Group 5 Final Report

Jiayi Lily Ma (jm4303), Lina Tian (yt2511), and Wen Ding (wd2288)

5/10/2019

Part 1: Introduction

1.1 Paper title:

Predicting the Present with Google Trends, by Hyunyoung Choi, Hal Varian, published Dec. 18, 2011 in The Economic Record, 2012.

1.2 Their motivation:

When releasing reports for certain economic activities, Government agencies tend to release them couple weeks late. In order to get a sense of the real-time economic condition, for exmple, the sales status of motor vehicle retail for May/2019 when still in May/2019, we normally look to the previous month and/or May/2018. Although overall, this method tends to be good in predicting what the current month's sale would look like (i.e. predicting the present), the authors of this paper propose that adding Google Trends data of related categories might help with the accuracy of "predicting the present."

Here, in our report, we choose to replicate one of their examples–"Motor vehicles and parts." They want to use the time period of 2004-2011, and the US Census Bureau tend to give report two weeks after the month of interest. They extracted weekly data from Google Trends and chose the first week of each month to represent the month so that they can get a 4-6 weeks lead in gettting the present sales.

1.3 Their model and method:

The models they used:

- base model reg0 is $y = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12}$
- Trend model reg1 is $y = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-12} + \beta_3 suvs + \beta_4 insurance$
  The base model is only using time series predictors, while the trend model adds Google Trends

The out of sample forcasting method they used:

- Rolling window forecast: Starting from the 18th month in the data, they use all the previous months to train a model and predict the next month.
- function to use in oosf.R: OutOfSampleForecast12()

Evalution method:

- MAE(mean absolute error)

1.4 Their Data source

- Google Trends weekly data

- Data from government or other official resources (for example, U.S. Census Bureau, US Department of Labor)

1.5 Their code

See `authors_original_code` for the original code.

1.5 Their Conclusion

"We have found that simple seasonal AR models that include relevant Google Trends variables tend to outperform models that exclude these predictors by 5% to 20%."

1.6 Code chunck to read in all the data

```r
dat <- read.csv("data/merged.csv") # paper data
merged0408 <- read.csv("data/merged_04_08.csv") # our 2004-2008 data
merged0913 <- read.csv("data/merged_09_13.csv") # our 2009-2013 data
merged1419 <- read.csv("data/merged_14_19.csv") # our 2014-2019 data
merged0408_v2 <- read.csv("data/merged_04_08_v2.csv") # our second version 2004-2008 data
```

1.7 Guide to their code and our original code

```r
source("oosf_new.R")
source("getEvaluation.R")
```

Their code:

- `oosf.R` is a helper file containing functions for out of sample forecast and producing MAE reports. There are functions named "OutOfSampleForecast12" and "MaeReport".
- `autos.St` has code for in-sample forecast (their models).
- `for-plots.St` has code for graphing their graphs.

Our code: Because the models that the paper implements are very simple, and we would come up with the mostly the same code for these regressions, we decided to write the following code:

- `oosf_new.R` is our new helper file containing more out of sample forecast functions that are based on our extended models. For example, OutOfSampleForecast0112_mw is a function for moving window forecast that we wrote based on their OutOfSampleForecast12 code.
- `getEvaluation.R` contains helper functions that take any data frame (authors or ours) and produce the lm object, MAE report, and graph altogether. There are functions for different out of sample forecast methods.

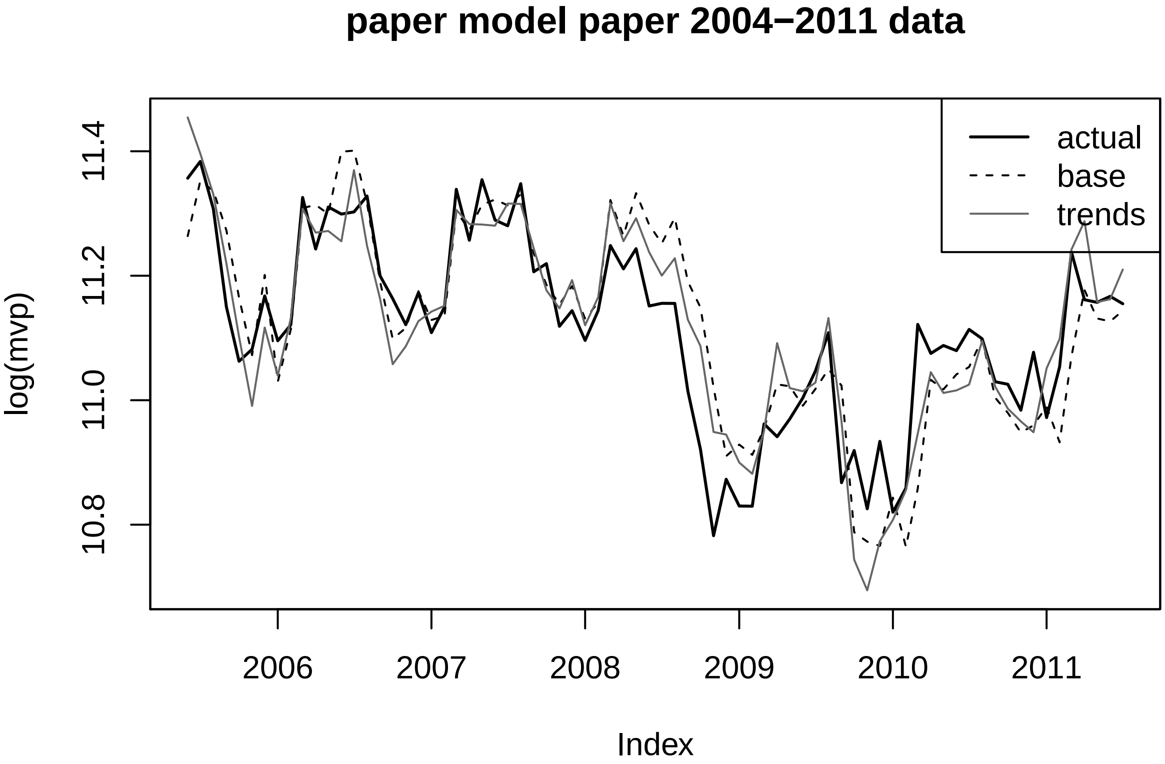
Part 2: Reproduction using paper's model and paper's data

Result 1:

For in sample forecasting, the paper sees an improvement in adjusted R-squared. For out of sample forecast, the paper finds out that "The MAE of the response variable $log(y_t)$ using the baseline AR-1 model is 6.34% while the MAE using the Trends data is 5.66%, an improvement of 10.6%"

```
getModelandEvaluation(dat, graph = 1, print = 1, plot_title = "paper model paper 2004-2011 data")
```

```
## [1] "Summary of base model"
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + lag(y, -12)))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.209554 -0.034684  0.002482  0.040477  0.220976
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.67266    0.76355   0.881 0.381117
## lag(y, -1)   0.64345    0.07332   8.776 3.59e-13 ***
## lag(y, -12)  0.29565    0.07282   4.060 0.000118 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07985 on 76 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.7185, Adjusted R-squared:  0.7111
## F-statistic:    97 on 2 and 76 DF,  p-value: < 2.2e-16
##
## [1] "Summary of trend model"
##
## Call:
## lm(formula = dyn(y ~ lag(y, -1) + lag(y, -12) + suvs + insurance),
##     data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.161327 -0.043774  0.002998  0.036651  0.159219
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.45798    0.78438  -0.584 0.561081
## lag(y, -1)   0.61947    0.06318   9.805 5.09e-15 ***
## lag(y, -12)  0.42865    0.06535   6.559 6.45e-09 ***
## suvs         1.05721    0.16686   6.336 1.66e-08 ***
## insurance   -0.52966    0.15206  -3.483 0.000835 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06509 on 74 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.8179, Adjusted R-squared:  0.808
## F-statistic: 83.08 on 4 and 74 DF,  p-value: < 2.2e-16
##
## [1] "paper model paper 2004-2011 data MAE"
##   mae.base mae.trends  mae.delta
## 0.06343984 0.05667658 0.10660890
```



Here, we get the exact same model, MAE, and graph as the paper. So we know that mainly their model is reproducible.

Result 2:

The paper finds that "if we look at the MAE during the recession (2007/12 - 2009/06), we find that the MAE without Trends data is 8.86% and with Trends data is 6.96%, an improvement of 21.4%"

```
getModelandEvaluation(dat, dateBegin = "2007-12-01", dateEnd = "2009-06-30", plot_title = "paper model
```

```
## [1] "paper model paper 2007-2009 data MAE"
##   mae.base mae.trends  mae.delta
## 0.08869325 0.06965812 0.21461753
```

We get the same improvement in MAE for the recession period–21.4%.

Part 3: Reproduction using paper's model and our data

3.1 Overview:

There are two main problems we had with how the paper dealt with their data:

1. The Google Trends data we downloaded are different from the Google Trends data the paper provides
2. The paper only looks at the MAE for 2007-2009, but it does not look at 2004-2007 or 2010-2011. We suspect that the reason why the trends model has an overall MAE improvement of 10% is that the significance of the improvement during the period of 2007-2009 outweighs the failure of the trends model in different time periods. This might mean that the trends model only works well during these economic recession periods, but does not help improving the base model during other periods.

To test our assumptions, we decided to reproduce the paper's model in couple ways:

- Paper's model, paper's 2004-2007 data
- Paper's model, paper's 2010-2011 data
- Paper's model, our 2004-2008 data
- Paper's model, our 2007-2008 data
- Paper's model, our 2004-2007 data
- Paper's model, our data from different time period

3.2 Paper's model, paper's data from different periods

In this section, we look at MAE improvement for different time ranges (2004-2007, 2010-2011, and compare with 2007-2009) using Paper's model and data

```
# before recession
getModelandEvaluation(dat, print = 0, dateBegin = "2004-01-01", dateEnd = "2007-12-01", plot_title = "pa
```

```
## [1] "paper model paper 2004-2007 data MAE"
##    mae.base  mae.trends   mae.delta
##  0.03980501  0.04359489 -0.09521124
```

```
# during recession
getModelandEvaluation(dat, print = 0, dateBegin = "2007-12-01", dateEnd = "2009-06-30", plot_title = "pa
```

```
## [1] "paper model paper 2007-2009 data MAE"
##   mae.base mae.trends  mae.delta
## 0.08869325 0.06965812 0.21461753
```

```
# after recession
getModelandEvaluation(dat, print = 0, dateBegin = "2010-01-01", dateEnd = "2011-7-01", plot_title = "pap
```

```
## [1] "paper model paper 2010-2011 MAE"
##   mae.base mae.trends  mae.delta
## 0.06416302 0.05054366 0.21226185
```

Observation:

- Before recession, Trend data doesn't help improvement (decrease) MAE, instead, MAE increases by about 9%.
- During recession, Trend data help decrease MAE by about 21% as the paper claims.
- After recession, Trend data also help decrease MAE by about 21%.

- Recall that overall, MAE improves by about 10%.

Our speculation:

Trend data may help during times of economic turbulence, like during recession (2007-2009) and when the economy is in the process of getting out of recession (2010-2011). The overall improvement in MAE may be caused by the relatively big decreases during periods of major economic changes. As the paper mentions in the second example of their paper, the trends model tend to help during the "turning point" of the economy, since the previous trends in time series cannot be maintained as a result of irregular economic behavior, but Google Trends data can keep track of the real-time interest of "Trucks & SUVs" better. As the search interest decreases, the sales of trucks and suvs are likely to decrease as well.

3.3 Paper's model, our data

In this section, we will use our data from 2004-2008, 2009-2013, 2014-2019 to reproduce their results. ### 3.3.1 difficulties in data processing

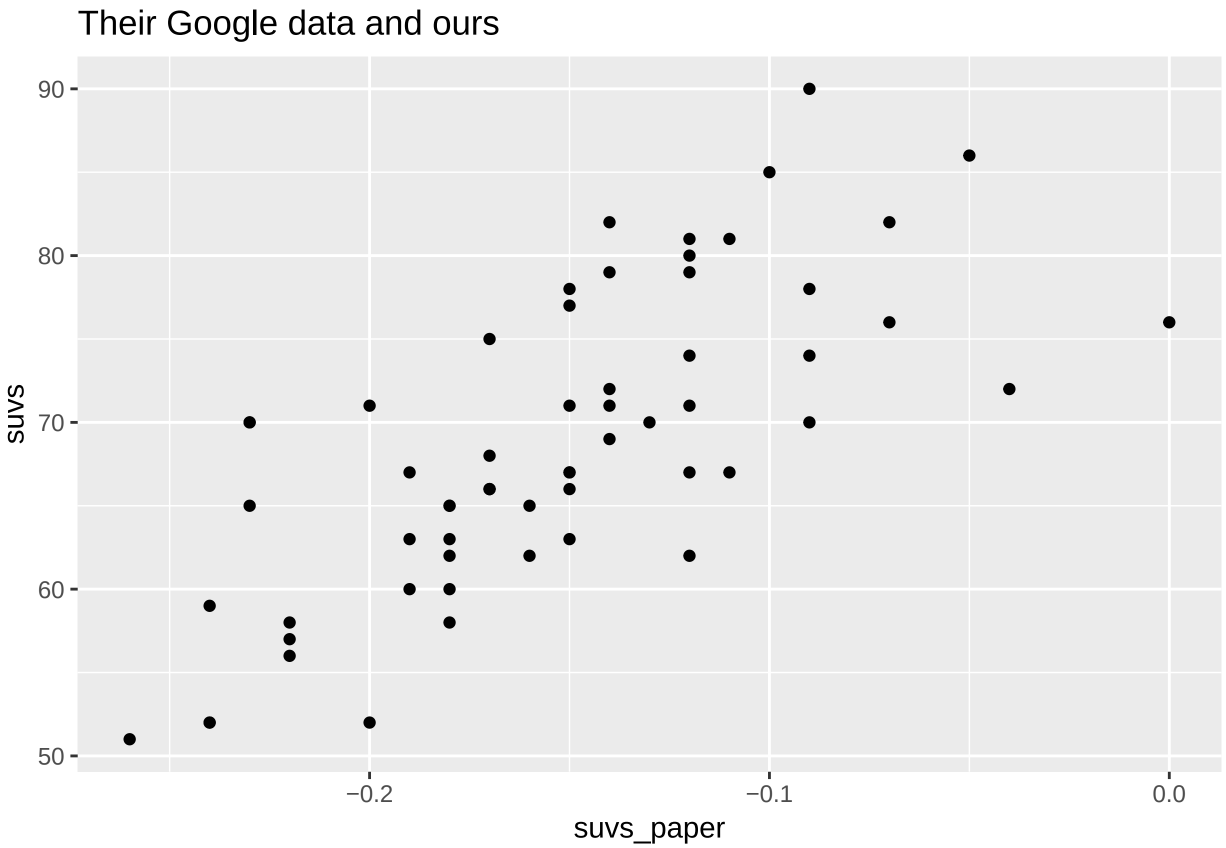
We tried to download the Google Trends data for the same category and the same time period, but there are two things that did not work out:

- Google Trends provides data in weekly format only when the searched time window is within 5 years. Also, they normalize their index for every query we do, so it is difficult for us to download data in 5-year ranges and combine them. As a result, we choose to only work with the data from 2004-2008 for replication, and 2009-2013, 2014-2019 for reproduction.
- Google Trends data we downloaded appear differently from the Google Trends data in the paper. For example, for the Trucks & suvs, the data we have is normalized in 0-100 range, but the data they had are all within (-0.5,0) range. Though they provided some explanation in how Google Trends normalize their index back then, it is very unclear. We also couldn't find much information about this issue online.
- As the paper stated, Google Trends uses a sampling method for their data, and as a result, data extracted on different days can vary by couple percent (but we have also verified on the side that they do not vary by too much, but this difference still influences the MAE. Due to time and space limitation, we would not include this in the report, but for more information on that, contact our group members.)
- The link for downloading Google Trends data only last for a day or so. As a result, our makefile will not work past a day. If anyone want to replicate our experiment and they try to find Google Trends data during the same period, they will not get the exact same ones due to the previous bulletin point.

(We do provide all our files in the data folder, and see Part 4.3 for more information)

Here is a graph showing the problem of the data and what might cause confusion in the future if anyone else were to replicate this paper.

```
trends_ours <- data.frame(merged0408)
trends_ours$Period = as.Date(trends_ours$Period)
trends_paper <- dat[1:60, ] # to make it 2004-2008
colnames(trends_paper) <- c("Period2", "sales2", "suvs_paper", "insurance_paper")
plot_data <- cbind(trends_ours, trends_paper)
ggplot(plot_data, aes(x = suvs_paper, y = suvs)) +
  geom_point() +
  ggtitle("Their Google data and ours")
```



```
# r-squared
summary(lm(suvs_paper ~ suvs, plot_data))$r.squared
```

```
## [1] 0.5236025
```

From the R-squared, we see that their google trends data and ours are not very correlated. It is hard for us to know why, but we see from the graph that they create a decent linear trend. We can also try rescaling the google data using $log(x)/100$.

3.3.2 paper model our 2004-2008 data both unscaled and scaled

We first test it out with 2004-2008 data: since we can only extract data within a 5 year period, we are going to compare our result with the paper's result using the 2004-2008 chunck of paper's data

```
#unscaled and compare
getModelandEvaluation(merged0408, print = 0, plot_title = "paper model our 2004-2008 data")
```

```
## [1] "paper model our 2004-2008 data MAE"
##   mae.base mae.trends  mae.delta
## 0.05938815 0.05384969 0.09325856
```

```
#getModelandEvaluation(merged0408, print = 1, plot_title = "paper model our 2004-2008 data")
#scaled
scaled_merged0408 <- merged0408
scaled_merged0408$suvs <- log(merged0408$suvs/100)
scaled_merged0408$insurance <- log(merged0408$insurance/100)
getModelandEvaluation(scaled_merged0408, plot_title = "paper model our scaled 2004-2008 data")
```

```
## [1] "paper model our scaled 2004-2008 data MAE"
##   mae.base mae.trends  mae.delta
## 0.05938815 0.05318478 0.10445462
```

```
#compared to the MAE of paper data from 04-08
getModelandEvaluation(dat,  dateBegin = "2004-01-04", dateEnd = "2008-12-02", plot_title = "paper model
```

```
## [1] "paper model paper 2004-2008 data MAE"
##   mae.base mae.trends  mae.delta
## 0.05910059 0.05306585 0.10210973
```

Observations:

For space saving purposes, only the MAEs are shown. The summary of base and trends model from our 2004-2008 data is not shown, but run the commented code `getModelandEvaluation(merged0408, print = 1, plot_title = "paper model our 2004-2008 data")` to check our observations regarding the coefficients of these models.

- Interestingly, if we use the data we found for years 2004-2008, we are able to reproduce similar in-sample forecast results in the paper. Although the coefficients are different, the signs are the same; the significance of each predictor is similar too, except for insurance which is not signifcant here. Furthermore, we found that MAE decreased by a similar amount, i.e. about 10% with the addition of Trend data. Finally, the R^2 value is close to the paper's.

- Regarding the rescaling: We see about the same result for the unscaled and scaled data in terms of MAE and R^2, so scaling is not that helpful

- Right now we still see that over 2008the addition with Google Trend's data helps with prediction in the sense that we observe an overall decrease in MAE is only true for certain years, specifically, years before 2008.

3.3.3 paper model our 2009-2013, 2014-2019 data

```
getModelandEvaluation(merged0913, plot_title = "paper code 2009-2013 data")
```

```
## [1] "paper code 2009-2013 data MAE"
##    mae.base  mae.trends   mae.delta
##  0.03587004  0.04184915 -0.16668815
```

```
#getModelandEvaluation(merged0913, print =1, plot_title = "paper code 2009-2013 data")
getModelandEvaluation(merged1419, plot_title = "paper code 2014-2019 data")
```

```
## [1] "paper code 2014-2019 data MAE"
##    mae.base  mae.trends   mae.delta
##  0.02215164  0.02498907 -0.12809136
```

```
#getModelandEvaluation(merged1419,print = 1, plot_title = "paper code 2014-2019 data")
```

Observations:

For space saving purposes, only the MAEs are shown. The summary of base and trends model from our 2004-2008 data is not shown, but run the commented code `getModelandEvaluation(merged0913, print =1, plot_title = "paper code 2009-2013 data")` and `getModelandEvaluation(merged1419,print = 1, plot_title = "paper code 2014-2019 data")` to check our observations regarding the coefficients of these models.

- For in sample forecast for both time periods (i.e.the base and trends model), the coefficients of both the suvs and insurance predictors are very small (around 0.002, -0.002 for both periods.) Neither of these google trends data are significant enough.

- More importantly, for both 2009-2013 and 2014-2019, the MAE increased by a lot (16.7% and 12.8%).

Speculations:

Given our previous suspicion about the trends model being useful only during the recession period, the result in this section might indicate a similar conclusion, since these periods do not have major turning points in the economy. However, this remains a speculation until our next portion.

3.3.4 paper model our 2004-2007, 2007-2008 data

```
#before recession
getModelandEvaluation(merged0408, dateBegin = "2004-01-04", dateEnd = "2007-12-02", plot_title = "paper
```

```
## [1] "paper model our 2004-2007 data MAE"
##    mae.base  mae.trends   mae.delta
##  0.03989324  0.04442203 -0.11352295
```

```
#during recession
getModelandEvaluation(merged0408,  dateBegin = "2007-12-02", dateEnd = "2008-12-07", plot_title = "pape
```

```
## [1] "paper model our 2007-2008 data MAE"
##   mae.base mae.trends  mae.delta
## 0.10425488 0.07467627 0.28371439
```

```
#compare to paper's in the same time range
getModelandEvaluation(dat,  dateBegin = "2007-12-02", dateEnd = "2008-12-07", plot_title = "paper model
```

```
## [1] "paper model paper 2007-2008 data MAE"
##   mae.base mae.trends  mae.delta
## 0.10894751 0.07753248 0.28835020
```

Observations:

- From 3.3.2, we see that using our 2004-2008 data, there is an overall decrease of MAE by about 9%, close to the papers.
- However, before recession, we also found that Trend data doesn't help; MAE increased by about 11%.
- Finally, like the paper, we found that during recession, Trend data helps to decrease MAE by a lot more than the 9% overall. In our case, the value decreased by about 28% for a period of 2007/12-2008/12, 6 month short of paper's data.So we compared the paper's data's MAE improvement in the same time range and found that the increase is also about 28%. Thus, in some sense, we've sucessfully reproduced the paper's result even though we don't have all the data that the authors used.

Speculations:

At this point, our observations are all pointing to our suspicion at the beginning of part 3. Namely, trends data only seem to be helpful during 2007-2009 and can potentially perform worse than the base AR-1 model.

Part 4: Extending their model

4.1 Overview:

Motivations:

- They used lag1 and lag 12 in their base model, and we want to see if lag1 and lag2 can somehow be more helpful in predicting the present when interacting with the google trends data, since lag1 lag2 can also encode how the previous two months have changed. Compared to log1 lag12, lag1 lag2 might better model the changes during the beginning of the recession period since it is not helpful to look at the data from 12 month ago to predict a drastic change at present.
- They used a rolling window forecasting which also take into account all the months before the month of interest. We think that a moving window forecast might work better, especially during the recession period. Starting from the 18th month in the data, we use only the previous 17 months (1.5 years), instead of all the previous months, to train a model and predict the next month.

Therefore, we did two ways of extending their model and method, and we tested these model using paper's data:

- First,
  - we kept their rolling window out of sample forecast method,
  - but updated their trends model to $y = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + suvs + insurance$, and compare it against base model: $y = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2}$
  - function in use from oosf_new.R and getEvaluation.R: OutOfSampleForecast1_2(), getModelandEvaluation1_2
- Second,
  - we kept their base and trends model, note that the base model was also trained using moving window.
  - but we updated their out of sample forecast to moving window forecast
  - function in use from oosf_new.R and getEvaluation.R: OutOfSampleForecast0112_mw(), getEvaluation_0112mv

We also decided to keep dividing their data to three periods–before, during, and after recession so that we can further investigate our speculation in part 3.

4.2 our lag1 lag2 trends model paper data

```
#Before recession
getModelandEvaluation1_2(dat, dateBegin = "2004-01-01", dateEnd = "2007-12-01", plot_title = "our trend
```

```
## [1] "our trends model paper 2004-2007 data MAE"
##     mae.base   mae.trends    mae.delta
## 0.0687323802 0.0686754770 0.0008278957
```

```
#During recession
getModelandEvaluation1_2(dat, dateBegin = "2007-12-01", dateEnd = "2009-06-30", plot_title = "our trend
```

```
## [1] "our trends model paper 2007-2009 data MAE"
##   mae.base mae.trends  mae.delta
##  0.0706302  0.0591258  0.1628822
```

```
#After recession
getModelandEvaluation1_2(dat, dateBegin = "2010-01-01", dateEnd = "2011-7-01", plot_title = "our trends
```

```
## [1] "our trends model paper 2010-2011 data MAE"
##    mae.base  mae.trends   mae.delta
##  0.05741737  0.06696378 -0.16626332
```

```
#over 2004-2011
getModelandEvaluation1_2(dat, plot_title = "our trends model paper 2004-2011 data")
```

```
## [1] "our trends model paper 2004-2011 data MAE"
##   mae.base mae.trends  mae.delta
## 0.06821008 0.06602176 0.03208200
```

```
#paper model for comparison
getModelandEvaluation(dat, plot_title = "paper model and paper 2004-2011 data")
```

```
## [1] "paper model and paper 2004-2011 data MAE"
##   mae.base mae.trends  mae.delta
## 0.06343984 0.05667658 0.10660890
```

Observations:

- Before recession, this model indicates that the improvement in MAE before Recession is close to 0 with the addition of Trend data; compare this with the paper's model, which showed that MAE increased by 9% rather than getting smaller.
- During recession: note that this new model doesn't perform as well as the paper's model in improving MAE during Recession (~16% vs ~21%).
- After recession: another discrepancy is found after Recession: whereas the paper's model showed that Trend data helps to decrease MAE by about 21%, this model shows quite the opposite–MAE increased by 16%.
- Over 2004-2011: Our model does not perform as well (only 3% improvment compared to 9%)

Speculation:

- The result did not come out as we had expected. We are not completely sure why lag1 lag2 perform in such a way without more understanding about the models.
- But we did see that our model performs well during recession period, thus confirming once again that the google trends data is more helpful during the recession period.

4.3 paper's data our moving window forecast

```
#Before recession
getEvaluation_0112mv(dat, dateBegin = "2004-01-01", dateEnd = "2007-12-01", plot_title = "our forecast m
```

```
## [1] "our forecast method paper 2004-2007 data MAE"
##    mae.base  mae.trends   mae.delta
##  0.03945803  0.04156328 -0.05335427
```

```
#During recession
getEvaluation_0112mv(dat, dateBegin = "2007-12-01", dateEnd = "2009-06-30", plot_title = "our forecast m
```

```
## [1] "our forecast method paper 2007-2009 data MAE"
##   mae.base mae.trends  mae.delta
## 0.07480836 0.05465174 0.26944344
```

```
#After recession
getEvaluation_0112mv(dat, dateBegin = "2010-01-01", dateEnd = "2011-7-01", plot_title = "our forecast me
```

```
## [1] "our forecast method paper 2010-2011 data MAE"
##   mae.base mae.trends  mae.delta
## 0.06719328 0.06027477 0.10296425
```

```
# over 2004-2011
getEvaluation_0112mv(dat, plot_title = "our forecast method and paper 2004-2011 data")
```

```
## [1] "our forecast method and paper 2004-2011 data MAE"
##   mae.base mae.trends  mae.delta
## 0.06154527 0.05531033 0.10130653
```

```
#compare MAE from rolling window
getModelandEvaluation(dat, plot_title = "paper forecast method and paper 2004-2011 data")
```

```
## [1] "paper forecast method and paper 2004-2011 data MAE"
##   mae.base mae.trends  mae.delta
## 0.06343984 0.05667658 0.10660890
```

Observations:

- Here, we want to note the improvement in predicting the sales during 2007-2009 (from 21.4% to 26.9%)
- The rest are similar to what we have for the paper's model

Speculation:

- The reason of this improvement might be as we hypothesizes–that for recession period, it is better to not look at data from all the way past, since the data from 2004 cannot really inform what happens in 2007-2009 when economy is in turmoil

Part 5: Some other explorations we did and some potential future extensions

These are some more extensions that we tried out but did not have enough time to pursue, but we want to leave it here because we believe that with proper exploration, they will be helpful in perfecting the conclusion of the original paper even more. Please contact us if you want the code and more information about this.

1. We had questions why they ended up choosing “Trucks and Suvs” and “Auto Insurance”" as the two predictors. They explained only a little in their paper, but they did not provide code for it. So here are two things we explored:
    - Use forward algorithm and the step function in R or L1 regularization to better determine which predictor is most helpful. (We found that both “Trucks and Suvs”" and “Auto Insurance” are significant, contact us for more information.)
    - Add another predictor from Google Trends that are not directly related to motor vehicles but with the overall economy in general, since we found that the google trends model is more helpful during recession period. For example, we played around with the category “Currencies and Foreign Exchange.” We did not have enough data to test out models with more than 5 predictors, but we put in “currencies” in place of insurance, and we saw a similar improvement during the recession time.
2. We downloaded google trends data for the same categories over the same period of time multiple times during our research. But everytime they are slightly different as we have mentioned in Part 3. However, when we put these data in and compare the results, the MAE is highly influenced for 2004-2008 period, but it is not as influenced during the recession period

```
#compare 2004-2008 Google Trends data, taken a day apart
getModelandEvaluation(merged0408_v2, plot_title = "paper model our new set of 2004-2008 data")
```

```
## [1] "paper model our new set of 2004-2008 data MAE"
##    mae.base  mae.trends   mae.delta
##  0.05938815  0.06081592 -0.02404145
```

```
getModelandEvaluation(merged0408, plot_title = "paper model our first set of 2004-2008 data")
```

```
## [1] "paper model our first set of 2004-2008 data MAE"
##   mae.base mae.trends  mae.delta
## 0.05938815 0.05384969 0.09325856
```

```
#compare 2007-2009 Google Trends data, taken a day apart
getModelandEvaluation(merged0408_v2,  dateBegin = "2007-12-02", dateEnd = "2008-12-07", plot_title = "pa
```

```
## [1] "paper model our 2007-2008 data MAE"
##   mae.base mae.trends  mae.delta
## 0.10425488 0.08427153 0.19167786
```

```
getModelandEvaluation(merged0408,  dateBegin = "2007-12-02", dateEnd = "2008-12-07", plot_title = "paper
```

```
## [1] "paper model our 2007-2008 data MAE"
##   mae.base mae.trends  mae.delta
## 0.10425488 0.07467627 0.28371439
```

- **Observation and speculation:**
- The MAE is only relatively stable across the two different google trends data during the recession period. So, we suspect the validity of paper's result of a 10% improvement on MAE over the entire 2004-2011. Using our first version of data, we get a similar MAE improvement for 2004-2008 (9%), but useing the second version of data, we get a increase in MAE, which is not improved at all. This might be because of the specific evalution method they chose, i.e. MAE, so maybe changing an evaluation parameter will be helpful in the future.

3. Other models we also explored yet not included: we also tested out lag1, lag2, lag3, and adding I(lag1^2), but we chose to include what we have in paper because it is the most important finding, see our `oosf_new.R` file for more information.

Part 6: Conclusion

1. We can reproduce their experiment using their model and their data. However, with a different set of data, their model does not work as well, especially during time periods that are not recession times. Even using their own data, if divided into different time chuncks, the google trends model only works well during 2007-2009, but not 2004-2007.
2. Our extensions of the model make sense since they do show improvement during the recession period. Moving window forecast is especially helpful during the recession period.

So overall, we speculate that:

1. Google trends model is more helpful during the recession period. In other periods, the trends model either does not help or decreases the MAE of the base model.
2. As a result of the sampling method Google trends uses to come up with their index, the data vary, and MAE is heavily influenced during times outside of recession. So, their conclusion of improvement in MAE over 2004-2011 can be due to the specific Google trends data set they get at the day of research.