Lancaster University
Management School

# Managmeent Science

## Working Paper 2018:2

# Online supplement: Demand forecasting with user-generated online information

Oliver Schaer, Nikolaos Kourentzes and Robert Fildes

*The Department of Management Science*
*Lancaster University Management School*
*Lancaster LA1 4YX*
*UK*

Online supplement: Demand forecasting with user-generated online information

Oliver Schaer[a,*], Nikolaos Kourentzes[a], Robert Fildes[a]

[a] *Department of Management Science, Lancaster University Management School, UK*

## A. Supporting tables for the literature review

The following Tables A1 to A5 provide detailed insights on the literature surveyed for each of the forecasting applications used in Section 2. For each reviewed study we list the *target variable*, data *source*, identified *best model* and *benchmarks* used. Studies with distinct datasets have been split into multiple lines. The column *type* names the internet platform and the column *measure* indicates the nature of the data, e.g. volume or sentiment information. We also report the data *frequency*, number of *series*, *maximum lag* the user generated information variable is incorporated and *forecast horizon*. The last column reports the relative *maximum error reduction* achieved, compared to the best performing benchmark without any. All forecasting model abbreviations are listed in Table A6.

*Correspondance: O Schaer, Department of Management Science, Lancaster University Management School, Lancaster, Lancashire, LA1 4YX, UK. Tel.: +44 1524 592911

*Email addresses:* o.schaer@lancaster.ac.uk (Oliver Schaer), n.kourentzes@lancaster.ac.uk (Nikolaos Kourentzes), r.fildes@lancaster.ac.uk (Robert Fildes)

Table A1: Forecasting economic indicators

| Paper | Target Variable | Type* | Measure§ | Best model | Benchmark model(s) | Freq.† | num. series | max. lag | max. horizon | max. % err. red. |
|---|---|---|---|---|---|---|---|---|---|---|
| Choi and Varian (2009)[a,c] | Home sales USA | GTD | Pop. | **ARX** | AR | m | 1 | 0 | 1 | 12% MAE |
| | Retail sales USA | GTD | Pop. | **sARX** | sAR | m | 1 | 0 | 1 | 18% MAE |
| Vosen and Schmidt (2011)[c,d] | Private consumption USA | GTD | Pop. | **ARX** | AR, ARX | m | 1 | 3 | 1 | 76% RMSE |
| Choi and Varian (2012)[a,c] | Motor vehicles and parts USA | GTD | Pop. | **sARX** | sAR | m | 1 | 0 | 1 | 10.5% MAE |
| | Unemployment benefits claims USA | GTD | Pop. | AR | **ARX** | w | 1 | 0 | 1 | AR better |
| | Consumer confidence Index AUS | GTD | Pop. | **ARX** | AR | m | 1 | 0 | 1 | 9.3% MAE |
| Barreira et al. (2013)[a] | Unemployment rate (var. countries) | GTD | Pop. | **ARX** | AR | m | 4 | 2 | 36 | 13.4% RMSE |
| Scott and Varian (2015)[a,b,c,d] | Consumer sentiment USA | GTD | Pop. | **BSTS** | AR, LR | m | 1 | 0 | 1 | 13.5% MAPE |
| | Gun sales USA | GTD | Pop. | **BSTS** | BSTS | m | 1 | 0 | 1 | 55.9% MAE |
| Wu and Brynjolfsson (2015)[a,c,d] | Home sales USA | GTD | Pop. | **ARX** | ARX | q | 1 | 1 | 1 | 2.9% MAE |
| | Home price index USA | GTD | Pop. | **ARX** | ARX | q | 1 | 1 | 1 | 7.1% MAE |
| Brynjolfsson et al. (2016)[a,c] | Unemployment benefits claims USA | GTD | Pop. | **ARX** | AR | w | 1 | 0 | 1 | 3.9% MAE |
| Li (2016)[a,c] | Unemployment benefits claims USA | GTD | Pop. | **DFM** | AR, DFM, RW | w | 1 | 2 | 2 | 0.04% RMSE |
| Limnios and You (2016)[c] | Home prices USA | GTD | Pop. | DW | **ARX**, ARX, **DW** | m | 1 | 6 | 1 | DW better |
| | Price-rent ratio USA | GTD | Pop. | IAC | **ARX**, ARX, **IAC** | m | 1 | 6 | 1 | IAC better |
| Smith (2016)[a,c,d] | Unemployment rate GBR | GTD | Pop. | **AR-MIDAS** | RW w. drift, Survey | m/w | 1 | 1 | 1 | 11% RMSE |
| D'Amuri and Marcucci (2017)[c,d] | Unemployment rate USA | GTD | Pop. | **ARX** | AR, ARX, SETAR, LSTAR, AAR | m | 1 | 4 | 12 | 39% RMSE |
| Bulut (2017)[a,c,d] | Exchange rate movements | GTD | Pop. | RW | **LRX**, LRX | m | 11 | 0 | 1 | RW better |
| Elshendy et al. (2017)[c] | Crude oil price | GTD; TWR; GDELT; WIA | Pop.; Vol.; Val.[2] | **ARIMAX** | ARIMA | d | 1 | 3 | 1 | 88.9% MAPE |
| Yu et al. (2018)[c] | Oil consumption | GTD | Pop. | **SVR** | LR, SVR,ELM,NN | m | 1 | 6 | 1 | 1.4% MAPE |

**Boldface** = Model containing online information;
[a] Nowcasting model;
[b] Only insample evaluation;
[c] Rolling origin evaluation;
[d] Statistical model testing;
* GTD = Google Trends, TWR = Twitter, WIA = Wikipedia;
§ Rat. = Rating, Pop. = Popularity, Vol. = Volume, Val. = Valence ([1] = Lexicon, [2] = Machine learning, [3] = n-grams, [4] = Self declared);
† d = daily, w = weekly, m = monthly.

Table A2: Forecasting financial markets

| Paper | Target Variable | Type* | Measure§ | Best model | Benchmark model(s) | Freq.† | num. series | max. lag | max. horizon | max. % err. red. |
|---|---|---|---|---|---|---|---|---|---|---|
| Bollen et al. (2011)[c,d] | DJIA Index | TWR | Val.[1] | **NN** | NN | d | 1 | 3 | 1 | 7.7% MAPE |
| Rao and Srivastava (2013)[c] | US Oil Funds Index | GTD; TWR | Pop.; Val.[1] | **ARX** | AR | w | 4 | 4 | 1 | 4.1% MAPE |
| | DJIA Index | GTD; TWR | Pop.; Val.[1] | **ARX** | AR | w | 4 | 4 | 1 | 26.8% MAPE |
| | NASDAQ-100 Index | GTD; TWR | Pop.; Val.[1] | **ARX** | AR | w | 4 | 4 | 1 | 2.6% MAPE |
| | Gold Price | GTD; TWR | Pop.; Val.[1] | **ARX** | AR | w | 2 | 4 | 1 | 1.99% MAPE |
| | EURO | GTD; TWR | Pop.; Val.[1] | **ARX** | AR | w | 2 | 4 | 1 | 39.8% MAPE |
| Hamid and Heiden (2015)[c,d] | Index volatility | GTD | Pop. | **ESC** | AR, HAR, ARFIMA | w | 1 | 1 | 1 | 6% MSE |
| Dimpfl and Jank (2016)[c,d] | Index volatility | GTD | Pop. | **VHAR**, **VAR** | AR, HAR | d | 1 | 1 | 14 | 3.5% MSE |
| Bijl et al. (2016)[b,c,d] | Stock returns | GTD | Pop. | **LR** | LR | d | 431 | 5 | 1 | 32% $R^2$ |
| Ho et al. (2017)[c,d] | Stock returns | FOM | Val.[4] | **SUR** | SUR | d | 45 | 1 | 1 | 36.4% MAE |
| Perlin et al. (2017)[c] | Index returns | GTD | Pop. | **VAR** | ARMA-GARCH, BH | m | 4 | 5 | 1 | 1014% Return |

**Boldface** = Model containing online information;
[a] Nowcasting model;
[b] Only insample evaluation;
[c] Rolling origin evaluation;
[d] Statistical model testing;
* BAU = Baidu, FOM = Forum, GTD = Google Trends, TWR = Twitter;
§ Pop. = Popularity, Vol. = Volume, Val. = Valence ([1] = Lexicon, [2] = Machine learning, [3] = n-grams, [4] = Self declared);
† d = daily, w = weekly, m = monthly, q = quarterly.

Table A3: Forecasting public health and environment

| Paper | Target Variable | Type* | Measure§ | Best model | Benchmark model(s) | Freq.† | num. series | max. lag | max. horizon | max. % err. red. |
|---|---|---|---|---|---|---|---|---|---|---|
| Ginsberg et al. (2009)[a,c] | Influenza outbreak USA | GTD | Pop. | **LR** | - | w | 1 | 0 | 1 | - |
| Lampos and Cristianini (2012)[a,c] | Influenza outbreak GBR | TWR | Val.[3] | **LR** | - | w | 3 | 0 | 1 | - |
| | Daily Rainfall GBR | TWR | Val.[3] | **LR** | - | d | 5 | 0 | 1 | - |
| Won et al. (2013) | Suicide events in Korea | BLG | Val.[1] | **ARX** | - | 3-day bins | 1 | 1 | 121 | - |
| Araz et al. (2014)[c] | Influenza outbreak Nebraska | GTD | Pop. | **LR** | sARIMA, HW | w | 1 | 1 | 1 | 72.8 % RMSE |
| Lazer et al. (2014)[a,c] | Influenza outbreak USA | GTD | Pop. | **ARX** | AR | w | 10 | 3 | 1 | 25.4% MAE |
| Preis and Moat (2014)[a,c] | Influenza outbreak USA | GTD | Pop. | **ARX** | AR | w | 1 | 0 | 1 | 21.3% MAE |
| Santillana et al. (2015)[c] | Influenza outbreak USA | GTD; TWR | Pop.; Vol. | **ABT** | **LR**, **SVM**, AR | w | 1 | 4 | 2 | 38.7% MAPE |
| Brynjolfsson et al. (2016)[a,c] | Influenza outbreak USA | GTD | Pop. | **LR** | AR | w | 1 | 3 | 1 | 9.9% MAE |
| Chen et al. (2017)[c] | Smog health hazard China | WEO | Val.[1] | **NN** | NN, SVM, RF | d | 8 | 0 | 1 | 26.0% RMSE |
| Xu et al. (2017)[a,c] | Influenza outbreak Hong Kong | GTD | Pop. | **BMA** | **ARIMAX**, **GLM**, **NN** | w | 1 | 0 | 2 | - |

**Boldface** = Model containing online information;
[a] Nowcasting model;
[b] Only insample evaluation;
[c] Rolling origin evaluation;
[d] Statistical model testing;
* BLG = Blog, GTD = Google Trends, TWR = Twitter, WEO = Weibo;
§ Pop. = Popularity, Vol. = Volume, Val. = Valence ([1] = Lexicon, [2] = Machine learning, [3] = n-grams, [4] = Self declared);
† d = daily, w = weekly.

Table A4: Forecasting services

| Paper | Target Variable | Type* | Measure§ | Best model | Benchmark model(s) | Freq.† | num. series | max. lag | max. horizon | max. % err. red. |
|---|---|---|---|---|---|---|---|---|---|---|
| Choi and Varian (2009)[a,b] | Visitor arrivals to HK | GTD | Pop. | **ARX** | - | m | 9 | 0 | - | - |
| Pan et al. (2012)[c] | Hotel room demand | GTD | Pop. | **ARX** | AR, ARIMA, RW | w | 1 | 1 | 1 | 27% MAPE |
| Hand and Judge (2012) | Cinema admissions | GTD | Pop. | **sARX** | **sARX** | m | 1 | 0 | - | - |
| Bangwayo-Skeete and Skeete (2015)[c,d] | Tourist arrivals in the Caribbean | GTD | Pop. | **AR-MIDAS** | AR, sARIMA | m/w | 5 | - | 12 | - |
| Bughin (2015)[a,c] | Telecom sales in BEL | GTD; TWR | Pop.; Val. | **ARX** | AR | m | 9 | 2 | 1 | 21% RMSE |
| Kim and Shin (2016)[c] | Air passengers | NAR | Pop. | **LR** | - | m | 1 | 8 | 8 | - |
| Önder and Gunter (2016)[c,d] | Tourist arrivals to Vienna | GTD | Pop. | **ADL** | sAR, HW, sRW | m | 7 | 12 | 12 | 56.4% MAE |
| Pan and Yang (2016)[c,d] | Hotel occupancy | GTD | Pop. | **ARMAX** | ARMA, MSDR | w | 1 | 2 | 2 | 4.73% MAPE |
| Peng et al. (2016)[c] | Tourist volume Jiuzhai Valley | BAU | Vol. | **ARX** | **ARX** | d | 1 | 3 | 1 | - |
| Rivera (2016) | Hotel registrations Puerto Rico | GTD | Pop. | **DLM**, HW | sRW, HW, sARIMA | m | 1 | 1 | 12 | 9.18% MAPE |
| Huang et al. (2017)[a,c] | Tourist volume Forbidden City | BAU | Vol. | **ARX** | ARIMA, ARX | d | 1 | 2 | 1 | 14.5% RMSE |
| Li et al. (2017)[c] | Tourist arrivals to Beijing | BAU | Vol. | **ARMAX** | ARMA, **ARMAX** | w | 1 | 5 | 4 | 37% MAPE |
| Önder (2017)[c,d] | Tourist arrivals to Countries | GTD | Pop. | **ADL** | AR, HW, RW | m | 6 | 12 | 12 | 3% MAE |
| | Tourist arrivals to Cities | GTD | Pop. | **ADL** | AR, HW, RW | m | 10 | 12 | 12 | 36.7% MAE |
| Padhi and Pati (2017)[c] | Tourist arrivals to Kerela | GTD | Pop. | **ARIMAX** | ARIMA, **ARX**, **VARX** | m | 1 | 4 | 4 | 53.8% MAPE |
| Park et al. (2017)[c,d] | Tourist arrivals to South Korea | GTD | Pop. | **sARIMAX** | sARIMAX, HW | m | 1 | 6 | 1 | 16.7% MAE |
| Zeynalov (2017)[c] | Tourist arrivals to Prague | GTD | Pop. | **AR-MIDAS** | ARIMA, **ARIMAX** | m/w | 1 | 1 | 1 | 19.7% MAE |
| | Overnight stays in Prague | GTD | Pop. | **AR-MIDAS** | ARIMA, **ARIMAX** | m/w | 1 | 1 | 1 | 8.5% MAE |

**Boldface** = Model containing online information;
[a] Nowcasting model;
[b] Only insample evaluation;
[c] Rolling origin evaluation;
[d] Statistical model testing;
* BAU = Baidu, GTD = Google Trends, NAR = Naver, TWR = Twitter;
§ Pop. = Popularity, Vol. = Volume, Val. = Valence;
† d = daily, w = weekly, m = monthly, q = quarterly.

Table A5: Forecasting consumer goods

| Paper | Target Variable | Type* | Measure§ | Best model | Benchmark model(s) | Freq.† | num. series | max. lag | max. horizon | max. % err. red. |
|---|---|---|---|---|---|---|---|---|---|---|
| Brand or company level | | | | | | | | | | |
| Choi and Varian (2009)[a,c] | Car sales USA | GTD | Pop. | **sARX** | sAR | m | 27 | 0 | 1 | 12.5% MAE |
| Seebach et al. (2011)[c] | Car sales DEU | GTD | Pop. | **ARX** | AR, RW | m | 2 | 6 | 3 | 50.3% MAE |
| Barreira et al. (2013)[a,c,d] | Car sales var. countries | GTD | Pop. | **ARX** | AR | m | 4 | 0 | 36 | 8.2% RMSE |
| Carrière-Swallow and Labbé (2013)[a,c,d] | Car sales CHL | GTD | Pop. | **ARX** | AR, ARMA | m | 9 | 18d | 1 | - |
| Skodda and Benthaus (2015)[c] | Car sales DEU | GTD; TWR | Pop.; Val.tnote1 | **LR** | **LR** | m | 2 | 6 | 1 | - |
| Fantazzini and Toktamysova (2015)[c,d] | Car sales DEU | GTD | Pop. | **VECM**, **VAR**, **BVAR** | RW w. drift, AR, VARX, VECMX, BVARX, LSTAR, ESTAR, AAR | m | 22 | 12 | 24 | 66% MSE |
| Cui et al. (2017)[c,d] | Online apparel sales | FBK | Vol.; Val.[2] | **RF** | RF, LR, SVM, GB, Company Forecast | d | 1 | 7 | 7 | 20.5% MAPE |
| Jun et al. (2017) | Netbook sales | GTD; NAR | Pop.; Pop. | HW | **Analogy**, sRW, MA | q | 1 | 0 | 1 | HW better |
| Geva et al. (2017)[c,d] | Car sales USA | GTD; FOM | Pop.; Vol.; Val.[1] | **NN** | sARX, NN | m | 23 | 2 | 1 | 5.6% MAPE |
| Product level | | | | | | | | | | |
| Boone et al. (2015)[a,b] | Speciality-food SKUs sales | GTD | Pop. | **ARX** | ARX | w | 2 | 0 | - | 5.3% RMSE |
| Schneider and Gupta (2016)[c] | Tablet sales | ORW | Rat.; Val.[2] | **LR** | **SVM**, LR | w | 231 | 0 | 1 | 77.2% MAPE |
| See-To and Ngai (2016)[a,c] | Fashion sales | ORW | Val.[1] | **ARX** | MA | $\frac{1}{2}$d | 2527 | 0 | 1 | 44.2% RMSE |
| Geva et al. (2017)[c,d] | Car sales USA | GTD; FOM | Pop.; Vol.; Val.[1] | **sARX** | sARX | m | 78 | 2 | 1 | 1.9% MAPE |
| Jun et al. (2017) | Nintendo Wii sales | GTD; NAR | Pop.; Pop. | HW | **Analogy**, sRW, MA | q | 3 | 0 | 1 | HW better |
| Lau et al. (2017)[c] | E-commerce platform USA | GTD; ORW | Val.[1] | **PELM** | **LR**, **SVM**, **ELM** | w | 11428 | 1 | 1 | - |
| | E-commerce platform CHN | BAU; ORW | Val.[1] | **PELM** | **LR**, **SVM**, **ELM** | w | 8115 | 1 | 1 | - |
| Boone et al. (2018)[c] | Speciality-food SKUs sales | GTD | Pop. | **ARX** | ARX | w | 5 | 1 | 1 | 7.66% |

**Boldface** = Model containing online information;
[a] Nowcasting model;
[b] Only insample evaluation;
[c] Rolling origin evaluation;
[d] Statistical model testing;
* BAU = Baidu, FBK = Facebook, FOM = Forum, GTD = Google Trends, NAR = Naver, ORW = Online Reviews, TWR = Twitter;
§ Rat. = Rating, Pop. = Popularity, Vol. = Volume, Val. = Valence ([1] = Lexicon, [2] = Machine learning, [3] = n-grams, [4] = Self declared);
† d = daily, w = weekly, m = monthly, q = quarterly.

Table A6: Abbreviations of forecasting models

| Model | Abbreviation |
|---|---|
| AAR | Additive AR |
| ABT | AdaBoost |
| ADL | Autoregressive Distributed Lag |
| AR | Autoregressive |
| ARFIMA | AR Fractional Integrated Moving Average |
| ARIMA | AR Integrated Moving Average |
| BH | Buy and Hold strategy |
| BMA | Bayesian Model Averaging |
| BSTS | Bayesian Structural Time Series |
| BVAR | Bayesian VAR |
| DLM | Dynamic Linear Model |
| DW | Dieci and Westerhoff Model |
| ELM | Extreme Learning Machine |
| ESC | Empirical Similarity Concept |
| IAC | Housing a la Iacoviello |
| GARCH | Generalised AR Conditional Heteroskedasticity |
| GB | Gradient Boosting |
| GLM | Generalised Linear Model |
| HAR | Heterogeneous AR |
| HW | Holt Winter |
| LR | Linear Regression |
| LSTAR | Logistic Smooth Transition AR |
| MA | Moving Average |
| MSDR | Markov Switching Dynamic Regression |
| NN | Neural Network |
| PELM | Parallel co-evolutionary ELM |
| RF | Random Forest |
| RW | Random Walk |
| SETAR | Self-Exciting Threshold AR |
| SUR | Seemingly Unrelated Regression |
| SVM | Support Vector Machine |
| VAR | Vector AR |
| VECM | Vector Error Correction Model |
| VHAR | Vector Heterogeneous AR |