The nowcasting of economic conditions in Japan using machine learning

Quantitative Methods
in the Social Sciences,
Columbia University

Taketo Muroya

Introduction

Since the official economic statistics in Japan are usually published more than one month later than the period to which they refer, it is difficult to assess the current economic conditions in a **timely** and **accurate** manner.

Therefore, this study tries to **nowcast the Index of Business Conditions (IBC)**, which represents monthly economic conditions, using text data on **economic reports**, and numeric values from **Google searches** and **electricity usages**.

The nowcasting of the IBC is mainly made by machine learning methods, **Random Forest** and **Recurrent Neural Network with Long Short-Term Memory (RNN-LSTM)**, in addition to the traditional econometrics model.

Data 1: Monthly Economic Report

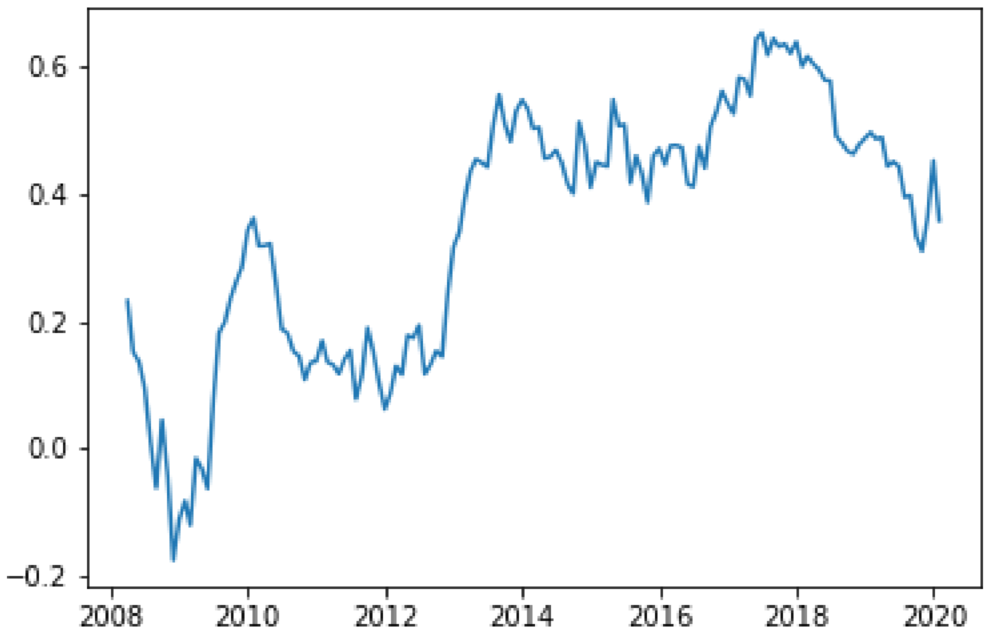
- The first predictor is the text data on Monthly Economic Report issued by the government of Japan
- Need to transform the text data into numeric values by Natural Language Processing (NLP)
- Sentiment score has similar shape as economic conditions

Word Cloud

留意
必要
先行き
雇用
情勢
増加
輸入
増加
推移
減少
緩やか
輸出
運営
基調
物価
動向
前月比
経済
4月
予算
企業
アメリカ
緩やか
横ばい
景気
上昇
悪化
3月
雇用
企業
収益持ち直して
設備投資
生産
2月
動き
見込まれ
投資

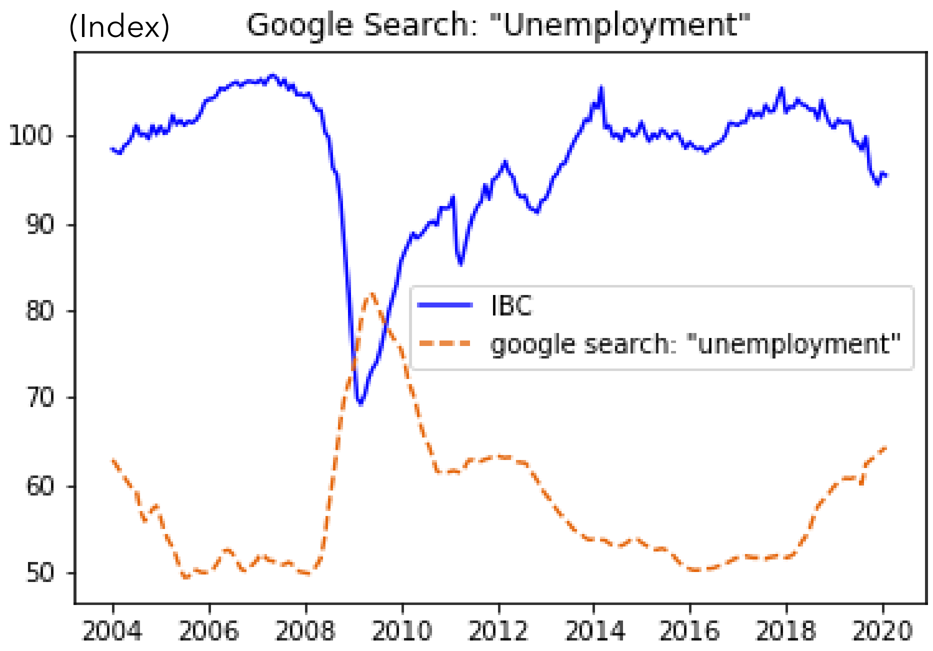
(translation: economy, increase, decrease, production, labor and so on)

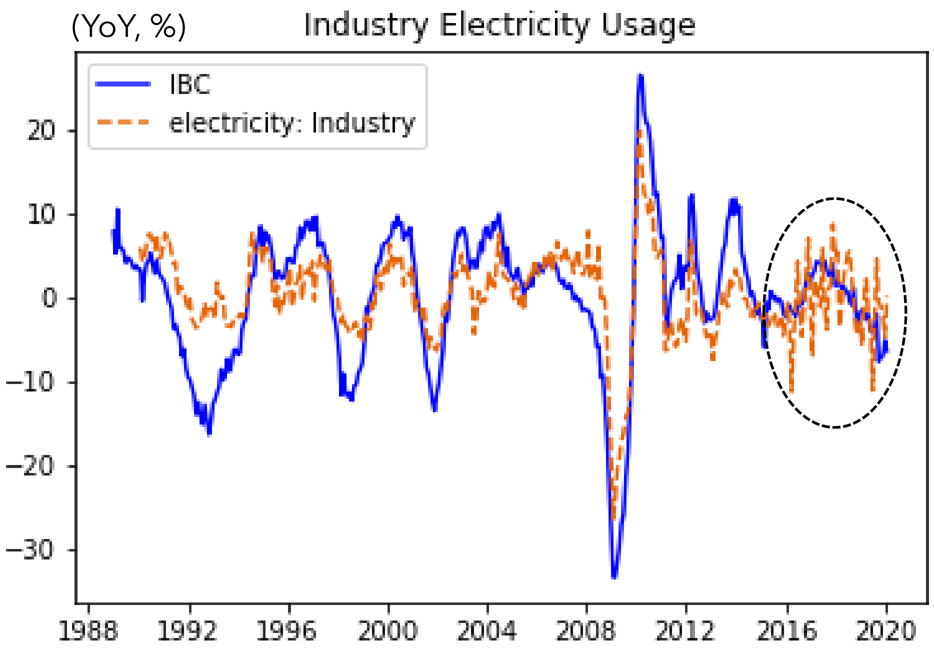
Sentiment score



Data 2 & 3: Google Search & Electricity Usage

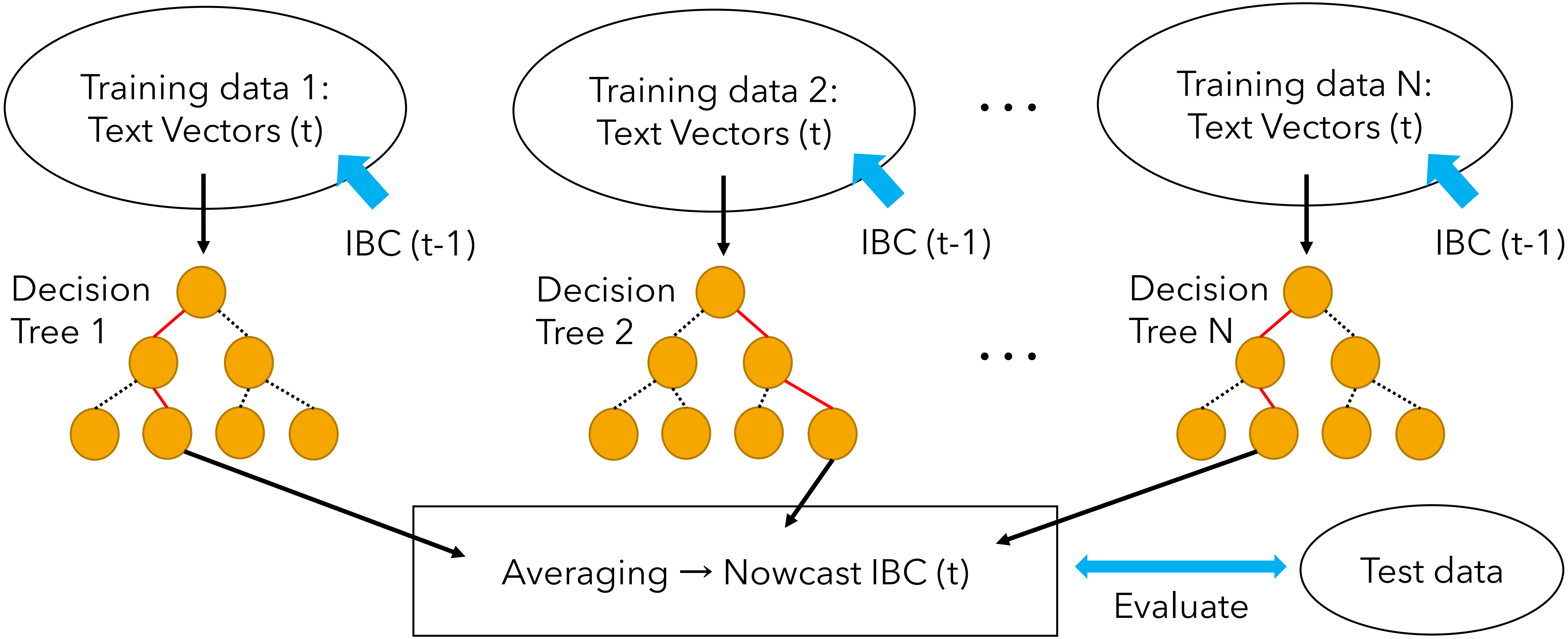
- The second predictor is Google search volume of “unemployment,” negatively correlated with IBC
- The third predictor is the electricity usage, especially industry usage, positively correlated with IBC
- After March 2016, the industry usage is an estimation





Methods 1: Random Forest

- Random forest nowcasts IBC using randomized training data and making many decision trees (e.g. 1,000)
- In addition, this model incorporates the previous value of IBC as predictors to deal with autocorrelation issues
- Output is evaluated by test data in terms of R-squared



Training data 1: Text Vectors (t)

Training data 2: Text Vectors (t)

. . .

Training data N: Text Vectors (t)

IBC (t-1)

IBC (t-1)

IBC (t-1)

Decision Tree 1

Decision Tree 2

. . .

Decision Tree N

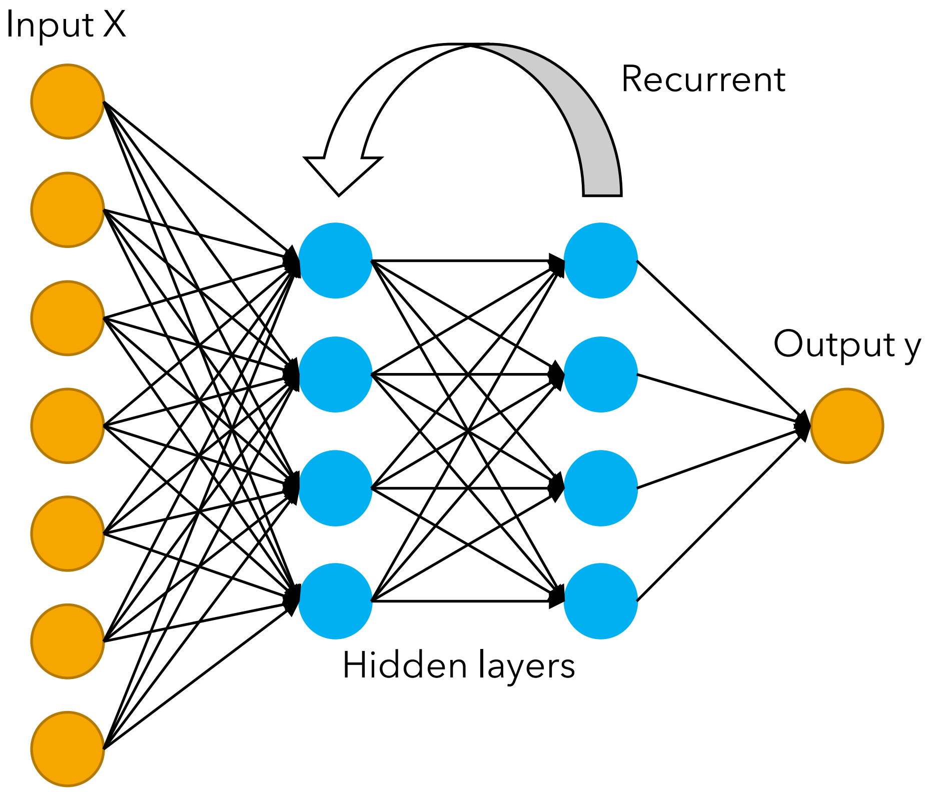
Averaging → Nowcast IBC (t)

Evaluate

Test data

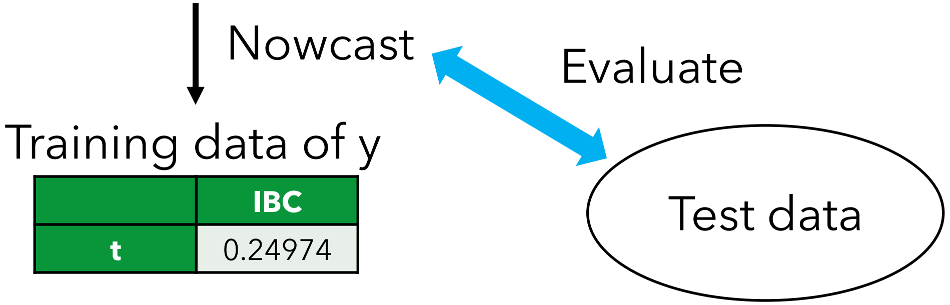
Methods 2: RNN-LSTM

- RNN-LSTM nowcasts IBC (y) using predictors (X) through the recurrent network built in the hidden layers
- Training data X consists of not only past IBC and past predictors but also current predictors
- Output is evaluated by test data in terms of R-squared



Training data of X

	IBC	Text V1	Text V2	Text V3	…
t-4	1.01782	0.32868	-0.21462	0.82724	…
t-3	1.04494	0.30665	-0.19517	0.75472	…
t-2	0.74373	0.37481	-0.19603	0.84563	…
t-1	0.65939	0.40235	-0.20816	0.94303	…
t	?	0.40593	-0.19796	0.92115	…



Nowcast

Evaluate

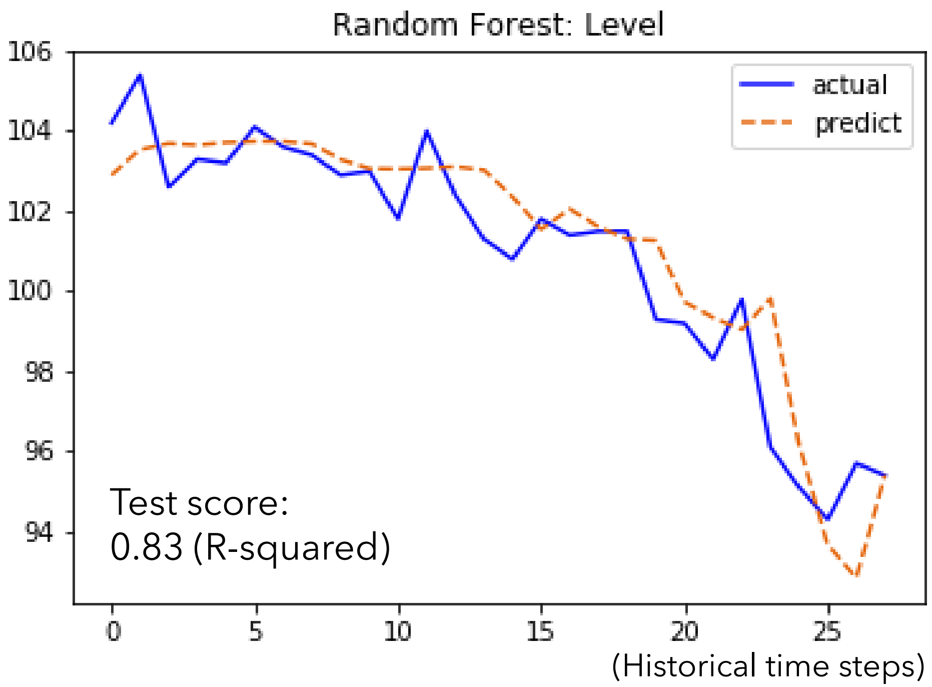
Training data of y

	IBC
t	0.24974

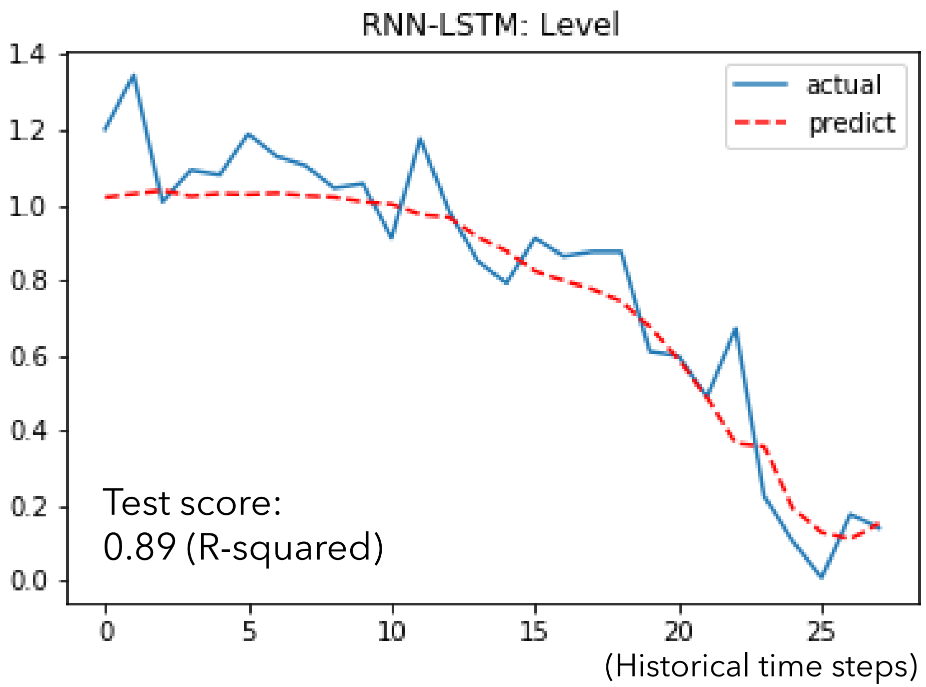
Test data

Result 1: Text Vector (Economic Report)

- Using tf-idf vectorizer, text data on economic reports are transformed into numeric vectors as predictors
- Both Random Forest and RNN-LSTM nowcast IBC with high test score (0.83-0.89)
- However, the economic report is not so timely published


Random Forest: Level

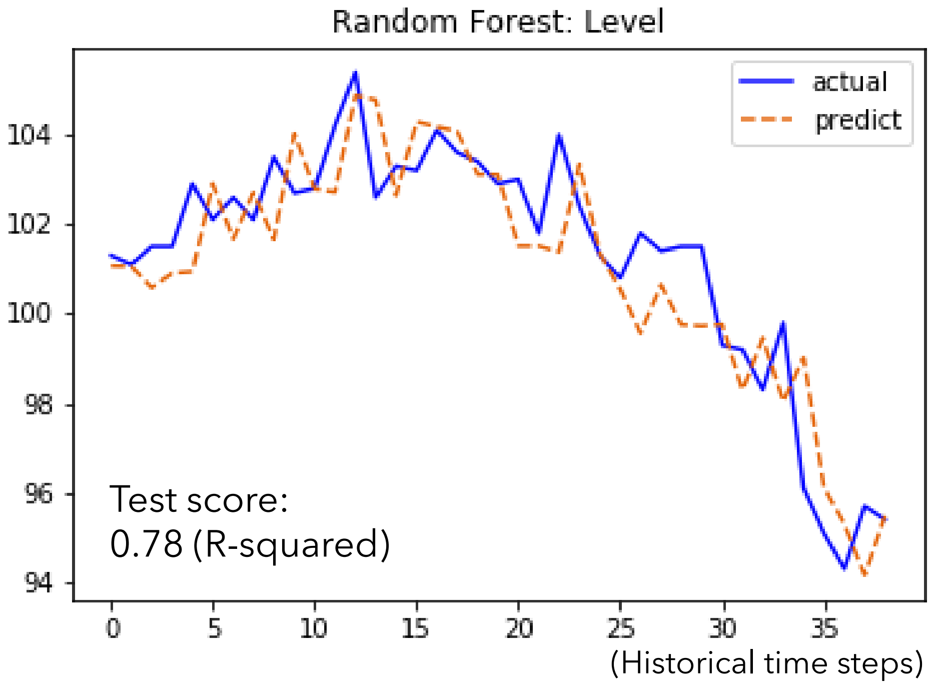
Test score:
0.83 (R-squared)


RNN-LSTM: Level

Test score:
0.89 (R-squared)

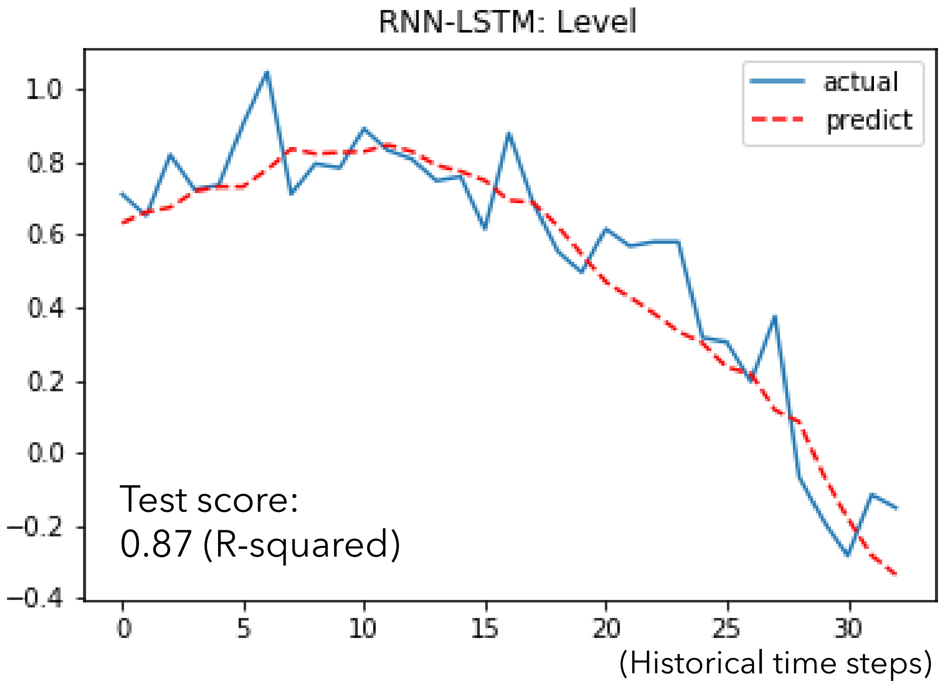
Result 2: Google Search

- Search volume of “unemployment” nowcasts IBC using RNN-LSTM with relatively high test score (0.87)
- Google search data is available even on a daily basis
- If other query words correlated with IBC were detected, the performance of nowcasting might have improved


Random Forest: Level

Test score:
0.78 (R-squared)

(Historical time steps)

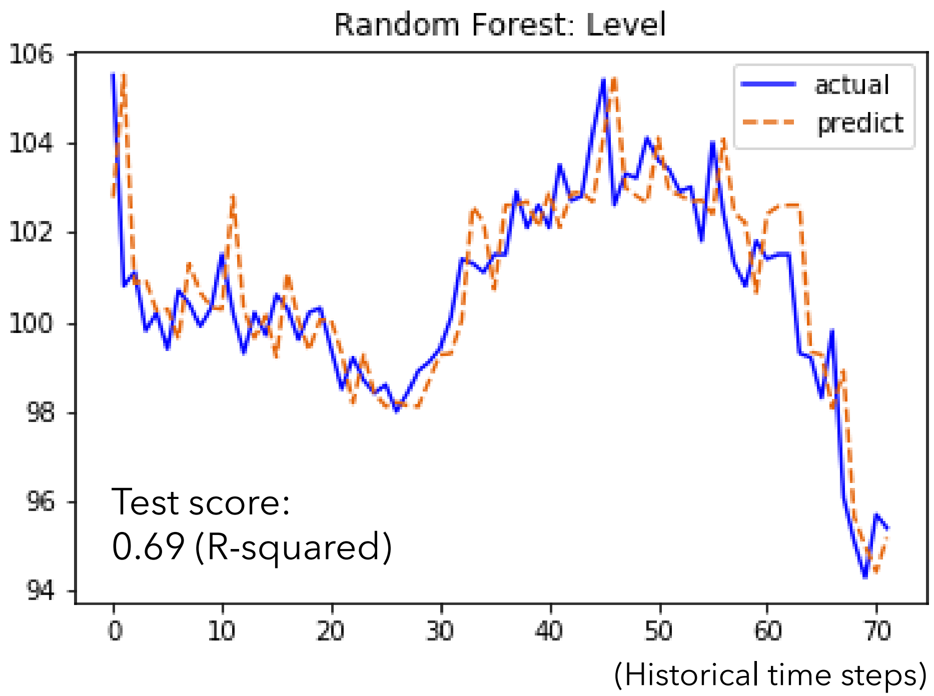

RNN-LSTM: Level

Test score:
0.87 (R-squared)

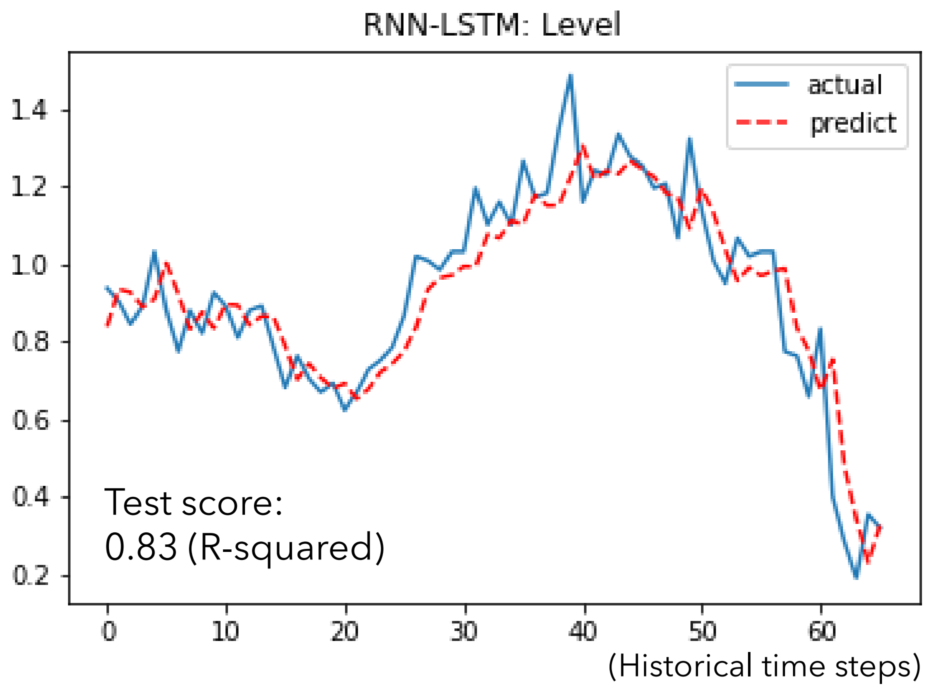
(Historical time steps)

Result 3: Electricity Usage

- Electricity usage nowcasts IBC using RNN-LSTM with relatively high test score (0.83)
- Electricity data is also available even on a daily basis
- For the data up to March 2016 where industry usage is still available, the test score is 0.91 using RNN-LSTM


Random Forest: Level

Test score:
0.69 (R-squared)


RNN-LSTM: Level

Test score:
0.83 (R-squared)

Discussion & Conclusion

The text data from economic reports nowcasts IBC with high test scores. Google search and electricity usage also nowcast IBC with relatively high test scores in a timelier manner. In addition, **RNN-LSTM performs nowcasting more successfully** than Random Forest model for all cases.

For electricity usage, the performance of nowcasting would have increased if separated data for industry and household electricity, which has been suspended after March 2016, could have been available as predictors.

Since the approaches in this study would be applicable for other predictors in general, as more and more big data related with economic activities become available in the future, **the application coverage of this study would be enlarged**.