Appendix C: Monte Carlo simulation for feature selection

To find the most relevant features for each research question composite scores were constructed using principal component analysis (PCA). PCA tries to identify the space in which the data points approximately lie (Jolliffe, 2011). It computes new variables called principle components which are obtained from linear combinations of the original features. By doing so, the goal of the PCA is to extract the most important features (Abdi & Williams, 2010). To find how many principal components should be computed a subset of the training data, which included the sliding windows and additional features for consumer sentiment and the value of Y_t, was used to compute the proportion of variance explained.

The principal components with the highest weights were then used in the Monte Carlo simulation. Monte Carlo is a simulation method that relies on repeated random sampling. The algorithm creates subsets of randomly chosen features and divides the objects in each subset in train and test sets (Komorowski, 2015). For each combination of features 10-fold cross validation was performed and the Mean Squared Errors (MSE) were computed on the test set. The combination of features with the lowest test MSE score were the features considered most relevant to each research question.

C1. MSE per feature combination

Experiment	Target value	Features	MSE
1 & 3.1	$\text{change}_{bpt+3;bpt+2}$	cs_{max}, cs_{min}, cs_{median}, cs_{CQD}	1.228 E-4
2	rs_{t+3}	cs_{max}, cs_{min}, cs_{CQD}, cs_{t+1}	8.811 E-2
3.2	$\text{change}_{cit+3;cit+2}$	cs_{max}, cs_{min}, cs_{CR}, ci_t	4.255 E-2
4.1	$\text{change}_{cst+3;cst+2}$	bp_{min}, bp_{CR}, bp_{CQD}, cs_t	17.577
4.2	$\text{change}_{cst+3;cst+2}$	rs_{max}, rs_{min}, rs_t	17.637
4.3	$\text{change}_{cst+3;cst+2}$	$ci_{\bar{x}}$,, ci_{σ^2}, , ci_{t+2}, cs_t	17.456

C2. Predictions with feature selection

Experiment 1

Part	Model	MSE train	MSE test	Parameters
I	Baseline	2.721	8.675	-
	Elastic Net	1.230	3.754	*alpha = 0.2121425; lambda = 0.001205047*
	SVM	1.227	3.770	*method = eps-regression; kernel = radial; C = 1; gamma = 0.25; epsilon = 0.1; support vectors: 2591*
	Random Forest	0.993	3.440	*ntree = 5000; importance = TRUE*
II	Baseline	2.135	5.826	-
	Elastic Net	1.021	2.892	*alpha = 0.006356115; lambda = 0.006691759*
	SVM	0.983	2.830	*method = eps-regression; kernel = radial; C = 1; gamma = 0.25; epsilon = 0.1; support vectors: 316*
	Random Forest	0.809	2.680	*ntree = 5000; importance = TRUE*

Experiment 2

Part	Model	Acc. train	Acc. test	F1 train	F1 test	Parameters
I	Baseline	0.985	0.985	0.991	0.991	-
	SVM	0.902	0.895	0.944	0.940	*method = C-classification; kernel = radial; C = 1; gamma = 0.25; support vectors: 1302*
	PART	0.940	0.935	0.966	0.963	-
	Bagging	0.941	0.937	0.966	0.964	-
	RandomForest	0.941	0.938	0.966	0.964	*ntree = 5000; importance = TRUE*
	k-NN	0.940	0.938	0.965	0.964	*method = knn; trControl = cv (number: 5); k = 5*
II	Baseline	0.894	0.901	0.0217	-	-
	SVM	0.946	0.951	-	-	*method = C-classification; kernel = radial; C = 1; gamma = 0.25; support vectors: 105*
	PART	0.955	0.947	0.345	0.286	-
	Bagging	0.957	0.951	0.400	0.364	-
	RandomForest	0.957	0.951	0.400	0.364	*ntree = 5000; importance = TRUE*
	k-NN	0.953	0.947	0.286	0.211	*method = knn; trControl = cv (number: 5); k = 7*

Experiment 3.1

Cl.	Model	MSE train	MSE test	Parameters
1	Baseline	3.034	52.600	-
	Elastic Net	1.469	25.782	*alpha = 0.06905604; lambda = 0.00229073*
	SVM	1.476	26.220	*method = eps-regression; kernel = radial; C = 1; gamma = 0.25; epsilon = 0.1; support vectors: 616*
	Random Forest	1.202	23.097	*ntree = 5000; importance = TRUE*
2	Baseline	7.705	15.539	-
	Elastic Net	3.722	5.899	*alpha = 0.2426044; lambda = 6.790757*
	SVM	3.688	5.794	*method = eps-regression; kernel = radial; C = 1; gamma = 0.25; epsilon = 0.1; support vectors: 253*
	Random Forest	2.391	6.544	*ntree = 5000; importance = TRUE*
3	Baseline	0.967	0.977	-
	Elastic Net	0.444	0.437	*alpha = 0.1274343; lambda = 0.001274972*
	SVM	0.453	0.455	*method = eps-regression; kernel = radial; C = 1; gamma = 0.25; epsilon = 0.1; support vectors: 1285*
	Random Forest	0.226	0.307	*ntree = 5000; importance = TRUE*
4	Baseline	3.354	0.345	-
	Elastic Net	1.251	0.152	*alpha = 0.2827089; lambda = 0.002055782*
	SVM	1.236	0.149	*method = eps-regression; kernel = radial; C = 1; gamma = 0.25; epsilon = 0.1; support vectors: 532*
	Random Forest	0.953	0.185	*ntree = 5000; importance = TRUE*

Experiment 3.2

Cl.	Model	MSE train	MSE test	Parameters
1	Baseline	0.126	0.070	-
	Elastic Net	0.042	0.024	*alpha = 0.3073374; lambda = 0.00926526*
	SVM	0.041	0.023	*method = eps-regression; kernel = radial; C = 1; gamma = 0.25; epsilon = 0.1; support vectors: 1357*
	Random Forest	0.023	0.023	*ntree = 5000; importance = TRUE*
2	Baseline	0.144	0.134	-
	Elastic Net	0.057	0.046	*alpha = 0.8686318; lambda = 0.008117249*
	SVM	0.055	0.045	*method = eps-regression; kernel = radial; C = 1; gamma = 0.25; epsilon = 0.1; support vectors: 541*
	Random Forest	0.023	0.049	*ntree = 5000; importance = TRUE*
3	Baseline	0.124	0.187	-
	Elastic Net	0.046	0.070	*alpha = 0.4247356; lambda = 0.002565703*
	SVM	0.043	0.070	*method = eps-regression; kernel = radial; C = 1; gamma = 0.25; epsilon = 0.1; support vectors: 2188*
	Random Forest	0.025	0.072	*ntree = 5000; importance = TRUE*
4	Baseline	0.088	0.122	-
	Elastic Net	0.029	0.042	*alpha = 0.5303174; lambda = 0.007506053*
	SVM	0.029	0.042	*method = eps-regression; kernel = radial; C = 1; gamma = 0.25; epsilon = 0.1; support vectors: 1098*
	Random Forest	0.014	0.044	*ntree = 5000; importance = TRUE*

Experiment 4

Prt.	Model	MSE train	MSE test	Parameters
1	Baseline	35.284	35.291	-
	Elastic Net	17.567	17.614	*alpha = 0.1561852; lambda = 0.1073471*
	SVM	17.620	17.919	*method = eps-regression; kernel = radial; C = 1; gamma = 0.25; epsilon = 0.1; support vectors: 5185*
	Random Forest	4.578	10.578	*ntree = 5000; importance = TRUE*
2	Baseline	35.284	35.291	-
	Elastic Net	17.623	17.743	*alpha = 0.09788889; lambda = 0.005984826*
	SVM	17.660	17.795	*method = eps-regression; kernel = radial; C = 1; gamma = 0.25; epsilon = 0.1; support vectors: 5329*
	Random Forest	17.661	17.711	*ntree = 5000; importance = TRUE*
3	Baseline	35.284	35.291	-
	Elastic Net	17.432	17.526	*alpha = 0.2203477; lambda = 0.003436149*
	SVM	16.919	17.433	*method = eps-regression; kernel = radial; C = 1; gamma = 0.25; epsilon = 0.1; support vectors: 5235*
	Random Forest	2.742	11.486	*ntree = 5000; importance = TRUE*

Appendix D: States in state clusters

Cluster	Train set	Test set
1. Financial cluster	Connecticut, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Rhode Island, Vermont, Virginia	California
2. Oil cluster	Louisiana, North Dakota, Oklahoma, Texas	Alaska, Wyoming, New Mexico
3. Manufacturing cluster	Alabama, Illinois, Indiana, Iowa, Kansas, Kentucky, Michigan, Minnesota, Mississippi, Missouri, Ohio, Pennsylvania, South Carolina, Tennessee, West Virginia, Wisconsin	Washington, Montana
4. Mixed economy cluster	Arkansas, Delaware, Florida, Georgia, Hawaii, Nebraska, North Carolina, South Dakota	Oregon, Idaho, Colorado, Nevada, Arizona, Utah

Appendix E: Multiple Imputation

Variable	Imputation method
sixmonthsout	Logistic regression
Oil price	Bayesian linear regression
Oil state	PMM
Agriculture	PMM
Mining	PMM
Construction	PMM
Manifacutring	PMM
Durable goods	PMM
Nondurable goods	PMM
Current coincident index	PMM

Appendix F: Results of experiment 1

Experiment 1.1

RQ1.1

Feature	Coefficient
Intercept	0.00230
cs_t	0.00018
cs_{t+1}	0.00006
cs_{t+2}	0.00004
cs_{max}	0.00006
cs_{min}	0.00001
cs_{σ^2}	0.00092
$cs_{\bar{x}}$	0.00011
cs_{median}	0.00023
cs_{CR}	-0.00019
cs_{CQD}	-0.00029
bp_t	0.00229

Elastic Net coefficients RQ1.1

RQ1.1

Feature	Importance
cs_t	7.460
cs_{t+1}	2.219
cs_{t+2}	1.492
cs_{max}	2.403
cs_{min}	-
cs_{σ^2}	40.009
$cs_{\bar{x}}$	4.208
cs_{median}	9.682
cs_{CR}	7.670
cs_{CQD}	12.276
bp_t	100.000

Elastic Net feature importance RQ1.1

RQ1.1

Feature	Importance
cs_t	71.211
cs_{t+1}	81.072
cs_{t+2}	79.080
cs_{max}	65.682
cs_{min}	67.426
cs_{σ^2}	98.415
$cs_{\bar{x}}$	73.329
cs_{median}	73.616
cs_{CR}	85.855
cs_{CQD}	89.923
bp_t	70.350

Random Forest feature importance RQ1.1

Experiment 1.2

RQ1.2

Feature	Coefficient
Intercept	-0.00012
cs_t	-
cs_{t+1}	-
cs_{t+2}	-
cs_{max}	-
cs_{min}	-
cs_{σ^2}	-
$cs_{\bar{x}}$	-
cs_{median}	-
cs_{CR}	-
cs_{CQD}	-
bp_t	-

Elastic Net coefficients RQ1.2

RQ1.2

Feature	Importance
cs_t	-
cs_{t+1}	-
cs_{t+2}	-
cs_{max}	-
cs_{min}	-
cs_{σ^2}	-
$cs_{\bar{x}}$	-
cs_{median}	-
cs_{CR}	-
cs_{CQD}	-
bp_t	-

Elastic Net feature importance RQ1.2

RQ1.2

Feature	Importance
cs_t	30.655
cs_{t+1}	20.624
cs_{t+2}	46.361
cs_{max}	21.741
cs_{min}	31.887
cs_{σ^2}	28.235
$cs_{\bar{x}}$	22.810
cs_{median}	23.275
cs_{CR}	25.958
cs_{CQD}	28.758
bp_t	-6.333

Random Forest feature importance RQ1.2

Appendix G: Results of experiment 2

Experiment 2.1

RQ2.1	
Feature	**Importance**
cs_t	0
cs_{t+1}	0
cs_{t+2}	0
cs_{max}	0
cs_{min}	1
cs_{σ^2}	0
$cs_{\bar{x}}$	0
cs_{median}	1
cs_{CR}	1
cs_{CQD}	0
rst	1

582PART feature importance RQ2.1

RQ2.1	
Feature	**Importance**
cs_t	14.733
cs_{t+1}	17.830
cs_{t+2}	17.781
cs_{max}	547.502
cs_{min}	571.022
cs_{σ^2}	17.582
$cs_{\bar{x}}$	532.179
cs_{median}	476.344
cs_{CR}	18.150
cs_{CQD}	22.460
rst	1277.657

Bagging feature importance RQ2.1

RQ2.1	
Feature	**Importance**
cs_t	28.899
cs_{t+1}	29.911
cs_{t+2}	28.133
cs_{max}	107.108
cs_{min}	138.206
cs_{σ^2}	18.001
$cs_{\bar{x}}$	88.650
cs_{median}	59.538
cs_{CR}	20.517
cs_{CQD}	19.236
rst	773.681

Random Forest feature importance RQ2.1

RQ2.1	
Feature	**Importance**
cs_t	56.60
cs_{t+1}	59.05
cs_{t+2}	60.53
cs_{max}	61.21
cs_{min}	61.89
cs_{σ^2}	-
$cs_{\bar{x}}$	62.54
cs_{median}	62.13
cs_{CR}	27.07
cs_{CQD}	27.43
rst	100.00

k-NN feature importance RQ2.1

Experiment 2.2

RQ2.2	
Feature	**Importance**
cs_t	0
cs_{t+1}	0
cs_{t+2}	0
cs_{max}	0
cs_{min}	1
cs_{σ^2}	0
$cs_{\bar{x}}$	0
cs_{median}	1
cs_{CR}	1
cs_{CQD}	0
rst	0

PART feature importance RQ2.2

RQ2.2	
Feature	**Importance**
cs_t	11.020
cs_{t+1}	12.405
cs_{t+2}	9.641
cs_{max}	16.998
cs_{min}	16.407
cs_{σ^2}	15.325
$cs_{\bar{x}}$	7.106
cs_{median}	5.312
cs_{CR}	18.763
cs_{CQD}	17.881
rst	-

Bagging feature importance RQ2.2

RQ2.2	
Feature	**Importance**
cs_t	2.389
cs_{t+1}	2.677
cs_{t+2}	2.320
cs_{max}	2.548
cs_{min}	2.948
cs_{σ^2}	5.193
$cs_{\bar{x}}$	2.647
cs_{median}	2.839
cs_{CR}	5.472
cs_{CQD}	5.569
rst	-

Random Forest feature importance RQ2.2

RQ2.2	
Feature	**Importance**
cs_t	82.26
cs_{t+1}	78.36
cs_{t+2}	80.73
cs_{max}	85.33
cs_{min}	93.34
cs_{σ^2}	79.34
$cs_{\bar{x}}$	87.90
cs_{median}	85.19
cs_{CR}	99.08
cs_{CQD}	100.00
rst	-

k-NN feature importance RQ2.2

Appendix H: Results of experiment 3

Experiment 3.1

RQ3.1

Feature	Coefficient
Intercept	0.00248
cs_t	-
cs_{t+1}	-
cs_{t+2}	0.00042
cs_{max}	-
cs_{min}	-
cs_{σ^2}	0.00020
$cs_{\bar{x}}$	-
cs_{median}	-
cs_{CR}	-
cs_{CQD}	-
bp_t	0.00256

Elastic Net coefficients RQ3.1 - Cluster 1

RQ3.1

Feature	Importance
cs_t	-
cs_{t+1}	-
cs_{t+2}	16.475
cs_{max}	-
cs_{min}	-
cs_{σ^2}	7.922
$cs_{\bar{x}}$	-
cs_{median}	-
cs_{CR}	-
cs_{CQD}	-
bp_t	100.000

Elastic Net feature importance RQ3.1 - Cluster 1

RQ3.1

Feature	Importance
cs_t	39.455
cs_{t+1}	51.794
cs_{t+2}	53.928
cs_{max}	41.509
cs_{min}	39.952
cs_{σ^2}	44.923
$cs_{\bar{x}}$	44.720
cs_{median}	43.813
cs_{CR}	42.749
cs_{CQD}	45.261
bp_t	52.540

Random Forest feature importance RQ3.1 - Cluster 1

RQ3.1

Feature	Coefficient
Intercept	0.00388
cs_t	0.00033
cs_{t+1}	0.00028
cs_{t+2}	-
cs_{max}	0.00023
cs_{min}	-
cs_{σ^2}	0.00232
$cs_{\bar{x}}$	0.00015
cs_{median}	0.00036
cs_{CR}	-0.00022
cs_{CQD}	-0.00037
bp_t	0.00443

Elastic Net coefficients RQ3.1 - Cluster 2

RQ3.1

Feature	Importance
cs_t	7.463
cs_{t+1}	6.301
cs_{t+2}	-
cs_{max}	5.215
cs_{min}	-
cs_{σ^2}	52.475
$cs_{\bar{x}}$	3.374
cs_{median}	8.107
cs_{CR}	5.114
cs_{CQD}	8.248
bp_t	100.000

Elastic Net feature importance RQ3.1 - Cluster 2

RQ3.1

Feature	Importance
cs_t	36.917
cs_{t+1}	32.903
cs_{t+2}	36.041
cs_{max}	36.177
cs_{min}	34.714
cs_{σ^2}	34.773
$cs_{\bar{x}}$	41.640
cs_{median}	34.401
cs_{CR}	35.380
cs_{CQD}	36.010
bp_t	3.209

Random Forest feature importance RQ3.1 - Cluster 2

RQ3.1

Feature	Coefficient
Intercept	0.00201
cs_t	0.00014
cs_{t+1}	0.00009
cs_{t+2}	-
cs_{max}	0.00008
cs_{min}	0.00005
cs_{σ^2}	0.00004
$cs_{\bar{x}}$	0.00008
cs_{median}	0.00010
cs_{CR}	-
cs_{CQD}	-
bp_t	0.00056

Elastic Net coefficients RQ3.1 - Cluster 3

RQ3.1

Feature	Importance
cs_t	24.026
cs_{t+1}	16.636
cs_{t+2}	-
cs_{max}	14.143
cs_{min}	9.507
cs_{σ^2}	6.811
$cs_{\bar{x}}$	14.405
cs_{median}	18.528
cs_{CR}	-
cs_{CQD}	-
bp_t	100.000

Elastic Net feature importance RQ3.1 - Cluster 3

RQ3.1

Feature	Importance
cs_t	89.817
cs_{t+1}	86.980
cs_{t+2}	88.329
cs_{max}	72.722
cs_{min}	80.380
cs_{σ^2}	120.125
$cs_{\bar{x}}$	83.692
cs_{median}	80.356
cs_{CR}	113.053
cs_{CQD}	117.660
bp_t	117.701

Random Forest feature importance RQ3.1 - Cluster 3

RQ3.1

Feature	Coefficient
Intercept	0.00187
cs_t	0.00016
cs_{t+1}	0.00019
cs_{t+2}	0.00003
cs_{max}	0.00015
cs_{min}	0.00002
cs_{σ^2}	0.00059
$cs_{\bar{x}}$	0.00014
cs_{median}	0.00027
cs_{CR}	-
cs_{CQD}	-
bp_t	0.00147

Elastic Net coefficients RQ3.1 - Cluster 4

RQ3.1

Feature	Importance
cs_t	10.913
cs_{t+1}	13.026
cs_{t+2}	2.117
cs_{max}	9.991
cs_{min}	1.079
cs_{σ^2}	40.204
$cs_{\bar{x}}$	9.674
cs_{median}	18.041
cs_{CR}	-
cs_{CQD}	-
bp_t	100.000

Elastic Net feature importance RQ3.1 - Cluster 4

RQ3.1

Feature	Importance
cs_t	25.177
cs_{t+1}	32.803
cs_{t+2}	30.940
cs_{max}	22.138
cs_{min}	28.151
cs_{σ^2}	36.063
$cs_{\bar{x}}$	22.662
cs_{median}	27.743
cs_{CR}	30.797
cs_{CQD}	33.887
bp_t	9.663

Random Forest feature importance RQ3.1 - Cluster 4

Experiment 3.2

RQ3.2	
Feature	**Coefficient**
Intercept	0.00126
cs_t	-
cs_{t+1}	-
cs_{t+2}	-
cs_{max}	-
cs_{min}	-
cs_{σ^2}	-
$cs_{\bar{x}}$	-
cs_{median}	-
cs_{CR}	-
cs_{CQD}	-
ci_t	-0.01594

Elastic Net coefficients RQ3.2 - Cluster 1

RQ3.2	
Feature	**Importance**
cs_t	-
cs_{t+1}	-
cs_{t+2}	-
cs_{max}	-
cs_{min}	-
cs_{σ^2}	-
$cs_{\bar{x}}$	-
cs_{median}	-
cs_{CR}	-
cs_{CQD}	-
ci_t	100.000

Elastic Net feature importance RQ3.2 - Cluster 1

RQ3.2	
Feature	**Importance**
cs_t	39.896
cs_{t+1}	36.362
cs_{t+2}	46.558
cs_{max}	36.385
cs_{min}	37.466
cs_{σ^2}	37.892
$cs_{\bar{x}}$	40.229
cs_{median}	39.030
cs_{CR}	36.579
cs_{CQD}	34.459
ci_t	58.357

Random Forest feature importance RQ3.2 - Cluster 1

RQ3.2	
Feature	**Coefficient**
Intercept	-0.00041
cs_t	-
cs_{t+1}	-0.04367
cs_{t+2}	0.03794
cs_{max}	0.00192
cs_{min}	0.06428
cs_{σ^2}	0.05880
$cs_{\bar{x}}$	-
cs_{median}	-0.07384
cs_{CR}	-0.04943
cs_{CQD}	-
ci_t	-0.03451

Elastic Net coefficients RQ3.2 - Cluster 2

RQ3.2	
Feature	**Importance**
cs_t	-
cs_{t+1}	59.145
cs_{t+2}	51.377
cs_{max}	2.594
cs_{min}	87.059
cs_{σ^2}	79.630
$cs_{\bar{x}}$	-
cs_{median}	100.000
cs_{CR}	66.944
cs_{CQD}	-
ci_t	46.740

Elastic Net feature importance RQ3.2 - Cluster 2

RQ3.2	
Feature	**Importance**
cs_t	24.080
cs_{t+1}	23.418
cs_{t+2}	21.541
cs_{max}	19.196
cs_{min}	18.977
cs_{σ^2}	22.545
$cs_{\bar{x}}$	22.346
cs_{median}	22.986
cs_{CR}	22.543
cs_{CQD}	23.462
ci_t	19.543

Random Forest feature importance RQ3.2 - Cluster 2

RQ3.2

Feature	Coefficient
Intercept	0.00018
cs_t	-0.00203
cs_{t+1}	-0.00104
cs_{t+2}	-
cs_{max}	-
cs_{min}	-
cs_{σ^2}	0.05051
$cs_{\bar{x}}$	-
cs_{median}	-
cs_{CR}	-
cs_{CQD}	-0.05214
ci_t	-0.04133

Elastic Net coefficients RQ3.2 - Cluster 3

RQ3.2

Feature	Importance
cs_t	3.895
cs_{t+1}	1.994
cs_{t+2}	-
cs_{max}	-
cs_{min}	-
cs_{σ^2}	96.871
$cs_{\bar{x}}$	-
cs_{median}	-
cs_{CR}	-
cs_{CQD}	100.000
ci_t	79.270

Elastic Net feature importance RQ3.2 - Cluster 3

RQ3.2

Feature	Importance
cs_t	62.331
cs_{t+1}	53.597
cs_{t+2}	61.540
cs_{max}	62.584
cs_{min}	58.320
cs_{σ^2}	42.867
$cs_{\bar{x}}$	63.918
cs_{median}	60.370
cs_{CR}	52.421
cs_{CQD}	52.225
ci_t	131.339

Random Forest feature importance RQ3.2 - Cluster 3

RQ3.2

Feature	Coefficient
Intercept	0.00060
cs_t	-
cs_{t+1}	0.00265
cs_{t+2}	-
cs_{max}	-
cs_{min}	-
cs_{σ^2}	-
$cs_{\bar{x}}$	-
cs_{median}	-
cs_{CR}	-
cs_{CQD}	-
ci_t	-0.01978

Elastic Net coefficients RQ3.2 - Cluster 4

RQ3.2

Feature	Importance
cs_t	-
cs_{t+1}	-
cs_{t+2}	13.37
cs_{max}	-
cs_{min}	-
cs_{σ^2}	-
$cs_{\bar{x}}$	-
cs_{median}	-
cs_{CR}	-
cs_{CQD}	-
ci_t	100.000

Elastic Net feature importance RQ3.2 - Cluster 4

RQ3.2

Feature	Importance
cs_t	55.299
cs_{t+1}	57.078
cs_{t+2}	46.606
cs_{max}	49.063
cs_{min}	52.805
cs_{σ^2}	50.079
$cs_{\bar{x}}$	52.304
cs_{median}	56.769
cs_{CR}	48.915
cs_{CQD}	48.120
ci_t	53.226

Random Forest feature importance RQ3.2 - Cluster 4

Appendix I: Experiment 3.1 with 70/30 partitioning

Experiment 3.1 (10^{-4})

Cl.	**Model**	**MSE train**		**MSE test**		**Parameters**
1.	Baseline	6.382	-	8.053	-	-
	Elastic Net	2.820	(55.81%)	4.382	(45.56%)	*alpha* = 0.2954246; *lambda* = 0.001006681
	SVM	3.226	(49.45%)	4.572	(43.23%)	*method = eps-regression; kernel = radial; C = 1; gamma = 0.09090909; epsilon = 0.1; support vectors: 468*
	Random Forest	0.890	(86.05%)	0.433	(94.62%)	*ntree* = 5000; *importance* = TRUE
2.	Baseline	15.766	-	4.126	-	-
	Elastic Net	5.470	(65.31%)	1.928	(53.27%)	*alpha* = 0.1503563; *lambda* = 0.001694307
	SVM	5.524	(64.96%)	2.065	(49.95%)	*method = eps-regression; kernel = radial; C = 1; gamma = 0.09090909; epsilon = 0.1; support vectors: 233*
	Random Forest	2.472	(84.32%)	2.915	(29.35%)	*ntree* = 5000; *importance* = TRUE
3.	Baseline	0.963	-	0.846	-	-
	Elastic Net	0.433	(55.04%)	0.405	(52.13%)	*alpha* = 0.02349696; *lambda* = 0.001108925
	SVM	0.456	(52.65%)	0.431	(49.05%)	*method = eps-regression; kernel = radial; C = 1; gamma = 0.09090909; epsilon = 0.1; support vectors: 1001*
	Random Forest	0.112	(88.37%)	0.201	(76.24%)	*ntree* = 5000; *importance* = TRUE
4.	Baseline	2.843	-	0.696	-	-
	Elastic Net	0.920	(67.64%)	0.320	(54.02%)	*alpha* = 0.1480629; *lambda* = 0.001153624
	SVM	0.928	(67.36%)	0.329	(52.73%)	*method = eps-regression; kernel = radial; C = 1; gamma = 0.09090909; epsilon = 0.1; support vectors: 731*
	Random Forest	0.479	(83.15%)	0.331	(52.44%)	*ntree* = 5000; *importance* = TRUE

Appendix J: Experiment 3.2 with 70/30 partitioning

Experiment 3.2

Cl.	Model	MSE train		MSE test		Parameters
1.	Baseline	0.113	-	0.095	-	-
	Elastic Net	0.041	(63.72%)	0.040	(57.89%)	*alpha* = 0.03792504; *lambda* = 0.009532534
	SVM	0.040	(64.60%)	0.040	(57.89%)	*method = eps-regression; kernel = radial; C = 1; gamma = 0.09090909; epsilon = 0.1; support vectors: 1063*
	Random Forest	0.021	(81.41%)	0.043	(54.74%)	*ntree* = 5000; *importance* = TRUE
2.	Baseline	0.119	-	0.159	-	-
	Elastic Net	0.046	(61.34%)	0.069	(56.60%)	*alpha* = 0.8567382; *lambda* = 0.008085317
	SVM	0.044	(63.03%)	0.070	(55.97%)	*method = eps-regression; kernel = radial; C = 1; gamma = 0.09090909; epsilon = 0.1; support vectors: 694*
	Random Forest	0.020	(83.19%)	0.074	(53.46%)	*ntree* = 5000; *importance* = TRUE
3.	Baseline	0.124	-	0.095	-	-
	Elastic Net	0.050	(59.68%)	0.044	(53.68%)	*alpha* = 0.07239164; *lambda* = 0.003337746
	SVM	0.047	(62.10%)	0.043	(54.74%)	*method = eps-regression; kernel = radial; C = 1; gamma = 0.09090909; epsilon = 0.1; support vectors: 1684*
	Random Forest	0.026	(79.03%)	0.044	(53.68%)	*ntree* = 5000; *importance* = TRUE
4.	Baseline	0.090	-	0.076	-	-
	Elastic Net	0.034	(62.22%)	0.035	(53.95%)	*alpha* = 0.01827431; *lambda* = 0.006731816
	SVM	0.033	(63.33%)	0.035	(53.95%)	*method = eps-regression; kernel = radial; C = 1; gamma = 0.09090909; epsilon = 0.1; support vectors: 1306*
	Random Forest	0.019	(78.89%)	0.037	(51.32%)	*ntree* = 5000; *importance* = TRUE

Appendix K: Results of experiment 4

Experiment 4.1

RQ4.1

Feature	Coefficient
Intercept	0.03949
bp_t	-
bp_{t+1}	0.00289
bp_{t+2}	0.00226
bp_{max}	-
bp_{min}	-
bp_{σ^2}	-
$bp_{\bar{x}}$	-
bp_{median}	-
bp_{CR}	-0.05817
bp_{CQD}	0.00692
cs_t	-0.30534

Elastic Net coefficients RQ4.1

RQ4.1

Feature	Importance
bp_t	-
bp_{t+1}	0.948
bp_{t+2}	0.741
bp_{max}	-
bp_{min}	-
bp_{σ^2}	-
$bp_{\bar{x}}$	-
bp_{median}	-
bp_{CR}	19.050
bp_{CQD}	2.268
cs_t	100.000

Elastic Net feature importance RQ4.1

RQ4.1

Feature	Importance
bp_t	93.033
bp_{t+1}	93.779
bp_{t+2}	90.071
bp_{max}	87.340
bp_{min}	87.703
bp_{σ^2}	142.421
$bp_{\bar{x}}$	93.767
bp_{median}	94.672
bp_{CR}	102.711
bp_{CQD}	109.283
cs_t	258.137

Random Forest feature importance RQ4.1

Experiment 4.2

RQ4.2

Feature	Coefficient
Intercept	0.03949
rs_t	0.35703
rs_{t+1}	0.00248
rs_{t+2}	-0.54901
rs_{max}	-
rs_{min}	-
rs_{σ^2}	-0.10267
$rs_{\bar{x}}$	-
rs_{median}	0.01632
rs_{CR}	-0.06562
rs_{CQD}	-
cs_t	-0.45370

Elastic Net coefficients RQ4.2

RQ4.2

Feature	Importance
rs_t	65.032
rs_{t+1}	0.452
rs_{t+2}	100.000
rs_{max}	-
rs_{min}	-
rs_{σ^2}	18.699
$rs_{\bar{x}}$	-
rs_{median}	2.973
rs_{CR}	11.952
rs_{CQD}	-
cs_t	82.639

Elastic Net feature importance RQ4.2

RQ4.2

Feature	Importance
rs_t	35.366
rs_{t+1}	14.968
rs_{t+2}	39.025
rs_{max}	25.026
rs_{min}	18.867
rs_{σ^2}	34.862
$rs_{\bar{x}}$	28.787
rs_{median}	14.901
rs_{CR}	34.796
rs_{CQD}	24.016
cs_t	77.292

Random Forest feature importance RQ4.2

Experiment 4.3

RQ4.3	
Feature	**Coefficient**
Intercept	0.03949
ci_t	-0.33595
ci_{t+1}	-0.16706
ci_{t+2}	0.58840
ci_{max}	-0.00005
ci_{min}	0.01761
ci_{σ^2}	0.13447
$ci_{\bar{x}}$	-
ci_{median}	-0.01800
ci_{CR}	-0.02495
ci_{CQD}	0.01279
cs_t	-0.38059

Elastic Net coefficients RQ4.3

RQ4.3	
Feature	**Importance**
ci_t	57.096
ci_{t+1}	28.393
ci_{t+2}	100.000
ci_{max}	0.0084
ci_{min}	2.992
ci_{σ^2}	22.854
$ci_{\bar{x}}$	-
ci_{median}	3.059
ci_{CR}	4.241
ci_{CQD}	2.174
cs_t	64.682

Elastic Net feature importance RQ4.3

RQ4.3	
Feature	**Importance**
ci_t	116.425
ci_{t+1}	107.359
ci_{t+2}	115.332
ci_{max}	99.967
ci_{min}	111.285
ci_{σ^2}	108.662
$ci_{\bar{x}}$	100.335
ci_{median}	107.319
ci_{CR}	105.841
ci_{CQD}	103.137
cs_t	220.702

Random Forest feature importance RQ4.3

Appendix L: Sliding window approach

	Window 1	Window 2	Window 3	Window 4	Window 5
Year	2005 to 2018	2005 to 2018	2005 to 2018	2005 to 2017	2005 to 2017
Month 1	January	February	March	April	May
Month 2	February	March	April	May	June
Month 3	March	April	May	June	July
	Window 6	Window 7	Window 8	Window 9	Window 10
Year	2005 to 2017	2005 to 2017	2005 to 2017	2005 to 2017	2005 to 2017
Month 1	June	July	August	September	October
Month 2	July	August	September	October	November
Month 3	August	September	October	November	December

Appendix M – Software and packages

This appendix provides a global overview of the packages that were used in the experimental procedure. All analyses and experiments were implemented using R Studio.

mice. The "mice" package is used to perform multiple imputation. The *mice::quickpred()* function is used to quick select predictors from the data. The *mincor* parameter that specifies the minimum threshold is set to 0.25. The *mice::mice* is used to replace the missing values. The parameter of the number of imputations *m* is set to 1 with the number of iterations *maxit* set to 1 too. The seed is set to '314159'. *Mice::complete* extracts the subset of complete cases. **TSPred.** "TSPred" is used for the sliding window method. *TSPred::slidingWindows* extracts all possible subsequences of a time series. The parameter *swSize* is set to 3. **zoo.** The "zoo" package is used for the extraction of features from the sliding window data. With *zoo::rollapply* the functions for the construction of the features is applied to rolling margins of the data. The parameter *width* is set to 3. **stats.** The "stats" package is used to perform PCA with the use of *stats::prcomp*. This function performs a principal components analysis on the data and returns the weights of the components. **glmnet.** "glmnet" is used to create an OLS model for the Monte Carlo simulation. The parameter *intercept* is set to TRUE, the parameters *alpha* and *lambda* to 0 and *standardize* to FALSE. **e1071.** The package "e1071" is used for training SVM. The parameters of *e1071::svm* are presented in the results section. **RWeka.** *RWeka::PART* is used for the PART algorithm. **ipred.** *ipred::bagging* is used to implement the bagging classification model. **randomForest.** The "randomForest" package is used for the implementation of the Random Forest algorithm. The parameters of the function *randomForest::randomForest* are listed in the results section. **caret.** "caret" is used to fit Elastic Net and k-NN to the data. *caret::confusionMatrix* is used for calculating cross-tabulations of the observed and predicted classes. *caret::createDataPartition()* is used to partition the data for experiment 3. **ggplot2, tidyr.** Packages used for creating visualizations of the data.