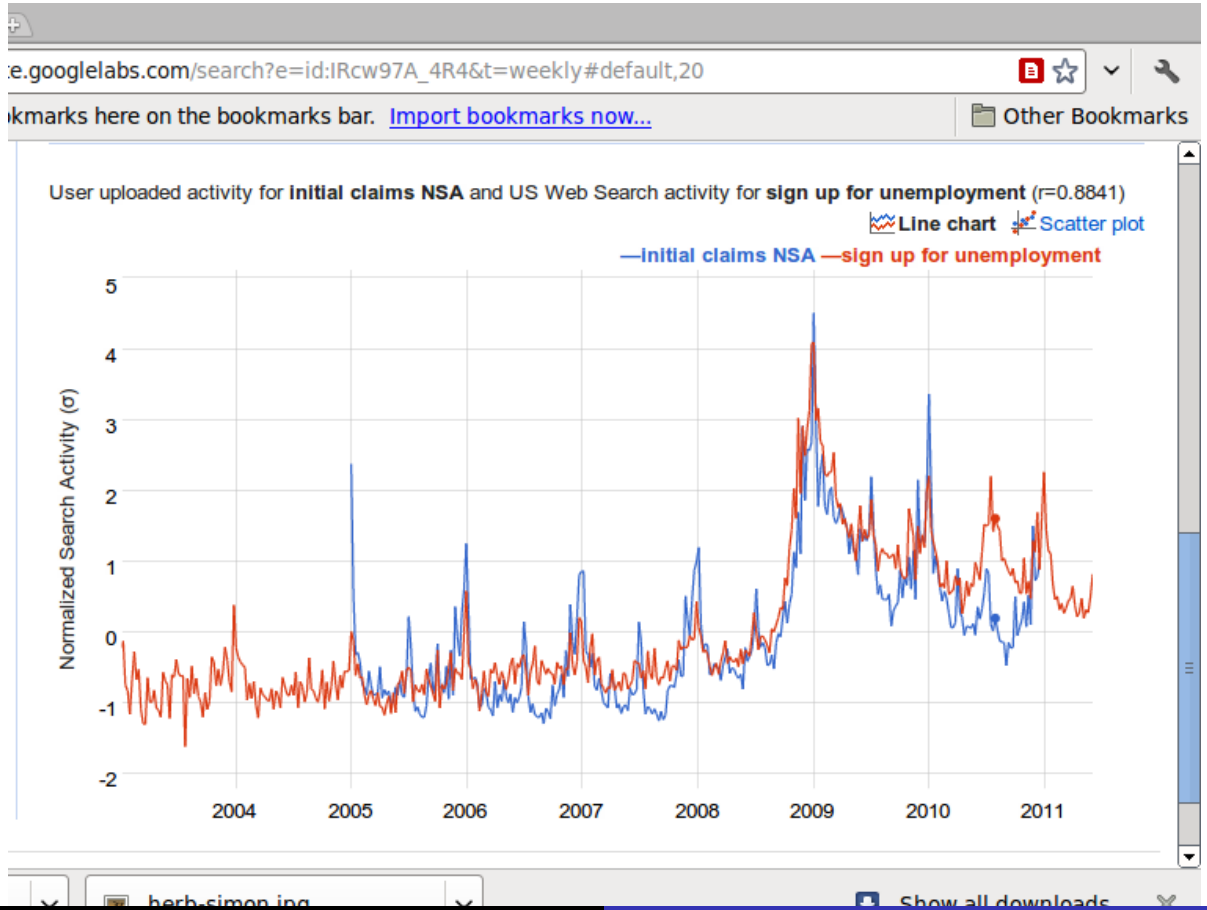


Problem motivation

- Want to use Google "Insights for Search" data to nowcast economic series
  - unemployment may be related to job search queries
  - auto purchases may be related to "vehicle shopping" queries
  - Even contemporaneous relationship is useful due to reporting lag
- Fat regression problem: there are many more predictors than observations
- Millions of queries, hundreds of categories
  - number of observations ~ 100 for monthly economic data
  - number of predictors ~ 150 for "economic" categories in I4S
- How do we choose which variables to include?

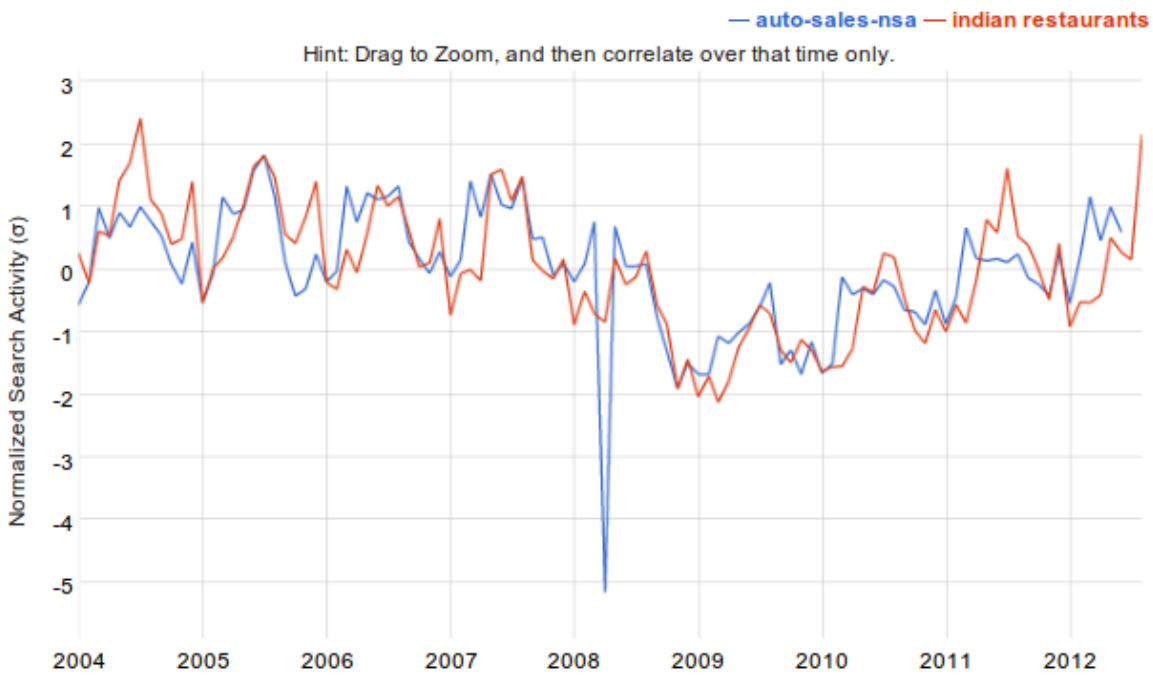
Example: unemployment

- Sometimes simple correlation works
- Google Correlate tool
- Initial claims for unemployment benefits (NSA)
- Query "sign up for unemployment"



Problems with simple correlation

- Need a regression, not a list of correlates
- Common trend or seasonality can lead to "spurious regression"
- Need to "whiten" the series by removing seasonality and trend



Approaches to variable selection

- Human judgment
- Significance testing (forward and backward stepwise regression)
- Information criteria (AIC, BIC)
- Principal components and factor models
- Lasso, ridge regression, penalized regression models

Our approach

- Original approach
  - forecast y_t using its own past values and human-chosen contemporaneous regressors from I4S
  - non-seasonal AR1: $y_t = a_1 y_{t-1} + b x_t + e_t$
  - seasonal AR1: $y_t = a_1 y_{t-1} + a_{12} y_{t-12} + b x_t + e_t$
- Current approach
  - Local linear trend and Kalman filter for time series
  - Spike and slab regression for variable selection
  - Bayesian model averaging for final forecast

Basic structural model with regression

- Classic time series model with constant level, linear time trend, regression
  - $y_t = \mu + bt + \beta x_t + e_t$
- "Local linear trend" is a stochastic generalization of this
  - Observation: $y_t = \mu_t + z_t + e_{1t}$
  - State 1: $\mu_t = \mu_{t-1} + b_{t-1} + e_{2t}$
  - State 2: $b_t = b_{t-1} + e_{3t}$
  - State 3: $z_t = \beta x_t$
  - In simple case where $b_t = \beta = 0$
    - $e_{2t} = 0$ is constant mean model where best estimate is sample average
    - $e_{1t} = 0$ is random walk where best estimate is current value
- Parameters to estimate: regression coefficients β and variances of (e_{it}) for $i = 1, \ldots, 3$
- Use these variances to construct optimal Kalman forecast: $\hat{y}_t = y_{t-1} + \beta x_t + k_t(\text{variances}) \times \text{forecast error at } t-1$

Advantages of Kalman

- No problem with unit roots or other kinds of nonstationarity
- No problem with missing observations
- No problem with mixed frequency
- No differencing or identification stage (easy to automate)
- Nice Bayesian interpretation
- Easy to compute estimates (particularly in Bayesian case)
- Nice interpretation of structural components
- Adaptive estimates (good for recession)
- Good forecast performance

Spike and slab regression for variable choice

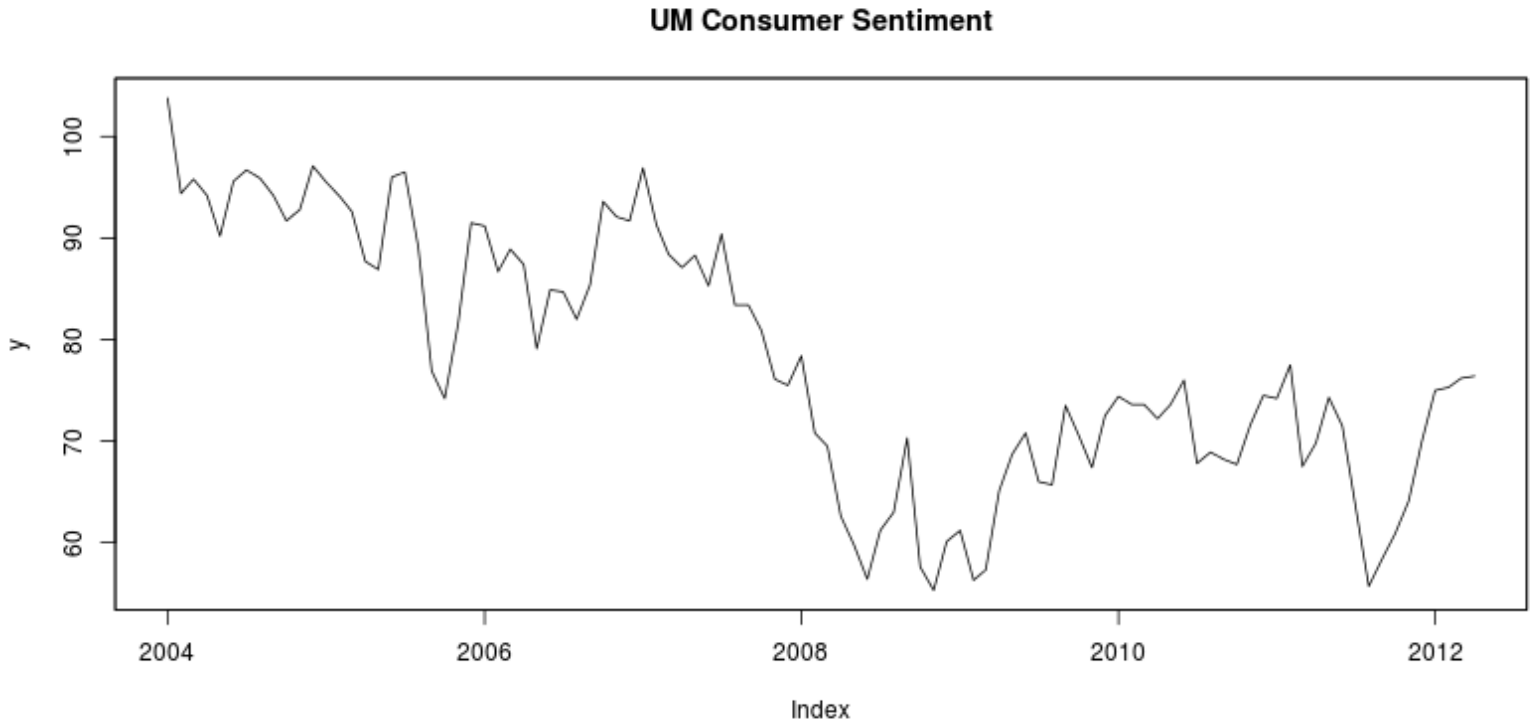
- Spike
  - Define vector γ that indicates variable inclusion
  - $\gamma_i = 1$ if variable i has non-zero coefficient in regression, 0 otherwise
  - Binomial prior distribution for γ
  - Can use an informative prior; e.g., expected number of predictors
- Slab
  - Conditional on being in regression ($\gamma_i = 1$) put a diffuse Normal prior on β_i
- Estimate posterior distribution of (γ, β) using MCMC

Bayesian model averaging

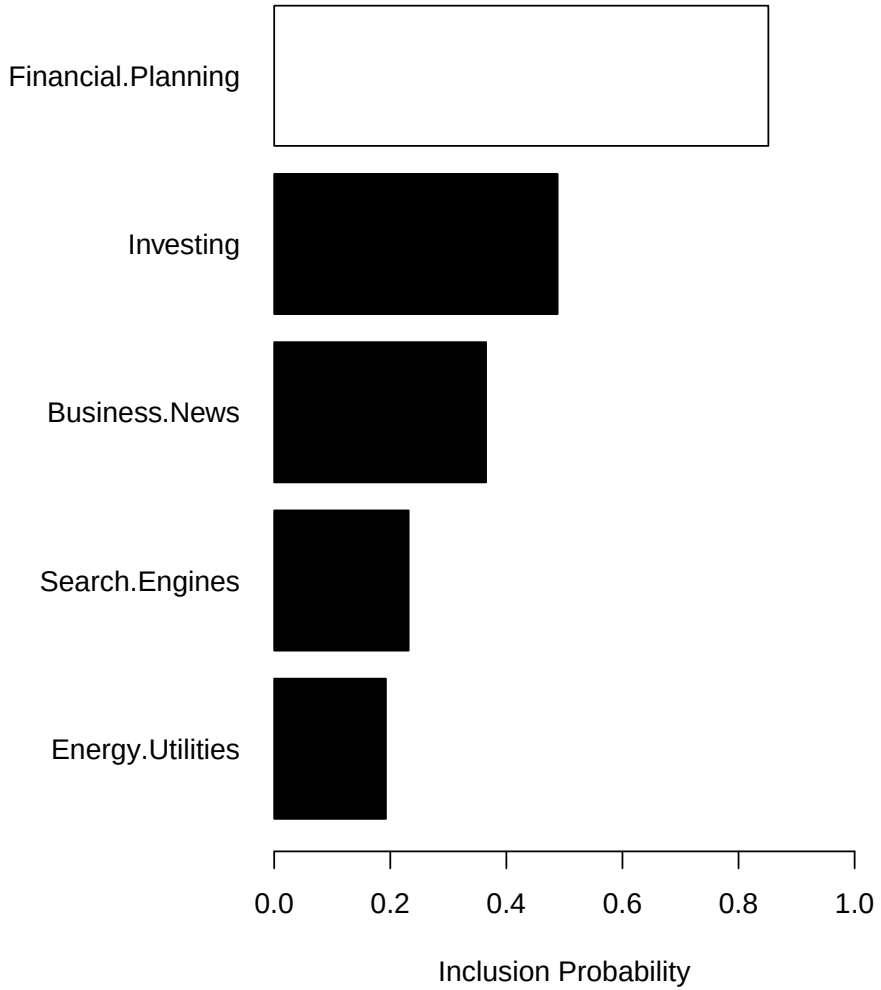
- We simulate draws from posterior using MCMC
- Each draw has a set of variables in the regression (γ) and a set of regression coefficients (β)
- Make a forecast of y_t using these coefficients
- Take average over all the forecasts for final prediction
- Take average over draws of γ to see which predictors have high probability of being in regression

Example: Consumer Sentiment

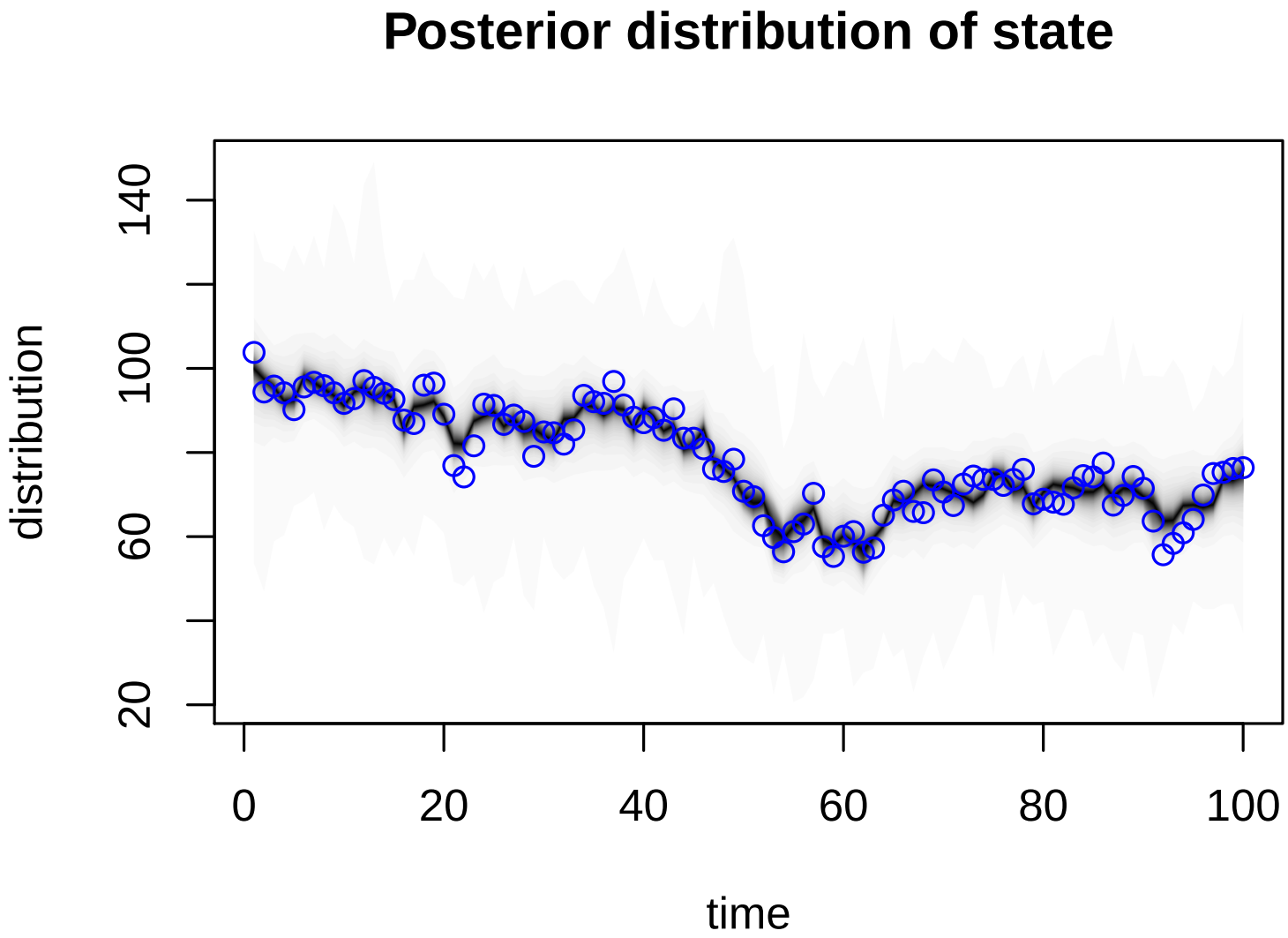
- Monthly UM Consumer sentiment from Jan 2004 to Apr 2012 ($n = 100$)
- Google Insights for Search categories related to economics ($k = 150$)
- No compelling intuition about what predictors should be



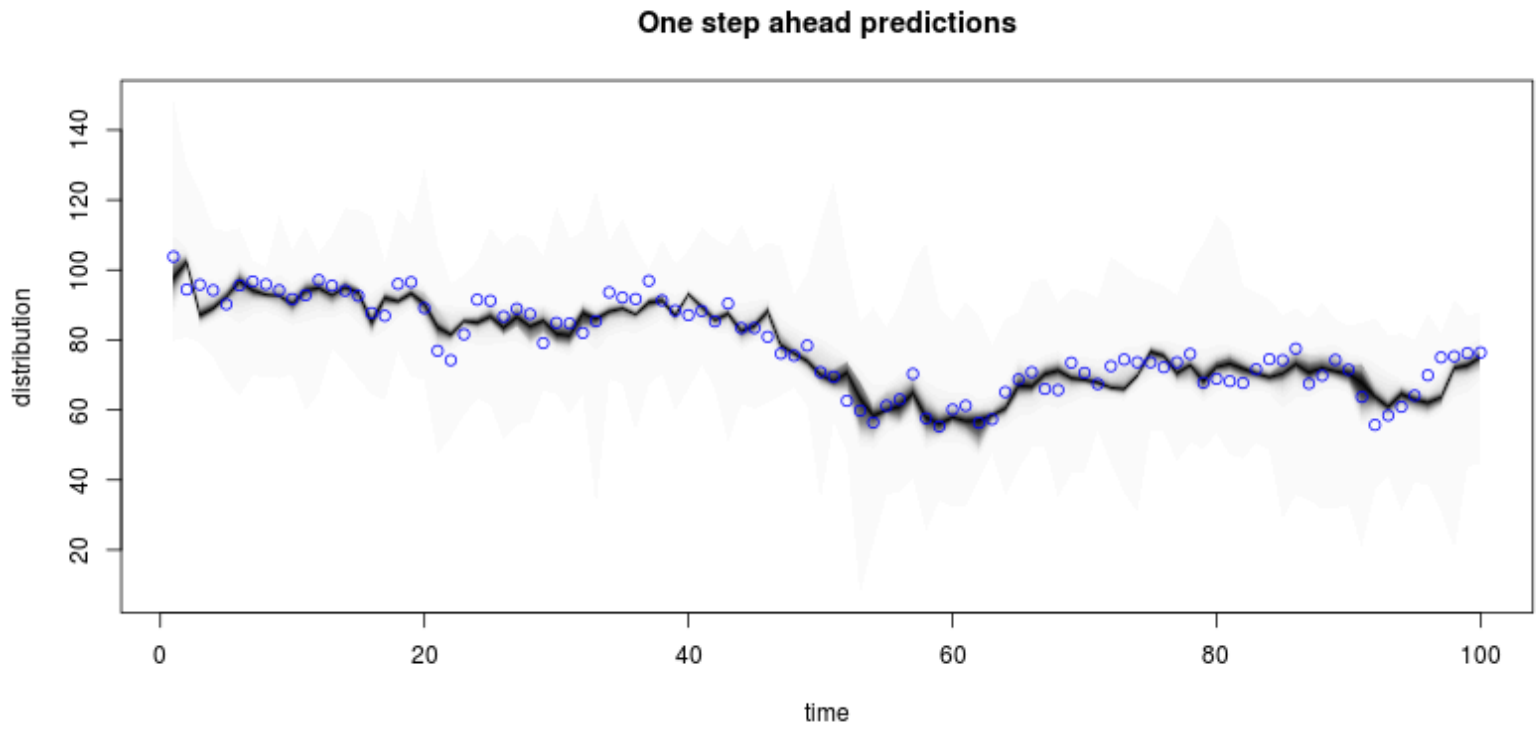
UM Consumer Sentiment



Posterior distribution of state and observations



Posterior distribution of one-step ahead forecast



Compare Actual, AR(1), reg and BSTS predictions

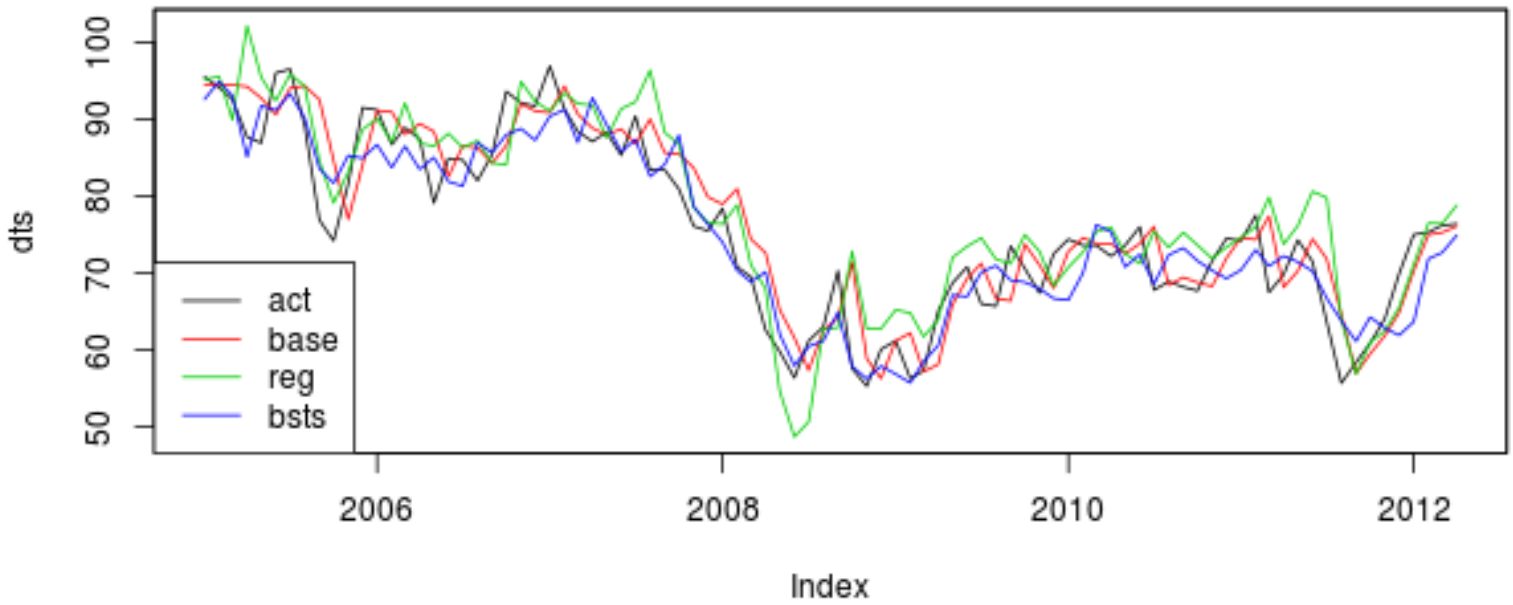


Figure: Actual, AR(1), regression, and BSTS one-step ahead predictions.

Decomposition of forecast into trend and regression

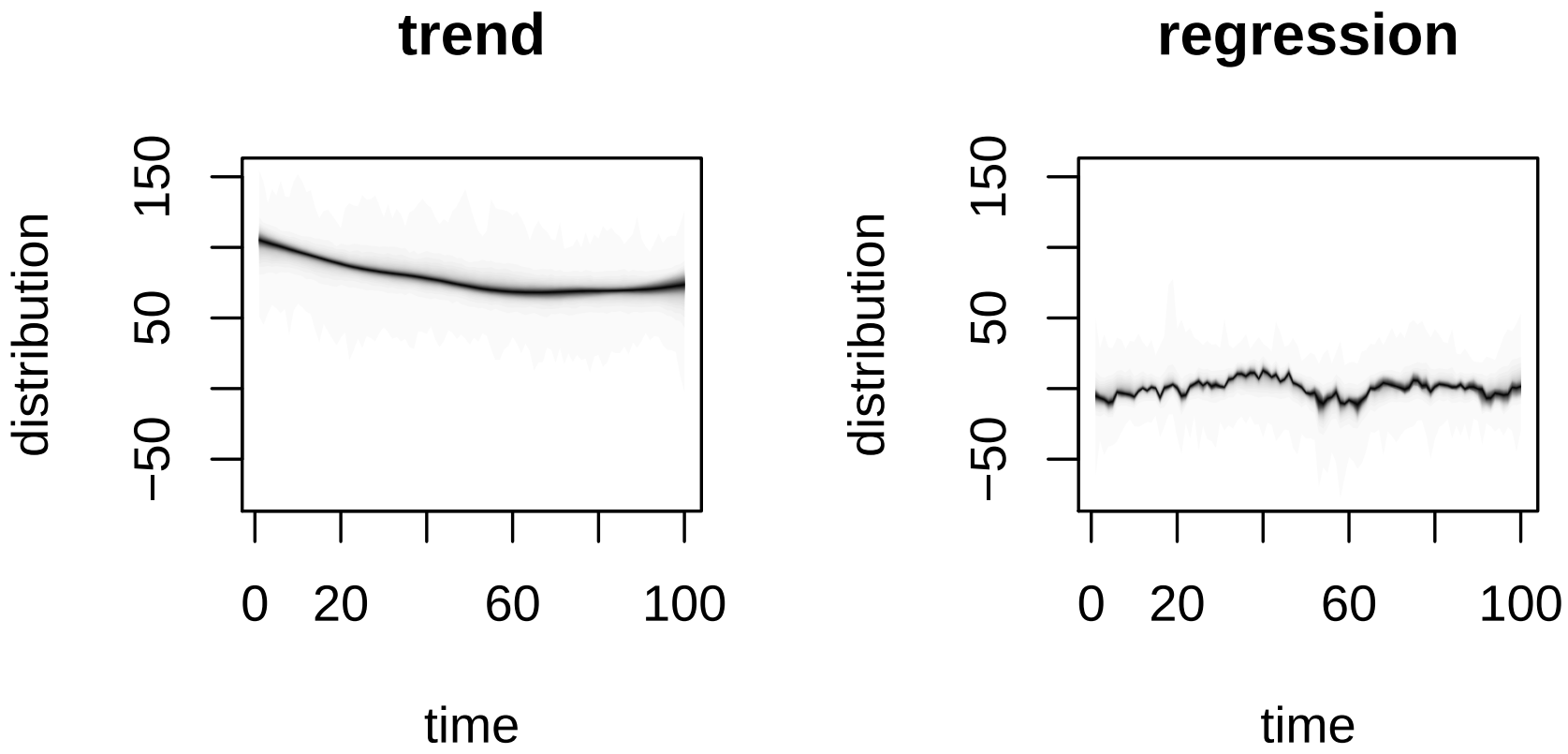


Figure: Decomposition of forecast.

Future work

- Seasonality — done
- Mixed frequency forecasting — almost done
- Fat tail distributions
- Parallel MCMC
- Automate the whole thing