Bayesian Variable Selection for Nowcasting Economic Time Series

Steve Scott
Hal Varian

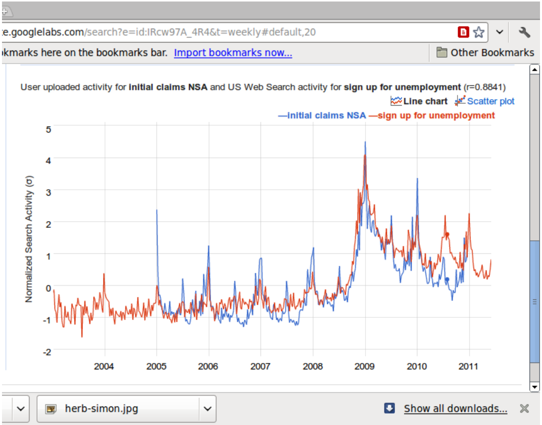
December 31, 2012

Problem motivation

- Want to use Google Trends data to nowcast economic series
  - unemployment may be predicted by "job search" queries
  - auto purchases may be predicted by "vehicle shopping" queries
- Fat regression problem: there are many more predictors than observations
- Millions of queries, hundreds of categories
  - number of observations $\sim$ 100 for monthly economic data
  - number of predictors $\sim$ 150 for "economic" categories in I4S
- How do we choose which variables to include?

Example: unemployment

- Sometimes Google Correlate works
- Load in: initial claims for unemployment benefits
- Get back 100 queries, including “sign up for unemployment”



e.googlelabs.com/search?e=id:IRcw97A_4R4&t=weekly#default,20

kmarks here on the bookmarks bar. Import bookmarks now...

User uploaded activity for **initial claims NSA** and US Web Search activity for **sign up for unemployment** (r=0.8841)

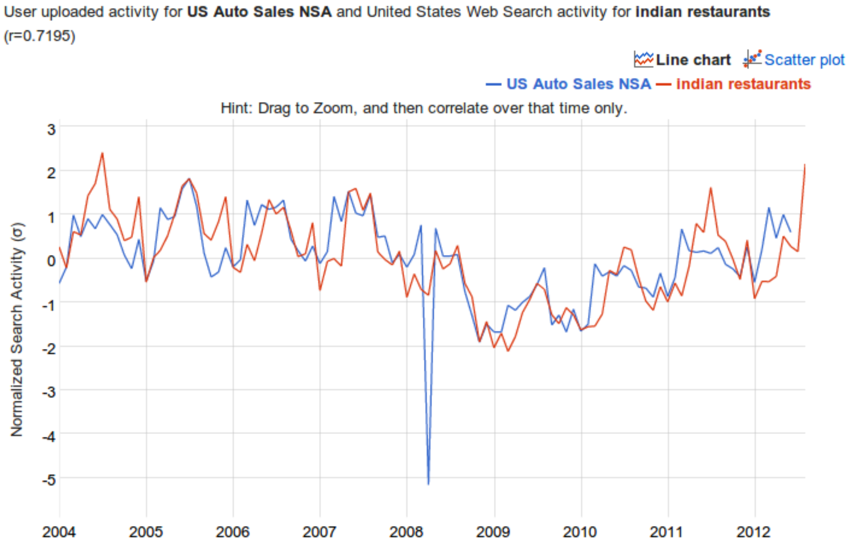
Build a simple AR model

- Use deseasonalized initial claims (y_t)
- Use deasonalized, detrended searches for "unemployment office" (x_t)

$$\begin{aligned} \text{base: } y_t &= a_0 + a_1 y_{t-1} + e_t \\ \text{regr: } y_t &= a_0 + a_1 y_{t-1} + b x_t + e_t \end{aligned}$$

- Estimate using rolling window
- One-step-ahead MAE during recession is about 8.7% lower when "unemployment office" query is included

But sometimes simple correlation doesn't work



User uploaded activity for **US Auto Sales NSA** and United States Web Search activity for **indian restaurants** (r=0.7195)

Hint: Drag to Zoom, and then correlate over that time only.

Avoid spurious regression

- How to control for trend and seasonality?
  - Build a model for the *predictable* part of time series ("whiten the series")
  - Find regressors that predict the *residuals*
- How to choose regressors?
  - Simple correlation is too limited
  - Human judgment doesn't scale

Approaches to variable selection

- Human judgment
- Significance testing (forward and backward stepwise regression)
- Information criteria (AIC, BIC)
- Principle component, partial least squares and factor models
- Lasso, ridge regression, penalized regression models

Our approach

- Original approach (simple autoregression)
  - forecast y_t using its own past values and human-chosen contemporaneous regressors from Google Trends
  - non-seasonal AR1: $y_t = a_1 y_{t-1} + bx_t + e_t$
  - seasonal AR1: $y_t = a_1 y_{t-1} + a_{12} y_{t-12} + bx_t + e_t$
- Current approach (Bayesian Structural Time Series)
  - Use Kalman filter to whiten 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, and regressors
  - $y_t = \mu + bt + \beta x_t + e_t$
- "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: $z_t = \beta x_t$
- Parameters to estimate: regression coefficients β and variances of (e_{it}) for $i = 1, \ldots, 2$
- Use these variances to construct optimal Kalman forecast: $\hat{y}_t = y_{t-1} + \beta x_t + k_t(\text{variances}) \times$ forecast error at $t - 1$

Intuition for Kalman filter

- Consider simple case without regressors and trend
  - Observation equation: $y_t = \mu_t + e_{1t}$
  - State equation: $\mu_t = \mu_{t-1} + e_{2t}$
- Two extreme cases
  - $e_{2t} = 0$ is constant mean model where best estimate is sample average up to t
  - $e_{1t} = 0$ is random walk where best estimate is current value
- In general, optimal forecast will be weighted average of past observations and current observation
- Weights depend on variances of the two error terms

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
- Easy to add seasonality
- Good forecast performance

Spike and slab regression for variable choice

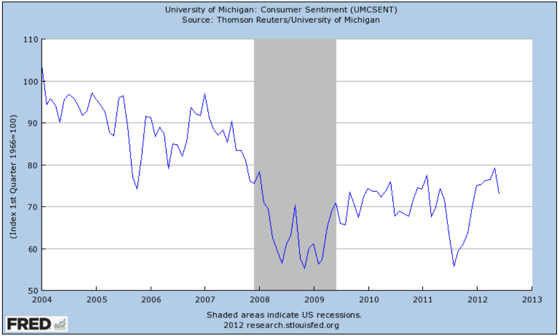
- Spike
  - Define vector γ that indicates variable inclusion
  - $\gamma_i = 1$ if variable i has non-zero coefficient in regression, 0 otherwise
  - Binomial prior distribution, $p(\gamma)$, for γ
  - Can use an informative prior; e.g., expected number of predictors
- Slab
  - Conditional on being in regression ($\gamma_i = 1$) put a (diffuse) prior on β_i, $p(\beta|\gamma)$.
- 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
- This gives the posterior forecast distribution
- Can take average over all the forecasts for final prediction
- Can take average over draws of γ to see which predictors have high probability of being in regression

Example 1: Consumer Sentiment

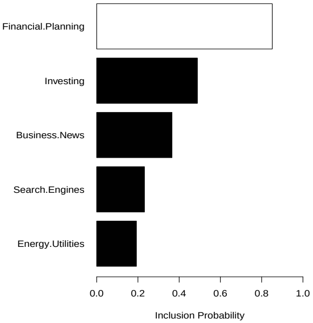
- Monthly UM Consumer sentiment from Jan 2004 to Apr 2012 ($n = 100$)
- Google Insights for Search categories related to economics ($k = 150$)
- No compelling intuition about what predictors should be



Variable selection

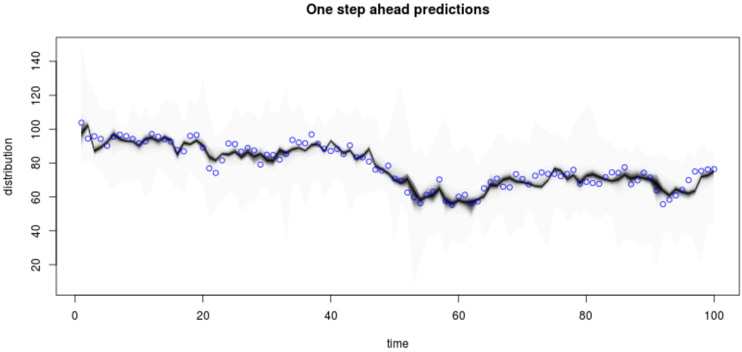
- Google Insights for Search categories related to economics ($k = 150$)
- Deseasonalize predictors using R command `stl`
- Detrend predictors using simple linear regression
- Let `bsts` choose predictors

UM Consumer Sentiment Predictors

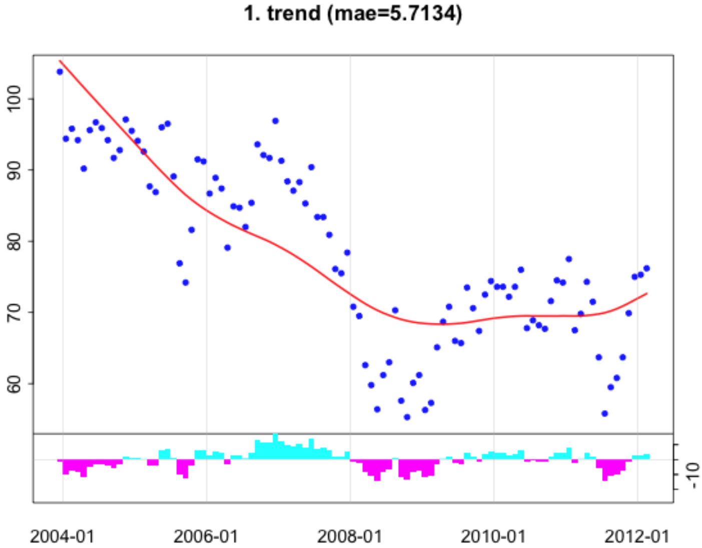


- Financial planning: schwab, 401k, ira, smith barney, fidelity, roth ira
- Investing: stock, gold, fidelity, stocks, silver, stock market, gold price, scottrade

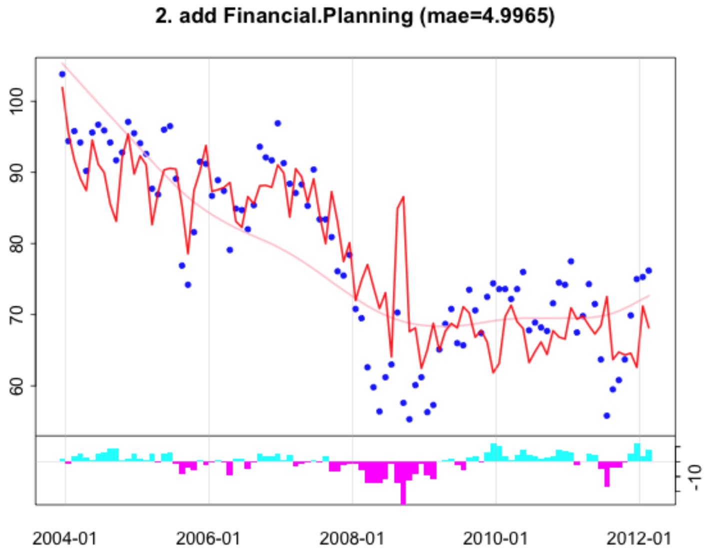
Posterior distribution of one-step ahead forecast



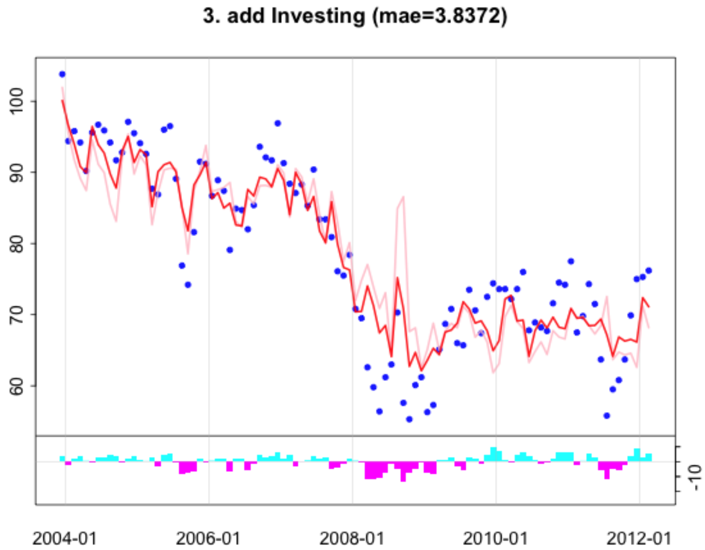
Start with Kalman trend



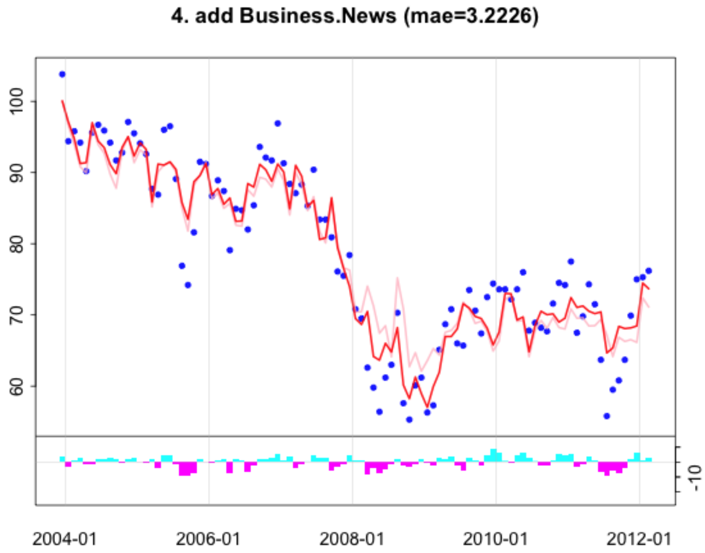
add Financial Planning



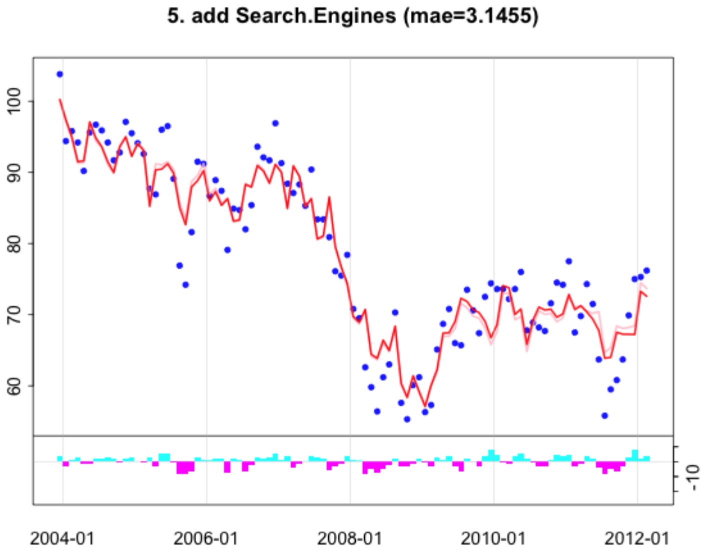
add Investing



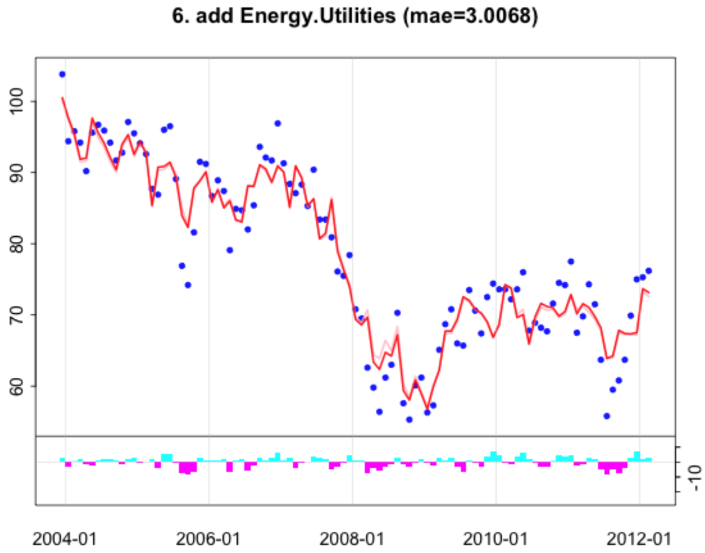
add Business News



add Search Engines



add Energy and Utilities



Fun with priors

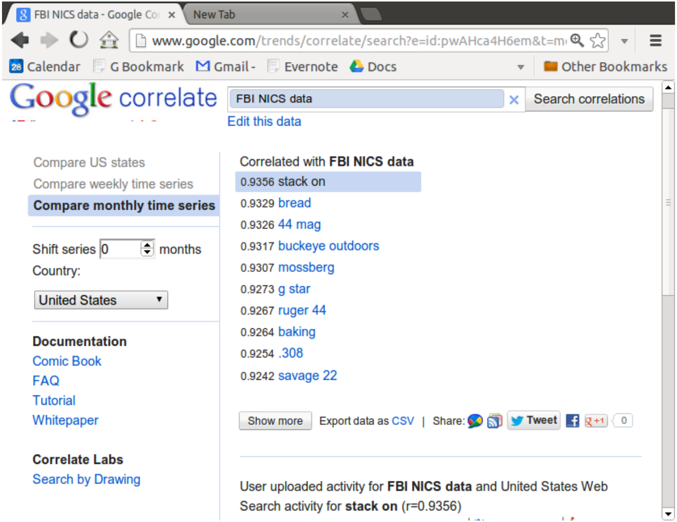
- Can use prior to influence variable choice in regression
  - Give higher weight to certain verticals
  - Influence the expected number of variables in regression
- Can use prior to improve estimate of trend component
  - Google data starts in 2004, only one recession
  - Can estimate parameters of trend model with no regressors
  - Use this as prior for estimate of trend in estimation period

Example of informative prior for trends

- UM Consumer Sentiment starting Jan 1996
- Google data starting Jan 2004
- Estimate variances for Kalman filter using data up to Jan 2004
- Use these parameters as informative prior for subsequent data
- Tends to give more weight to regressors

Example 2: gun sales

Use FBI's National Instant Criminal Background Check



FBI NICS data

Shift series 0 months

Country: United States

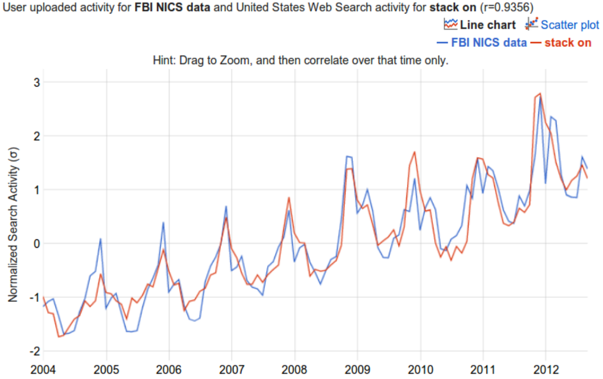
Correlated with **FBI NICS data**

0.9356	stack on
0.9329	bread
0.9326	44 mag
0.9317	buckeye outdoors
0.9307	mossberg
0.9273	g star
0.9267	ruger 44
0.9264	baking
0.9254	.308
0.9242	savage 22

User uploaded activity for **FBI NICS data** and United States Web Search activity for **stack on** (r=0.9356)

Google Correlate Results

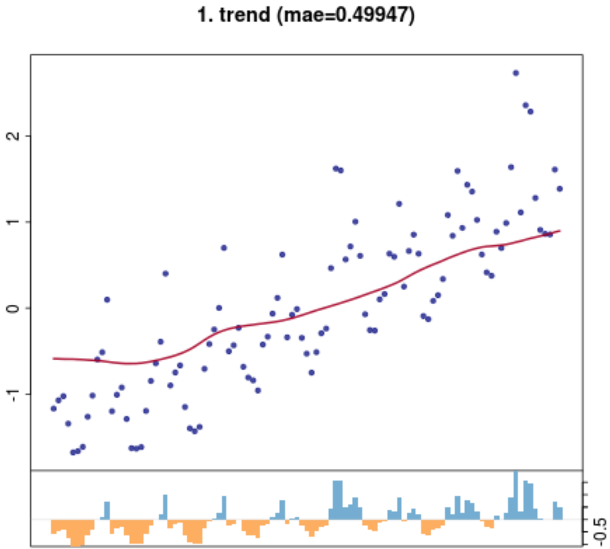
- [stack on] has highest correlation
- [gun shops] is chosen by bsts



User uploaded activity for **FBI NICS data** and United States Web Search activity for **stack on** (r=0.9356)

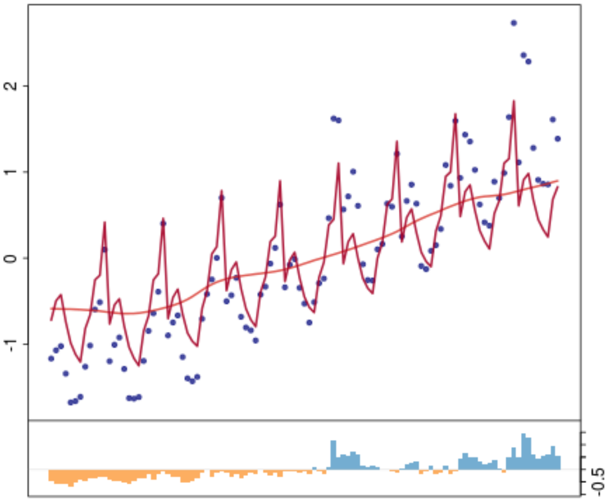
Hint: Drag to Zoom, and then correlate over that time only.

Trend

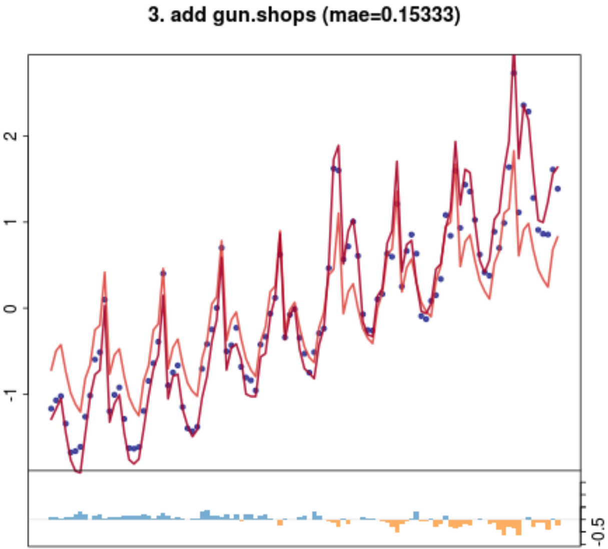


Seasonal

2. add seasonal (mae=0.33654)



Gun Shops



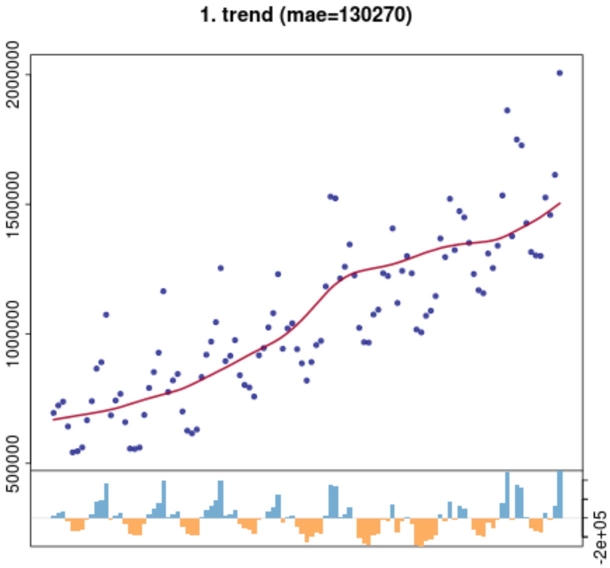
Google Trends predictors

- 586 Google Trends verticals, deseasonalized and detrended
- 107 monthly observations

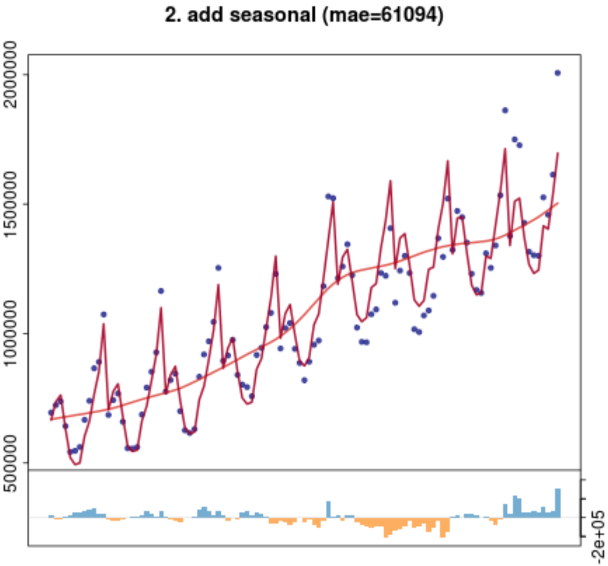
Category	mean	inc.prob
Recreation::Outdoors::Hunting:and:Shooting	1,056,208	0.97
Travel::Adventure:Travel	-84,467	0.09

Table : Google Trends predictors for NICS checks.

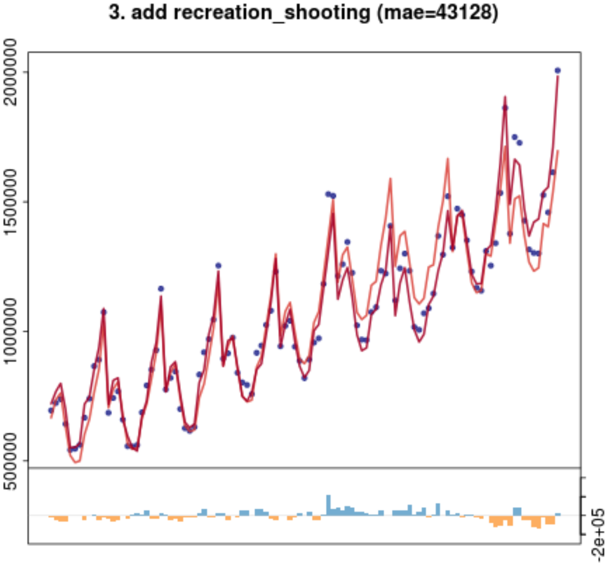
Trend



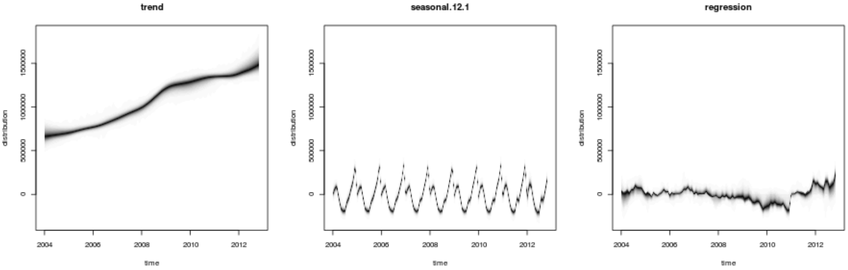
Seasonal


2. add seasonal (mae=61094)
500000
1000000
1500000
2000000
-2e+05

Hunting and Shooting



State decomposition



Future work

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