Forecasting the 2020 edition of the Boat Race

Rutger Lit[a,b] and Siem Jan Koopman[a,b]

[a]Time Series Lab

[b]Vrije Universiteit Amsterdam, The Netherlands

June 1, 2020

Abstract

We study the annual outcome of the Boat Race between Oxford and Cambridge and forecast the 2020 edition which was cancelled due to the COVID-19 outbreak. We find a strong presence of cyclical behaviour in the time series dynamics and model it through an autoregressive process with score-driven innovations. The inclusion of explanatory variables improve the fit of the time series further. In particular, the weight difference between the rowers in the boats of the two universities is a statistically significant predictor. All model computations are performed with the *Time Series Lab* software package and can be easily replicated.

Key words: Boat race, Time Series Lab, Unobserved components, Time Series, Forecasting, Score-driven models

1 Introduction

The Oxford and Cambridge Boat Race or "The Boat Race" is an annual rowing race between the Cambridge University Boat Club and the Oxford University Boat Club. The race takes place on the river Thames with open-weight rowing boats designed to take eight rowers and one cox. The rivalry between the Oxford and Cambridge boat is traditionally intense. The competitive spirit is at a very high level and it is common for Olympic rowers to compete in the event. The first race was in 1829 and the race is held annually since 1856 with pauses due to both world wars. However, due to the COVID-19 outbreak, the 2020 event was cancelled, something which had not occured since WWII.

As of 2019, Cambridge won the men's race 84 times and Oxford 79 times with one dead heat in 1877. We refer to `http://theboatrace.org` for more information on this event. The Boat Race has been modelled rigorously before by Mesters and Koopman (2015). In the current study, the model computations are performed using the Time Series Lab - Score Edition software package. Screen shots of the modelling steps in *Time Series Lab* are presented in the Appendix so that all reported results can be easily replicated.

2 The model

The Boat Race time series consists of either two values: 0 (Oxford Win) and 1 (Cambridge Win). We therefore assume that these observations are coming from the Bernoulli distribution with probability density function (pdf)

$$p(y_t|\pi_t) = \pi_t^{y_t}(1-\pi_t)^{1-y_t} \tag{1}$$

where π_t is a time-varying probability and $y_t \in \{0,1\}$ for $t = 1,\ldots,T$ where T is the length of the time series. We specify the unobserved time-varying probability π_t as a function of the dynamic process α_t and the regression effect $X_t\beta$, that is

$$\begin{aligned} \pi_{t+1} &= f(\theta_t), \qquad \theta_t = \alpha_t + X_t\beta, \\ \alpha_{t+1} &= \omega + \phi_1\alpha_t + \phi_2\alpha_{t-1} + \kappa s_t, \end{aligned} \tag{2}$$

where the link function $f(\cdot)$ is the logit link function so that π_t takes values between 0 and 1, that is $0 < \pi_t < 1$. The unknown coefficients include the constant ω, the autoregressive parameters ϕ_1 and ϕ_2, the updating parameter κ, and the parameters in the $k \times 1$ regression coefficient vector β. The innovations in the autoregressive process are s_t, and the explanatory variables are placed in the $1 \times k$ vector X_t, for $t = 1,\ldots,T$. All fixed coefficients are collected in the parameter vector ψ which is estimated by maximum likelihood. The autoregressive parameters ϕ_1 and ϕ_2 are constrained such that α_t is a stationary process.

The driving force behind the updating equation in (2) is the scaled score innovation s_t as given by

$$s_t = S_t \cdot \nabla_t, \qquad \nabla_t = \frac{\partial \log\, p(y_t|\pi_t, \mathcal{F}_{t-1};\psi)}{\partial\theta_t}, \tag{3}$$

for $t = 1,\ldots,T$ and where ∇_t is the score of the density $p(y_t|\pi_t, \mathcal{F}_{t-1};\psi)$.The innovation s_t can be regarded as a function of past observations. The information set $\mathcal{F}_{t-1}$ consists of lagged variables of π_t and contains exogenous variables as well. Equations (1) - (3) form a score-driven model in which the mechanism to update the time-varying parameter over time is the scaled score of the likelihood function. These so called score-driven models, or Generalized Autoregressive Score models, were proposed in the general case by Creal et al. (2013) and for time-varying location/scale volatility models by Harvey (2013).

In *Time Series Lab*, a wide range of unobserved components can be included in θ_t. Among these are non-stationary components like Trend and Seasonal, and stationary components like Autoregressive processes of order p. The score-driven framework easily takes explanatory variables into account as well, something which we will use for the modelling of the Boat Race time series. Furthermore, *Time Series Lab* allows the user to choose from several probability distributions $p(y_t|\mu_t,\sigma_t)$ but for the Boat Race time series we will only use the pdf in (1).

The score-driven approach provides a unified and consistent framework for introducing time-varying parameters in a wide class of nonlinear models. Score-driven models encompass several well-known models like the GARCH model of Engle (1982) and the ACD model of Engle and Russell (1998). Time Series Lab - Score Edition specializes in models with unobserved components driven by the score.

3 Data description

The annual Boat Race time series was downloaded from `http://theboatrace.org`. The series consist of binary variables $y_t \in \{0, 1\}$ for $t = 1, \ldots, T$ with T=191. A zero denotes a win for Oxford and a 1 denotes a win for Cambridge. Time $t = 1$ refers to the year 1829 in which the first Boat Race was held and $t = T$ is the year 2019. If the race is not held, for example due to both world wars, the corresponding observation is a missing value. The missing values are for the years 1830-1835, 1837, 1838, 1843, 1844, 1847, 1848, 1850, 1851, 1853, 1855, 1877 (dead heat), 1915-1919 (WWI) and 1940-1945 (WWII).

Explanatory variables are included in the model. We mainly follow Mesters and Koopman (2015) and include the average log difference in weight between the rowers in the Cambridge and Oxford boats, the outcome of the coin toss, and the distance by which the previous race was won. These three explanatory variables ($k = 3$) rely on information that is available just before the race starts.

4 Results

Estimation results are presented in Table 1 from which we find that the AR(2) parameters of the process (2) are estimated at $\phi_1 = 1.6434, \phi_2 = -0.9023$. An AR(2) process exhibits cyclical behaviour if $\phi_1^2 + 4\phi_2 < 0$, a condition that is satisfied for the Boat Race time series. The average period of the cycle is

$$\frac{2\pi}{\text{arc cos}(-\phi_1(1-\phi_2)/(4\phi_2))} \tag{4}$$

which is 12.01 years for the Boat Race. *Time Series Lab* provides this information and prints the following message on screen:

```
***** Information: the condition ϕ1^2 + 4ϕ2 < 0 is satisfied *****
As a result, the AR2 component for probability exhibits cyclical behaviour
with an average period of 12.01
```

The extracted probabilities are presented in Figure 1 and 2 with and without the influence of the explanatory variable respectively. The autocorrelation function of the residuals and scores are presented in Figure 3. They show no significant residual autocorrelation, a sign of a properly specified model. The regression coefficient β Diff Log W has a positive sign indicating that a heavier boat is faster. This can be explained by the fact that heavier rowers have on average more muscle mass and can have a stronger pull. This is backed up by light and heavy weight rowing divisions where the heavy weights are faster on average.

Who would have won the Boat Race in 2020?

Looking at the point forecast of the probability for 2020, there is an 80% probability that Cambridge would have won the Boat Race in 2020. It should be emphasized that this prediction is made without the effect of the weight difference of the crew since that information is not available to us. If Oxford had an weight advantage of 2kg per rower, the winning probability of Cambridge for 2020 decreased to 74%. Vice versa, if Cambridge would have the same 2 kg weight difference, their probability would increase to 86%. The probability that Cambridge would have won the 2020 edition of the Boat Race for a range of weight differences is given in Figure 4.

References

Creal, D. D., S. J. Koopman, and A. Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics 28(5)*, 777–795.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 987–1007.

Engle, R. F. and J. R. Russell (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 1127–1162.

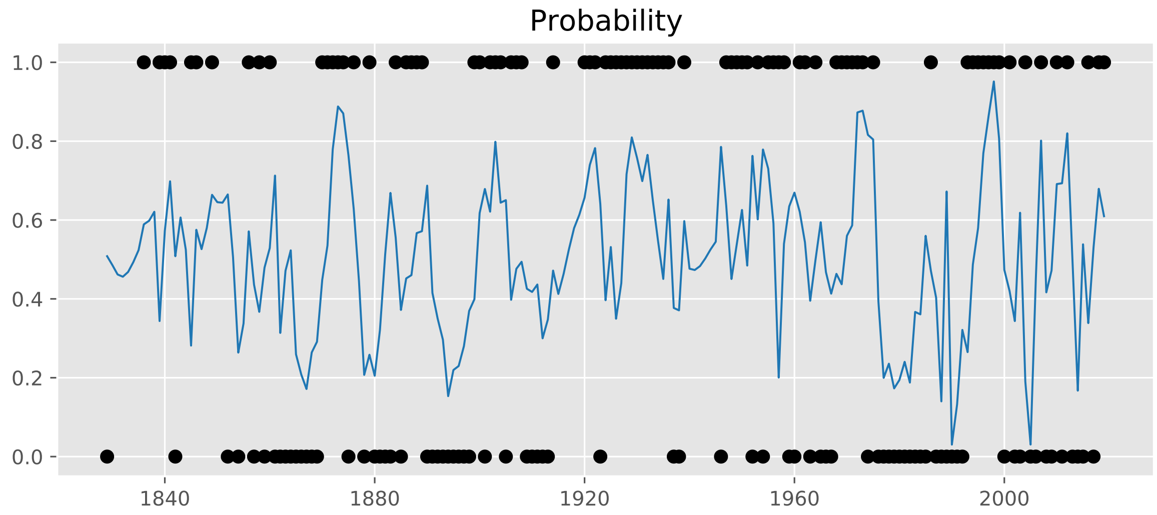
Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*, Volume 52. Cambridge: Cambridge University Press.

Lit, R., S. J. Koopman, and A. C. Harvey (2020). Time Series Lab - Score Edition. https://timeserieslab.com.

Mesters, G. and S. Koopman (2015). Forecasting the boat race. In S. J. Koopman and N. Shephard (Eds.), *Unobserved components and time series econometrics*, Chapter 7, pp. 90–114. Oxford: Oxford University Press.

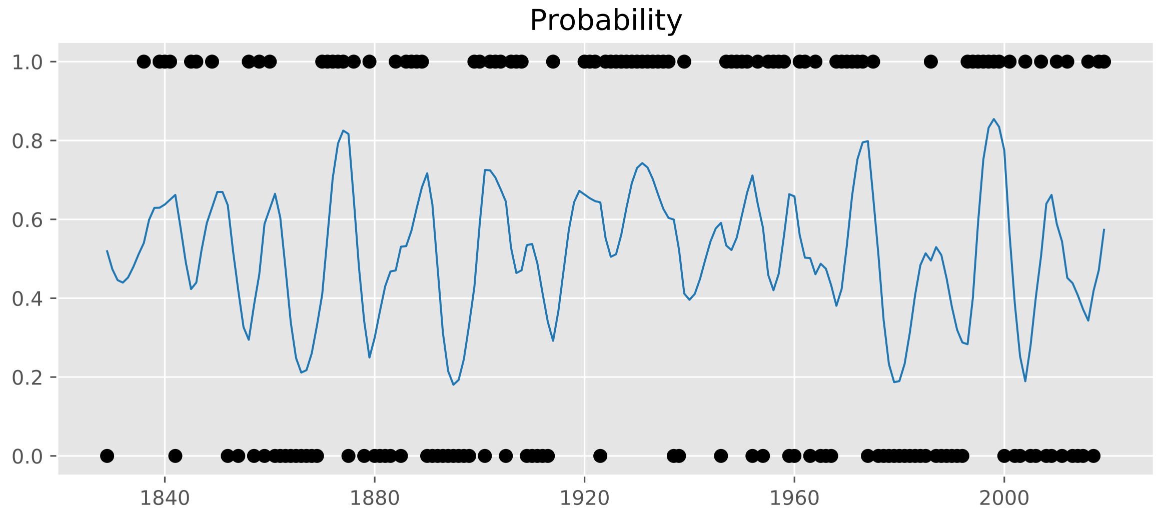
Tables and Figures

Figure 1
Extracted probability with the inclusion of $X\beta$



A dot at 0 denotes a win for Oxford and a dot at 1 denotes a win for Cambridge.

Figure 2
Extracted probability without the inclusion of $X\beta$



A dot at 0 denotes a win for Oxford and a dot at 1 denotes a win for Cambridge.

Figure 3
ACF of the Boat Race residuals and the score

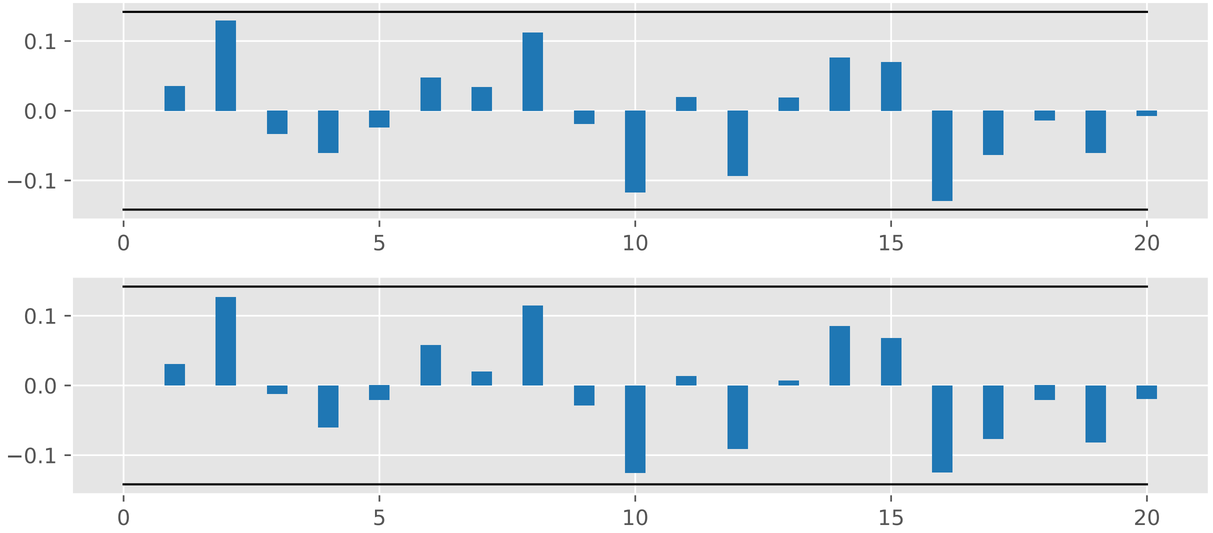


Table 1
Parameter estimates of the cyclical score-driven model

The table reports parameter estimates obtained from maximizing the likelihood function of the model in (1) - (3) with $k = 3$ explanatory variables. P-values should be taken with caution. They are directly interpretable for regression coefficients but, due to boundary issues, can give misleading results for parameters corresponding to dynamic components or variances.

Parameter type	Value	Sig.Lvl	Asymp.SE	p-value	-1.96 SE	+1.96 SE
Logit probability						
AR2 ω	0.0503		0.0782	0.5212	-0.1031	0.2036
AR2 κ	0.0763	*	0.0309	0.0144	0.0158	0.1368
AR2 ϕ_1	1.6412	***	0.0258	0.0000	1.5905	1.6919
AR2 ϕ_2	-0.8976	***	0.0720	0.0000	-1.0386	-0.7565
β Winner Toss	-0.1112		0.3639	0.7602	-0.8245	0.6020
β Diff Log W	18.2140	**	6.2221	0.0038	6.0187	30.4093
β Winning Dist..	-0.0114		0.0326	0.7268	-0.0754	0.0526

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 2
Parameter estimates of the cyclical score-driven model

The table reports parameter estimates obtained from maximizing the likelihood function of the model in (1) - (3) with $k = 1$ significant explanatory variables. P-values should be taken with caution. They are directly interpretable for regression coefficients but, due to boundary issues, can give misleading results for parameters corresponding to dynamic components or variances.

Parameter type	Value	Sig.Lvl	Asymp.SE	p-value	-1.96 SE	+1.96 SE
Logit probability						
AR2 ω	0.0232		0.0522	0.6565	-0.0790	0.1255
AR2 κ	0.0714	*	0.0284	0.0129	0.0157	0.1272
AR2 ϕ_1	1.6502	***	0.0226	0.0000	1.6059	1.6945
AR2 ϕ_2	-0.9094	***	0.0695	0.0000	-1.0456	-0.7732
β Diff Log W	18.4330	**	6.2404	0.0035	6.2018	30.6642

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 3
In-sample model fit for a range of model specifications

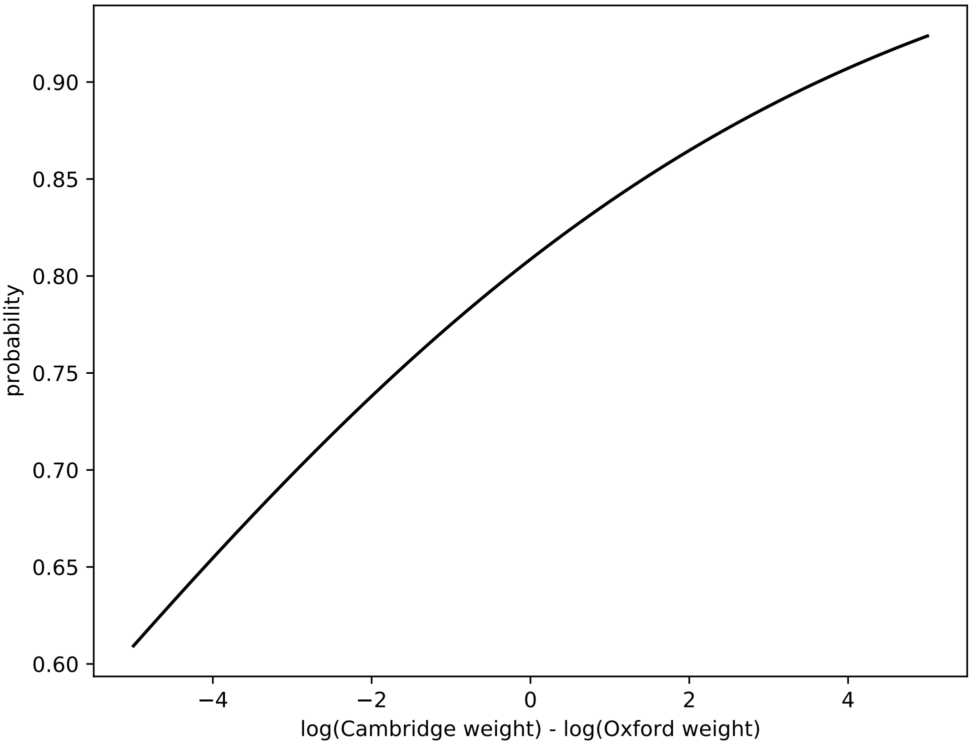
The table reports in-sample model fit for a range of model specifications. All models are estimated with Time Series Lab - Score Edition and all dynamic components are driven by the score. The Likelihood ratio (LR) test is defined as $-2\,(\text{Log L}_i - \text{Log L}_b)$ where Log L$_i$ is the smaller (nested) model of the benchmark model Log L$_b$.

Model description	Log L	# par	LR	RMSE	MAE
Distribution: Bernoulli					
Probability: logit(AR1)	-105.02	3	18.04^{***}	0.476	0.453
Probability: logit(AR1 + $X\beta$)	-100.00	4	5.86^{*}	0.458	0.421
Probability: logit(AR2)	-102.37	4	12.74^{***}	0.468	0.443
Probability: logit(AR2 + $X\beta$)	-97.07	5		0.451	0.410

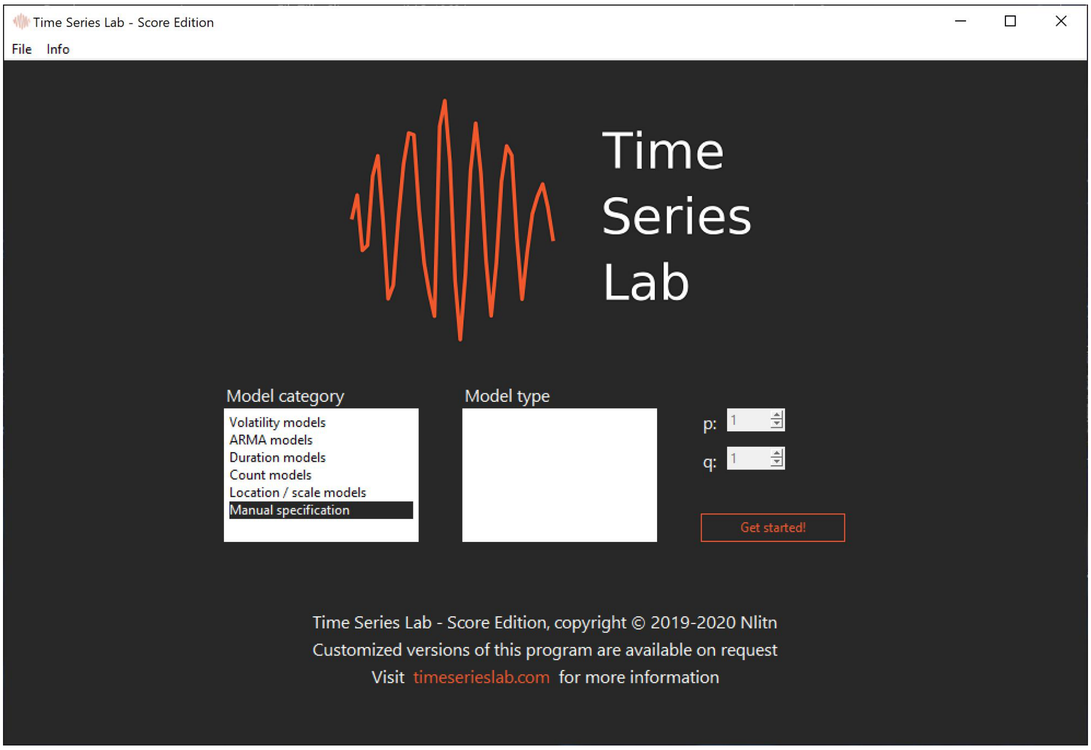
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Figure 4

Probability of Cambridge winning the Boat Race 2020



Time Series Lab - Screen Shots



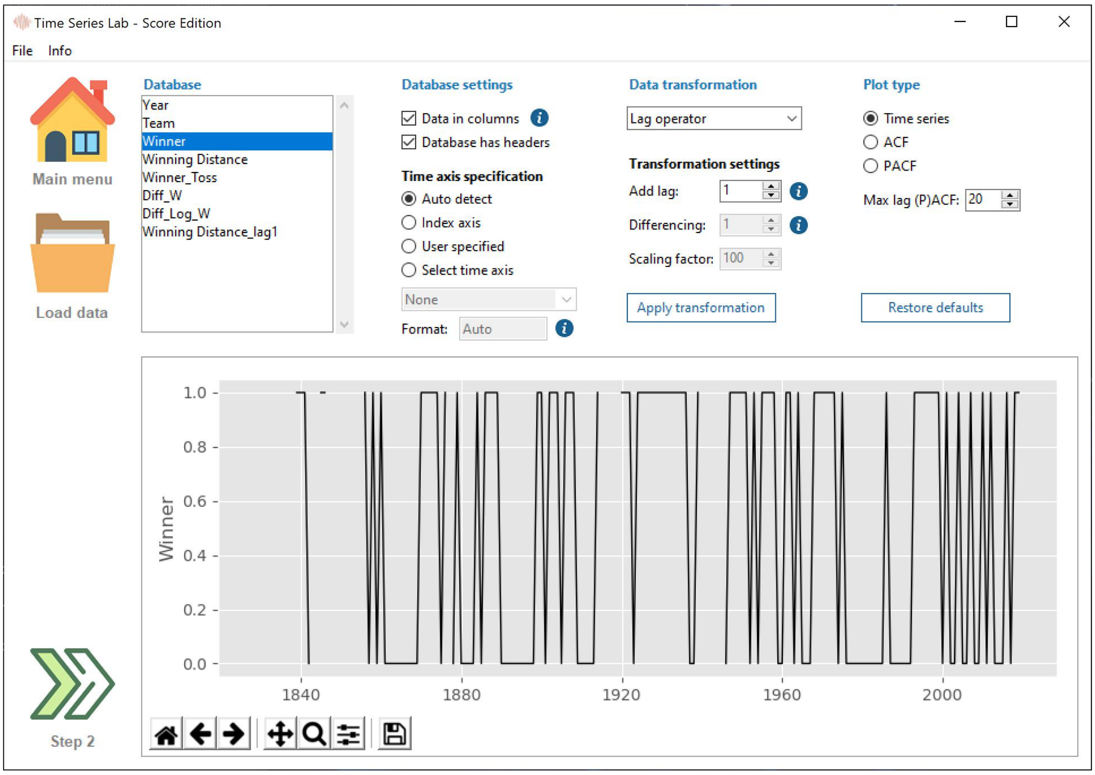
Model category
- Volatility models
- ARMA models
- Duration models
- Count models
- Location / scale models
- Manual specification

Model type

p: 1

q: 1

Time Series Lab - Score Edition, copyright © 2019-2020 Nlitn
Customized versions of this program are available on request
Visit timeserieslab.com for more information



Database
- Year
- Team
- Winner
- Winning Distance
- Winner_Toss
- Diff_W
- Diff_Log_W
- Winning Distance_lag1

Database settings

☑ Data in columns

☑ Database has headers

Time axis specification

● Auto detect

○ Index axis

○ User specified

○ Select time axis

None

Format: Auto

Data transformation

Lag operator

Transformation settings

Add lag: 1

Differencing: 1

Scaling factor: 100

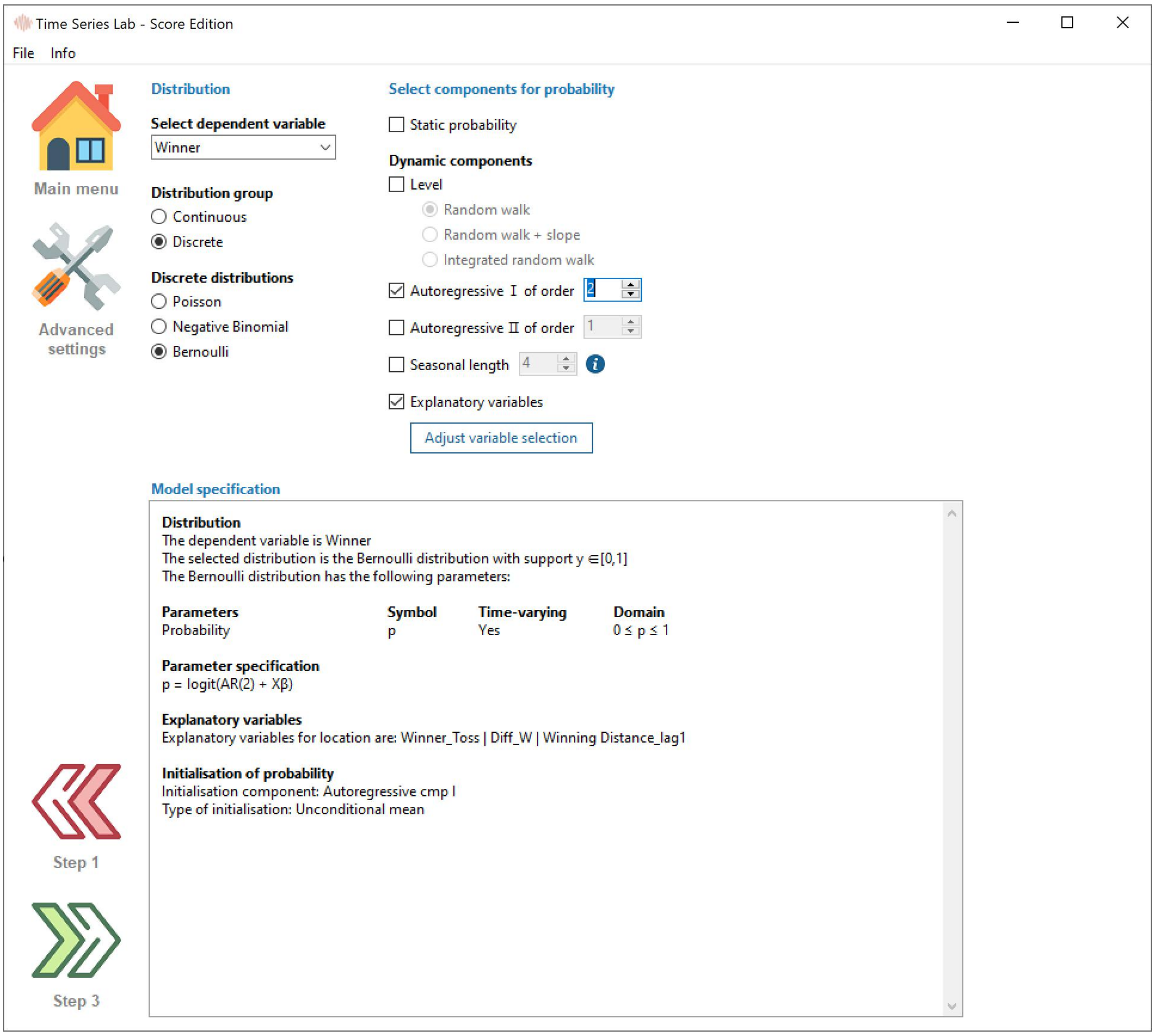
Plot type

● Time series

○ ACF

○ PACF

Max lag (P)ACF: 20



Distribution

Select dependent variable

Winner

Distribution group

○ Continuous

● Discrete

Discrete distributions

○ Poisson

○ Negative Binomial

● Bernoulli

Select components for probability

□ Static probability

Dynamic components

□ Level

● Random walk

○ Random walk + slope

○ Integrated random walk

☑ Autoregressive I of order 2

□ Autoregressive II of order 1

□ Seasonal length 4

☑ Explanatory variables

Model specification

Distribution

The dependent variable is Winner

The selected distribution is the Bernoulli distribution with support y ∈[0,1]

The Bernoulli distribution has the following parameters:

Parameters	Symbol	Time-varying	Domain
Probability	p	Yes	0 ≤ p ≤ 1

Parameter specification

p = logit(AR(2) + Xβ)

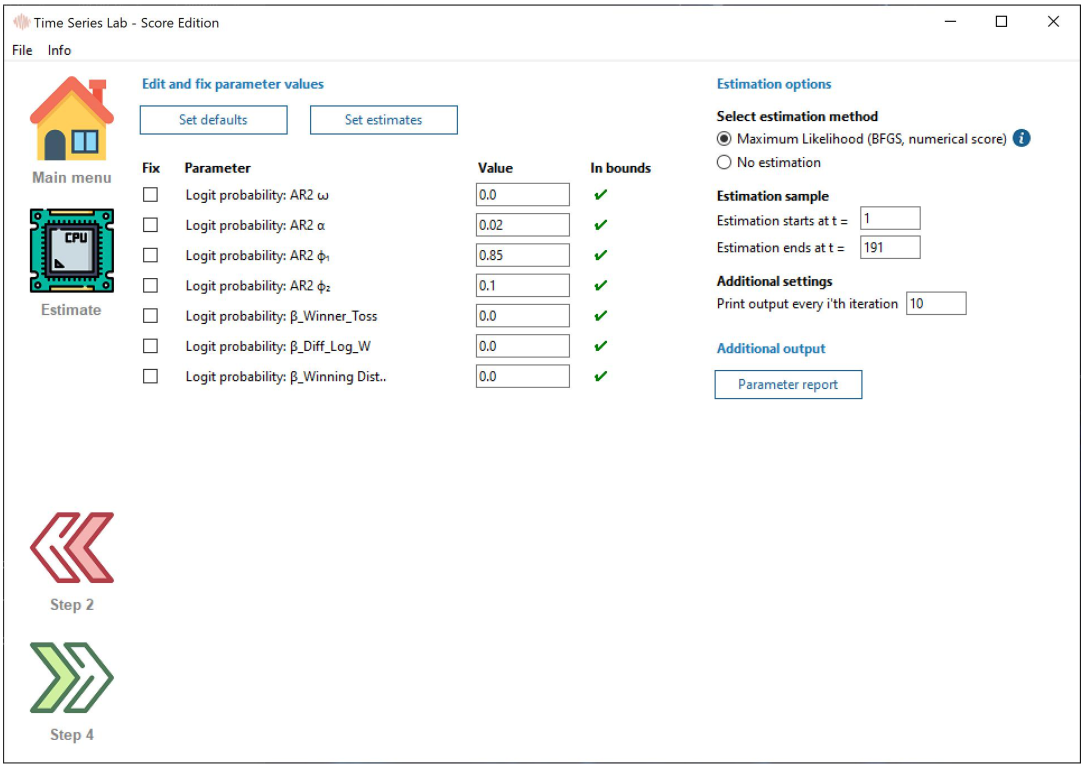
Explanatory variables

Explanatory variables for location are: Winner_Toss \| Diff_W \| Winning Distance_lag1

Initialisation of probability

Initialisation component: Autoregressive cmp I

Type of initialisation: Unconditional mean



Edit and fix parameter values

Set defaults | Set estimates

Fix	Parameter	Value	In bounds
□	Logit probability: AR2 ω	0.0	✓
□	Logit probability: AR2 α	0.02	✓
□	Logit probability: AR2 ϕ_1	0.85	✓
□	Logit probability: AR2 ϕ_2	0.1	✓
□	Logit probability: β_Winner_Toss	0.0	✓
□	Logit probability: β_Diff_Log_W	0.0	✓
□	Logit probability: β_Winning Dist..	0.0	✓

Estimation options

Select estimation method

● Maximum Likelihood (BFGS, numerical score)

○ No estimation

Estimation sample

Estimation starts at t = 1

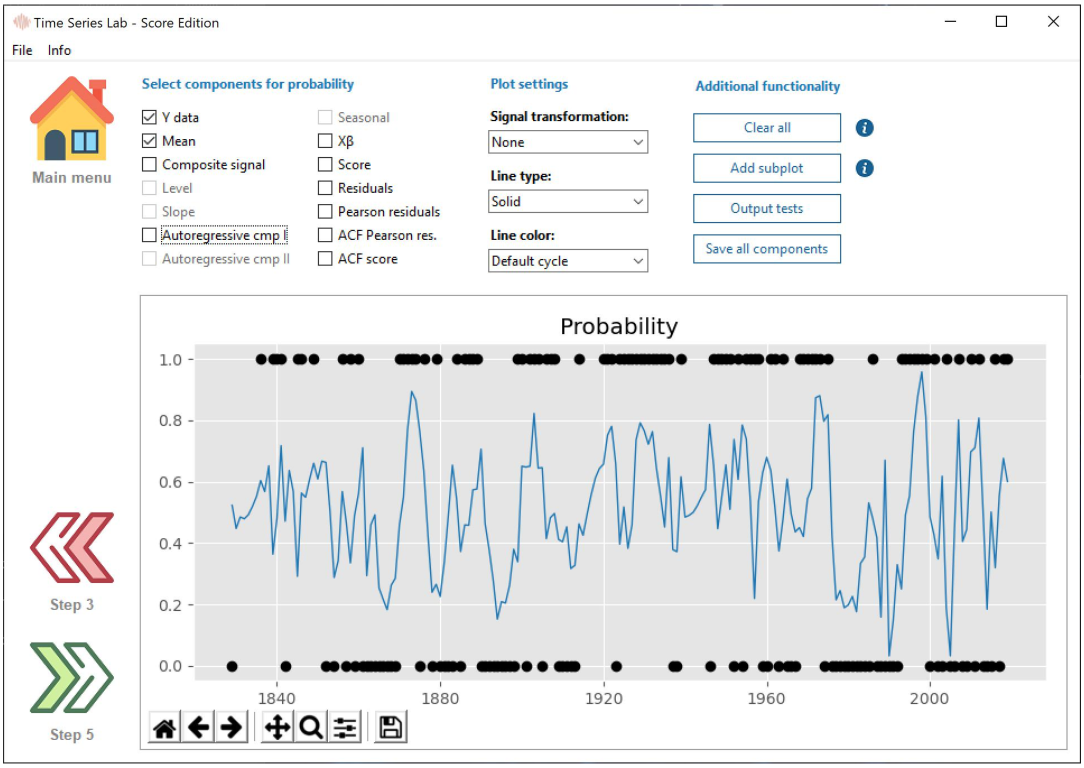
Estimation ends at t = 191

Additional settings

Print output every i'th iteration 10

Additional output

Parameter report



Select components for probability

☑ Y data

☑ Mean

□ Composite signal

□ Level

□ Slope

□ Autoregressive cmp I

□ Autoregressive cmp II

□ Seasonal

□ Xβ

□ Score

□ Residuals

□ Pearson residuals

□ ACF Pearson res.

□ ACF score

Plot settings

Signal transformation: None

Line type: Solid

Line color: Default cycle

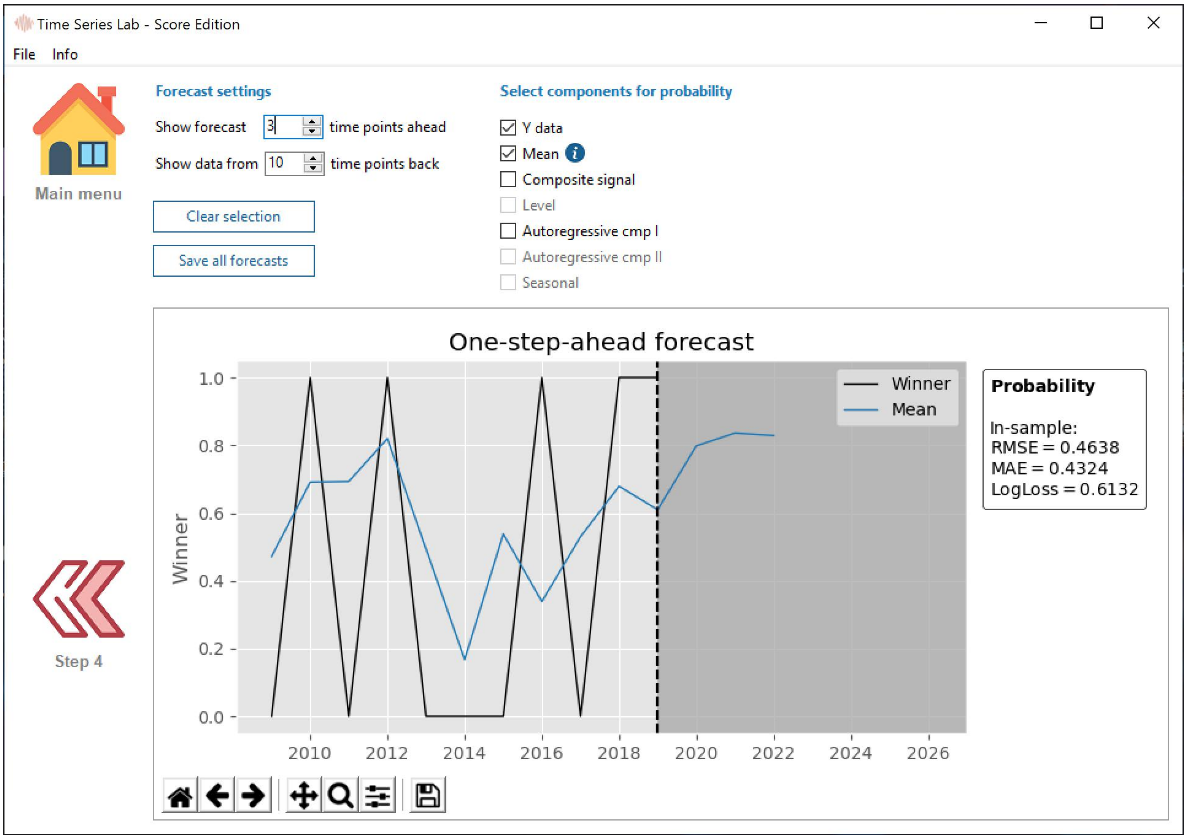
Additional functionality

Clear all

Add subplot

Output tests

Save all components



Forecast settings

Show forecast 3 time points ahead

Show data from 10 time points back

Select components for probability

☑ Y data

☑ Mean

☐ Composite signal

☐ Level

☐ Autoregressive cmp I

☐ Autoregressive cmp II

☐ Seasonal

Probability

In-sample:
RMSE = 0.4638
MAE = 0.4324
LogLoss = 0.6132