Predicting the Present with the Google Trends

Forcasting Flu Season

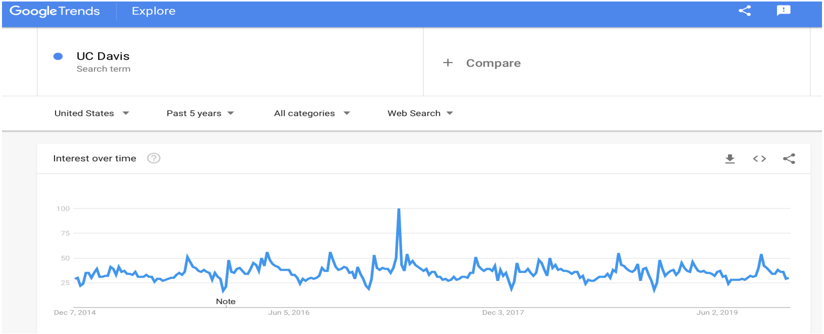
Kenneth Broadhead   Xuezhen Li

UC Davis

December 6, 2019

Introduction to Google Trends

Google Trends



UC Davis
Search term

United States ▾ Past 5 years ▾ All categories ▾ Web Search ▾

- Google Trends provides a time series index of the volume of queries users enter into Google in a given geographic area.

Google Trends

- The maximum query share in the time period specified is normalized to be 100.
- Google classifies search queries into about 30 categories at the top level and about 250 categories at the second level using a natural language classification engine.

Predicting the Present?

- Economic (time series) data often has a significant reporting lag of up to several weeks.
- Similarly, weekly CDC (time series) data on national flu activity takes time to compile and publish.
- It would be nice if researchers had more timely access to these reports.

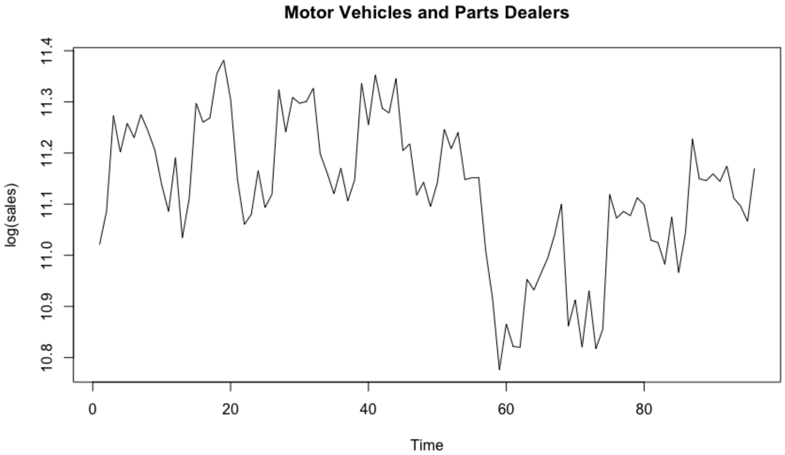
Predicting the Present?

- Google Trends data is available in almost real time.
- Perhaps we can forecast the present, but unknown, values of our time series using current Google Trends data.

Examples

Economic Data Example

- Sales in Motor Vehicles and Parts Dealers (2004-2011):



Economic Data Example

- We fit a simple seasonal AR(1) model to the log transformed data y_t.

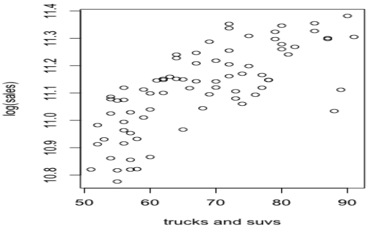
$$\text{Base Model: } y_t = b_0 + b_1 y_{t-1} + b_{12} y_{t-12} + e_t$$

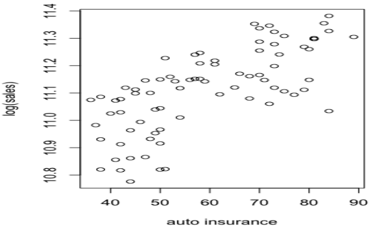
- We add two automotive-related categories from the Google trends. One is Trucks and SUVs, the other one is Auto Insurance.

$$\text{Trend Model: } y_t = b_0 + b_1 y_{t-1} + b_{12} y_{t-12} + a_1 g_{1,t} + a_2 g_{2,t} + e_t$$

Economic Data Example

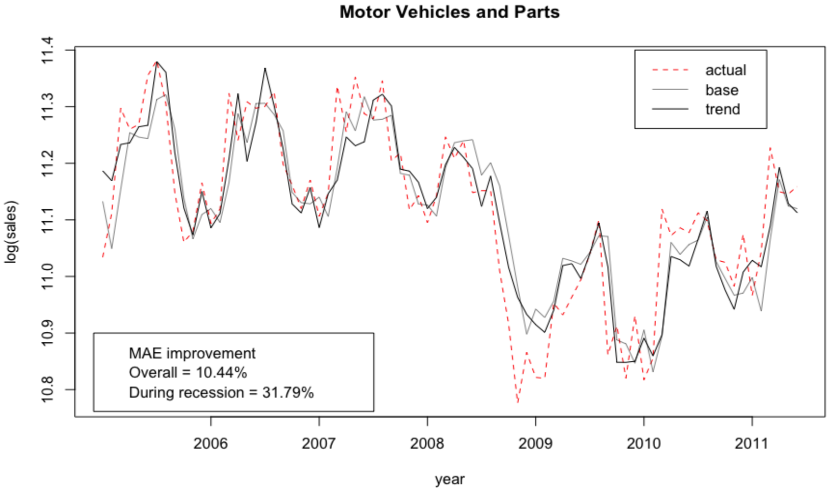
- Examine Correlations:





- R^2 increases from 0.7211 to 0.7821.
- Whether the Trends variables improve out-of-sample forecasting or not?
  - Overall MAE decrease from 6.38% to 5.71%.
  - During recession MAE decrease from 7.61% to 5.19%.

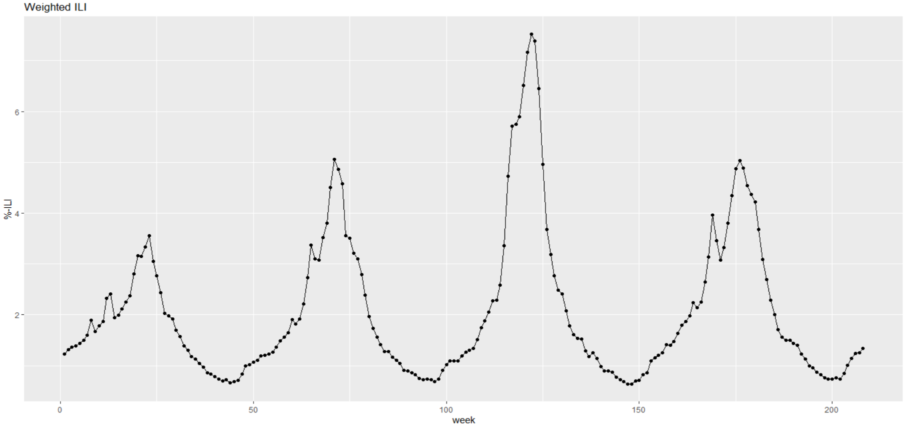
Economic Data Example



MAE improvement
Overall = 10.44%
During recession = 31.79%

CDC Flu Data Example

- Data: weekly CDC ILINet (a proxy for flu activity) data for previous 4 seasons.



CDC Flu Data Example

- We fit a simple seasonal AR(1) model to the inverse transformed data, x_t.

$$\text{Model 1: } x_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 x_{t-52}$$

- We then fit a model incorporating google trends data from the week we wish to forecast. Some experimenting suggests that google search data for "flu shot" provides the greatest improvement to in sample fit.

$$\text{Model 2: } x_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 x_{t-52} + \beta_3 g_t$$

- We then compare predictive power of each model using rolling window forecasts, and comparing the resulting MAEs for each model.

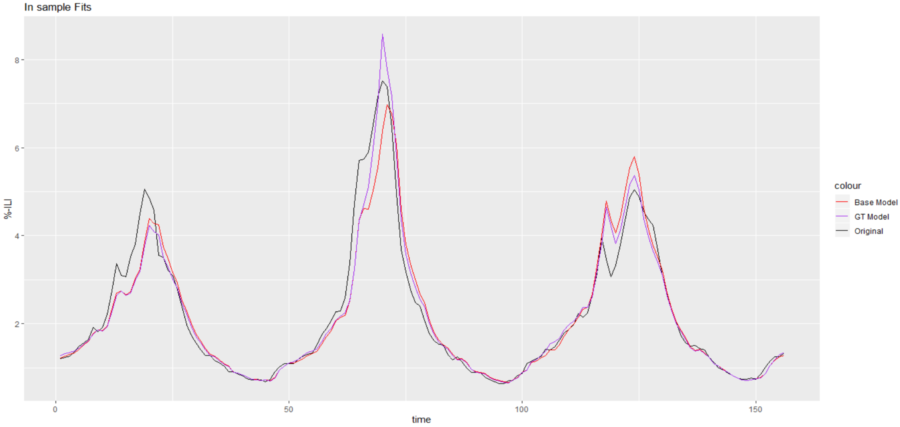
CDC Flu Data Results

- Model 1 had a good fit ($R_a^2 = 0.9773$).
- Model 2 didn't greatly improve fit, ($R_a^2 = 0.9797$).
- We saw an 8.9 percent improvement in MAE when using the GT data.

Conclusions

Final Thoughts and Limitations

- Final in sample forecasts for each model:



Final Thoughts and Limitations

- Potential relative Google trends have to be chosen carefully by prior experience.
- Forecasting the present is usually only useful if:
  - One's data set has a reporting delay;
  - Google searches are indicative of the activity of the mechanism driving the underlying Stochastic process.
- It may also be unhelpful if a base model provides a very good initial fit.
- However, even if overall fit isn't improved, forecasting accuracy of specific aspects of data may still be improved.