

University Degree in Computer Science
2017-2018

Bachelor Thesis

“Learning process’ analysis system: Learning Analytics”

Author: Alberto Blázquez Herranz

Supervisor: Miguel Ángel Patricio Guisado

Colmenarejo, Madrid, Spain

October 2018



This work is licensed under Creative Commons **Attribution – Non Commercial – Non Derivatives**

ACKNOWLEDGEMENTS

To all my loved ones: my family, my friends and my girlfriend, for their understanding with my commitment to this piece of work and, more especially, for relieving me from every feel of loneliness I have experienced during its development.

Thanks to you my hopes are more alive than ever.

LEARNING PROCESS' ANALYSIS SYSTEM: LEARNING ANALYTICS

Author: Alberto Blázquez Herranz

Abstract

As a consequence of nowadays' intensive data production, companies are setting its exploitation as a cornerstone for their growth, with new disciplines emerging with the intention of guiding this force of technological development, as it is the case of data intensive processes related with formative scenarios (academical or not).

Guidelines have been provided with the purpose of tackling current and upcoming challenges identified for the advancement of Learning Analytics. With special attention on its development and analytics facets, this project aims to take a step towards its feasible adoption.

With this purpose, an assessment of current literature's approach to this discipline's objectives has been conducted, concluding that, in order to capture a broader and effective picture of students' engagement to learning processes, a wide variety of information sources need to be considered, including qualitative ones.

Additionally, a set of scalable predictive models (involving regression and time series forecasting) related to students' interaction and outcomes have been developed with favourable results.

Finally, viability of the further development of these tasks and its inclusion in a real-world application are discussed.

INDEX

1	INTRODUCTION.....	16
1.1	Background and motivation	16
1.2	Objectives.....	17
1.3	Document structure	17
1.4	Socio-Economic Environment	18
1.5	Regulatory framework	18
2	STATE OF ART	19
2.1	Brief history of Learning Analytics and current challenges.....	19
2.2	Defining a domain: an approach to adoption of Learning Analytics	21
2.3	Theoretical reference for designing a tool.....	24
2.4	Previous work considered	25
3	DATA COLLECTION AND DATABASE SETTING	26
3.1	Original Database Model	27
3.1.1	Domain	28
3.1.2	Table and columns detail.....	28
3.1.3	Additional observations.....	34
3.2	Data pre-processing and other tasks performed	36
3.2.1	Ensuring data consistency	38
3.2.2	Adding format to date values	40
3.3	Result and first logical model.....	42
3.4	Database normalization.....	43
3.4.1	First Normal Form (1NF).....	44
3.4.2	Second Normal Form (2NF)	48
3.4.3	Third Normal Form (3NF)	52
3.4.4	Additional observations.....	53
3.5	Resulting Database Model	59
3.5.1	Logical model.....	60
3.5.2	Physical model	61
4	FEATURE SELECTION AND ENGINEERING	62
5	SETTING AN APPROACH	65
5.1	Assessing our case.....	65
5.2	Reinforcing statements.....	70
6	REGRESSION MODELS.....	75
6.1	Mean Interaction (prediction task).....	78

6.1.1	Treatment of categorical data.....	78
6.1.2	Feature selection.....	87
6.1.3	Cases of study.....	93
6.1.4	Selected algorithms	98
6.1.5	Results	98
6.2	Average score (prediction task).....	102
6.2.1	Treatment of categorical data.....	102
6.2.2	Feature selection.....	103
6.2.3	Cases of study.....	109
6.2.4	Selected algorithms	113
6.2.5	Results	113
7	TIME SERIES MODELS	118
7.1	Graphical assessment and definition of time series.....	119
7.1.1	Time series definition.....	144
7.2	Identification and definition of external regressors.....	151
7.2.1	Outliers' effects	151
7.3	Characteristics of the experiments conducted	153
7.3.1	Time constraints	153
7.3.2	Measure of error	154
7.3.3	Measure of residuals' correlation	154
7.4	Seasonality assessment.....	156
7.4.1	Time series decomposition and adjustment.....	160
7.4.2	Fourier series terms	165
7.5	Modelling	167
7.5.1	ARIMA	169
7.5.2	Neural Network	183
7.5.3	ETS.....	196
7.5.4	TBATS	200
7.5.5	Combination.....	203
7.5.6	Results	206
8	CONCLUSIONS.....	227
8.1	Further Work.....	227
9	APPENDIX: ORGANIZATION.....	228
9.1	Project's planification	228
9.2	Project's budget.....	228
10	REFERENCES.....	230

FIGURE INDEX

Figure 1. Learning Analytics emergence timeline	19
Figure 2. Pyramid graph: main factors influencing Learning Analytics adoption	21
Figure 3. Hype Cycle showing the placement of Predictive Analysis	22
Figure 4. Key areas using analytics within organizations [14]	23
Figure 5. .csv file format (viewed as text).....	26
Figure 6. .csv file format (Microsoft Excel visualization)	26
Figure 7. Database schema as shown in OU Analyse project’s website	27
Figure 8. Database element’s type legend.....	28
Figure 9. Merge R function.....	35
Figure 10. Inner join between studentInfo and studentRegistration’s tables	35
Figure 11. Fixing process of assessments table’s odd values.....	36
Figure 12. Fixed assessments table’s odd values	37
Figure 13. Calculation of “final_result” from student results	37
Figure 14. Redundant records (viewed as text).....	38
Figure 15. Data consistency definition.....	39
Figure 16. Correspondence between withdrawn students and their un-registration date	39
Figure 17. Missing un-registration date for a withdrawn student	40
Figure 18. Missing withdrawal tag for a student with an un-registration date assigned	40
Figure 19. Information on courses’ identification format.	41
Figure 20. Modification of date format	41
Figure 21. Database’s entity-relationship diagram.....	42
Figure 22. Database’s logical model	43
Figure 23. R function: "duplicated"	45
Figure 24. Loaded tables prior to row uniqueness checking	45
Figure 25. Result of the application of R’s “duplicated” function for row uniqueness checking.....	45
Figure 26. Change from the initial studentRegistration data arrangement (left) to its modification (right).	48
Figure 27. Partial dependency definition	48
Figure 28. StudentInfo’s table.....	49
Figure 29. Change from the initial studentInfo data arrangement (left) to its modification (right).	50
Figure 30. Unique user information from studentInfo	50
Figure 31. Second change from studentRegistration’s data arrangement (left) to its modification (right).....	51
Figure 32. Change from studentVle’s data arrangement (left) to its modification (right)	52
Figure 33. Change from studentAssessment’s data arrangement (left) to its modification (right)	52
Figure 34. Transitive dependency definition.....	52
Figure 35. Unique regions present in studentInfo’s table	53
Figure 36. Information on assessments’ types and weights.	54
Figure 37. Example of the occurrence of assessments with unassigned weight	54
Figure 38. SQL query showing that no assessments distinct from CMAs and TMAs were conducted for GGG modules.....	55
Figure 39. SQL query showing the number of assessments distinct from Exams conducted for GGG modules.....	55

Figure 40. SQL query showing the variety of results for GGG modules' students (limited to first 10 occurrences).	55
Figure 41. SQL query showing the number of students who finished a course performing any type of assessment.	56
Figure 42. SQL query showing the number of students who finished a course performing any type of assessment but exams.....	56
Figure 43. Capture from the result of and R anti-join between the results of querying for the students who finished a course performing any type of assessment and those who did any assessments but Exams	56
Figure 44. Individual query for each final result of the four students who only did an Exam (assessment type) during the course they were registered to.	57
Figure 45. Individual query for each assessment score of the four students who only did an Exam (assessment type) during the course they were registered to	57
Figure 46. Final database's logical model.....	60
Figure 47. Final database's physical model	61
Figure 48. Initial attributes: domain visualization.....	64
Figure 49. Diagram illustrating the traditional approach to student engagement	65
Figure 50. Boxplot for mean interaction and final result labelling	66
Figure 51. Resulting boxplot for mean interaction and final result labelling after outlier removal	67
Figure 52. Visualization of Cohen's D meaning (value of 1.017497)	69
Figure 53. Visualization of Cohen's D meaning (value of 1.60421)	69
Figure 54. Visualization of mean interaction against mean score values.....	71
Figure 55. Visualization of mean interaction against mean score values (outliers removed)	71
Figure 56. Mean interaction against mean score values: fitness of a linear model (left) and its correspondent residuals (right).....	72
Figure 57. Mean interaction against mean score values: fitness of a logarithmic model (left) and its correspondent residuals (right)	73
Figure 58. Diagram illustrating the approach decided for our project analytics tasks	74
Figure 59. Weight of Evidence's formula	76
Figure 60. Weight of Evidence's expression.....	79
Figure 61. Sample of the transformation "Region" variable from categorical to numerical.....	79
Figure 62. Formula for the silhouette of a cluster's point	80
Figure 63. Average silhouette's width for Age's band categories' clustering	80
Figure 64. Average silhouette's width for Highest education categories' clustering.....	81
Figure 65. Clustering of Highest education level.....	81
Figure 66. Explicit clustering of Highest education level	82
Figure 67. Average silhouette's width for Region categories' clustering	82
Figure 68. Clustering of Region.....	83
Figure 69. Explicit clustering of Region	84
Figure 70. Average silhouette's width for IMD's band categories' clustering	85
Figure 71. Clustering of IMD's band	86
Figure 72. Explicit clustering of IMD band	87
Figure 73. Features' importance with respect to mean interaction values (information gain)	88
Figure 74. Features' importance with respect to mean interaction values (information gain ratio)	89
Figure 75. Correlation coefficient matrix of variable set's (numerical).....	90
Figure 76. Cramer's V matrix of variable set's (categorical).....	91
Figure 77. Intra-Class Correlation matrix of variable set's (categorical vs. numerical)	92

Figure 78. Explicit clustering of Highest education level	102
Figure 79. Explicit clustering of IMD band	103
Figure 80. Explicit clustering of Region	103
Figure 81. Features' importance with respect to mean interaction values (information gain) ..	104
Figure 82. Features' importance with respect to mean interaction values (information gain ratio)	105
Figure 83. Correlation coefficient matrix of variable set's (numerical).....	106
Figure 84. Cramer's V matrix of variable set's (categorical).....	107
Figure 85. Intra-Class Correlation matrix of variable set's (categorical vs. numerical)	108
Figure 86. Fourier series formula.....	119
Figure 87. First terms of a Fourier series	119
Figure 88. Time series plot of course "AAA-2013J"	121
Figure 89. Time series plot of course "AAA-2014J"	120
Figure 90. Time series plot of course "BBB-2013B"	122
Figure 91. Time series plot of course "BBB-2013J"	121
Figure 92. Time series plot of course "BBB-2014B"	123
Figure 93. Time series plot of course "BBB-2014J"	122
Figure 94. Time series plot of course "CCC-2014B"	124
Figure 95. Time series plot of course "CCC-2014J"	123
Figure 96. Time series plot of course "DDD-2013B"	125
Figure 97. Time series plot of course "DDD-2013J"	124
Figure 98. Time series plot of course "DDD-2014B"	126
Figure 99. Time series plot of course "DDD-2014J"	125
Figure 100. Time series plot of course "EEE-2013J"	127
Figure 101. Time series plot of course "EEE-2014B"	126
Figure 102. Time series plot of course "EEE-2014J"	128
Figure 103. Time series plot of course "FFF-2013B"	127
Figure 104. Time series plot of course "FFF-2013J"	129
Figure 105. Time series plot of course "FFF-2014B"	128
Figure 106. Time series plot of course "FFF-2014J"	130
Figure 107. Time series plot of course "GGG-2013J"	129
Figure 108. Time series plot of course "GGG-2014B"	131
Figure 109. Time series plot of course "GGG-2014J"	130
Figure 110. Time series plot of course "AAA-2013J"	134
Figure 111. Time series plot of course "AAA-2014J"	133
Figure 112. Time series plot of course "BBB-2013B"	135
Figure 113. Time series plot of course "BBB-2013J"	134
Figure 114. Time series plot of course "BBB-2014B"	136
Figure 115. Time series plot of course "BBB-2014J"	135
Figure 116. Time series plot of course "CCC-2014B"	137
Figure 117. Time series plot of course "CCC-2014J"	136
Figure 118. Time series plot of course "DDD-2013B"	138
Figure 119. Time series plot of course "DDD-2013J"	137
Figure 120. Time series plot of course "DDD-2014B"	139
Figure 121. Time series plot of course "DDD-2014J"	138
Figure 122. Time series plot of course "EEE-2013J"	140
Figure 123. Time series plot of course "EEE-2014B"	139
Figure 124. Time series plot of course "EEE-2014J"	141

Figure 125. Time series plot of course “FFF-2013B”	140
Figure 126. Time series plot of course “FFF-2013J”	142
Figure 127. Time series plot of course “FFF-2014B”	141
Figure 128. Time series plot of course “FFF-2014J”	143
Figure 129. Time series plot of course “GGG-2013J”	142
Figure 130. Time series plot of course “GGG-2014B”	144
Figure 131. Time series plot of course “GGG-2014J”	143
Figure 132. Joint time series plot of courses “AAA-2013J” and “AAA-2014J”	145
Figure 133. Joint time series plot of courses “BBB-2013B” and “BBB-2014B”	145
Figure 134. Joint time series plot of courses “BBB-2013J” and “BBB-2014J”	146
Figure 135. Time series plot of course “CCC-2014B”	147
Figure 136. Time series plot of course “CCC-2014J”	146
Figure 137. Joint time series plot of courses “DDD-2013B” and “DDD-2014B”	147
Figure 138. Joint time series plot of courses “DDD-2013J” and “DDD-2014J”	147
Figure 139. Time series plot of course “EEE-2014B”	148
Figure 140. Joint time series plot of courses “EEE-2013J” and “EEE-2014J”	148
Figure 141. Joint time series plot of courses “FFF-2013B” and “FFF-2014B”	149
Figure 142. Joint time series plot of courses “FFF-2013J” and “FFF-2014J”	149
Figure 143. Time series plot of course “GGG-2014B”	150
Figure 144. Joint time series plot of courses “GGG-2013J” and “GGG-2014J”	150
Figure 145. Graphical representation of indicator variables’ confection	152
Figure 146. Figure showing discrepancies between Ljung-Box and Breusch-Godfrey test for the same residual evaluation (output from a simple ARIMA model on “AAA-2013J” time series)	155
Figure 147. Forecast modelling exemplification: “AAA-2013J” and “AAA-2014J” joint time series.....	157
Figure 148. Periodogram for “AAA-2013J” course time series and its most relevant spectrum spikes (with its correspondent time in days)	159
Figure 149. Forecast modelling exemplification: “AAA-2013J” and “AAA-2014J” joint time series decomposition (weekly and bimonthly seasonality)	161
Figure 150. Forecast modelling exemplification: “AAA-2013J” and “AAA-2014J” joint time series logarithmic decomposition (weekly and bimonthly seasonality).....	162
Figure 151. Forecast modelling exemplification: “AAA-2013J” and “AAA-2014J” adjusted (for weekly and bimonthly seasonality) joint time series (logarithm)	164
Figure 152. First terms of a Fourier series	165
Figure 153. Overlapping of Fourier time series (accounting for weekly and bimonthly seasonality) with “AAA-2013J” time series (logarithm)	166
Figure 154. Forecast modelling exemplification: “AAA-2013J” and “AAA-2014J” joint time series.....	168
Figure 155. Best forecasting model achieved by using an ARIMA model: “AAA-2013J” and “AAA-2014J” joint time series	171
Figure 156. Residuals resulting from the best forecasting model achieved by using an ARIMA model: “AAA-2013J” and “AAA-2014J” joint time series	172
Figure 157. Best forecasting model achieved by using an ARIMA model with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series	174
Figure 158. Resulting residuals from the best forecasting model achieved by using an ARIMA model with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series ...	175
Figure 159. Best forecasting model achieved by using a de-seasonalized ARIMA model: “AAA-2013J” and “AAA-2014J” joint time series	177

Figure 160. Resulting residuals from the best forecasting model achieved by using a de-seasonalized ARIMA model: “AAA-2013J” and “AAA-2014J” joint time series.....	178
Figure 161. Best forecasting model achieved by using an ARIMA model on de-seasonalized data with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series.....	180
Figure 162. Resulting residuals from the best forecasting model achieved by using an ARIMA model on de-seasonalized data with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series.....	181
Figure 163. Best forecasting model achieved by using a Neural Network model: “AAA-2013J” and “AAA-2014J” joint time series	184
Figure 164. Resulting residuals from the best forecasting model achieved by using a Neural Network model: “AAA-2013J” and “AAA-2014J” joint time series	185
Figure 165. Best forecasting model achieved by using a Neural Network model with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series.....	187
Figure 166. Resulting residuals from the best forecasting model achieved by using a Neural Network model with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series.....	188
Figure 167. Best forecasting model achieved by using a Neural Network model on de-seasonalized data: “AAA-2013J” and “AAA-2014J” joint time series.....	190
Figure 168. Best forecasting model achieved by using a Neural Network model on de-seasonalized data: “AAA-2013J” and “AAA-2014J” joint time series.....	191
Figure 169. Best forecasting model achieved by using a Neural Network model on de-seasonalized data with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series.....	193
Figure 170. Resulting residuals from the best forecasting model achieved by using a Neural Network model on de-seasonalized data with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series	194
Figure 171. Overfitted forecasting model achieved by using an ETS model: “AAA-2013J” and “AAA-2014J” joint time series	197
Figure 172. Resulting residuals from the overfitted forecasting model achieved by using an ETS model: “AAA-2013J” and “AAA-2014J” joint time series	198
Figure 173. Overfitted forecasting model achieved by using a TBATS model: “AAA-2013J” and “AAA-2014J” joint time series	201
Figure 174. Resulting residuals from the overfitted forecasting model achieved by using a TBATS model: “AAA-2013J” and “AAA-2014J” joint time series.....	202
Figure 175. Forecasting model achieved by merging (mean) both best ARIMA and Neural Network models: “AAA-2013J” and “AAA-2014J” joint time series.....	204
Figure 176. Residuals from the forecasting model achieved by merging (mean) both best ARIMA and Neural Network models: “AAA-2013J” and “AAA-2014J” joint time series.....	205

TABLE INDEX

Table 1. Table's description template.....	28
Table 2. Course's description table.....	29
Table 3. Assessment's description table	29
Table 4. Vle's description table	30
Table 5. Student Info's description table	31
Table 6. Student Registration's description table	32
Table 7. Student Assessments' description table.....	33
Table 8. Student Vle's description table.....	34
Table 9. Tables affected by date formatting process.....	41
Table 10. Entities' type definition	42
Table 11. Modified StudentRegistration table	47
Table 12. StudentUnregistration table.....	47
Table 13. Features' description template.....	62
Table 14. Features' description	63
Table 15. Welch's t-test and Cohen's D results (influence of interaction in final results).....	68
Table 16. Intra-Class correlation results (final results described by mean interaction)	70
Table 17. Appropriateness of linear and logarithmic models to describe final results based on mean interaction	73
Table 18. Regression tasks' cases of study	77
Table 19. Predictors employed for mean interaction's regression tasks.....	78
Table 20. Categorical predictors and their cardinality (mean interaction's regression tasks)	78
Table 21. Predictors resulting from applying an Information Gain filter (mean interaction's regression tasks).....	89
Table 22. Predictors resulting from applying an Information Gain Ratio filter (mean interaction's regression tasks).....	90
Table 23. Predictors resulting from applying a Collinearity filter based on Information Gain (mean interaction's regression tasks).....	93
Table 24. Predictors resulting from applying a Collinearity filter based on Information Gain Ratio (mean interaction's regression tasks)	93
Table 25. Predictors present in “RAW” case of study (mean interaction's regression tasks).....	93
Table 26. Predictors present in “CLUST” case of study (mean interaction's regression tasks) ..	94
Table 27. Predictors present in “NUM” case of study (mean interaction's regression tasks).....	94
Table 28. Predictors present in “IG” case of study (mean interaction's regression tasks).....	94
Table 29. Predictors present in “IGR” case of study (mean interaction's regression tasks)	94
Table 30. Predictors present in “COLL1” case of study (mean interaction's regression tasks)..	95
Table 31. Predictors present in “COLL2” case of study (mean interaction's regression tasks)..	95
Table 32. Predictors present in “COLL1+CLUST” case of study (mean interaction's regression tasks)	95
Table 33. Predictors present in “COLL1+NUM” case of study (mean interaction's regression tasks)	96
Table 34. Predictors present in “COLL2+CLUST” case of study (mean interaction's regression tasks)	96
Table 35. Predictors present in “COLL2+NUM” case of study (mean interaction's regression tasks)	96
Table 36. Predictors present in “IG+COLL” case of study (mean interaction's regression tasks)	97

Table 37. Predictors present in “IGR+COLL” case of study (mean interaction’s regression tasks)	97
Table 38. Results from “RAW” case of study (mean interaction’s regression tasks)	98
Table 39. Results from “CLUST” case of study (mean interaction’s regression tasks).....	98
Table 40. Results from “NUM” case of study (mean interaction’s regression tasks)	98
Table 41. Results from “IG” case of study (mean interaction’s regression tasks)	99
Table 42. Results from “IGR” case of study (mean interaction’s regression tasks)	99
Table 43. Results from “COLL1” case of study (mean interaction’s regression tasks).....	99
Table 44. Results from “COLL2” case of study (mean interaction’s regression tasks).....	99
Table 45. Results from “COLL1+CLUST” case of study (mean interaction’s regression tasks)99	
Table 46. Results from “COLL1+NUM” case of study (mean interaction’s regression tasks) ..	99
Table 47. Results from “COLL2+CLUST” case of study (mean interaction’s regression tasks)99	
Table 48. Results from “COLL2+NUM” case of study (mean interaction’s regression tasks) 100	
Table 49. Results from “IG+COLL” case of study (mean interaction’s regression tasks)	100
Table 50. Results from “IGR+COLL” case of study (mean interaction’s regression tasks).....	100
Table 51. Summary of best regression models and algorithms (mean interaction)	101
Table 52. Predictors employed for average score's regression tasks.....	102
Table 53. Predictors resulting from applying an Information Gain filter (average score's regression tasks)	105
Table 54. Predictors resulting from applying an Information Gain Ratio filter (average score's regression tasks)	105
Table 55. Predictors resulting from applying a Collinearity filter based on Information Gain (average score's regression tasks).....	108
Table 56. Predictors resulting from applying a Collinearity filter based on Information Gain Ratio (average score's regression tasks)	109
Table 57. Predictors present in “RAW” case of study (average score’s regression tasks).....	109
Table 58. Predictors present in “CLUST” case of study (average score's regression tasks)	109
Table 59. Predictors present in “NUM” case of study (average score's regression tasks)	110
Table 60. Predictors present in “IG” case of study (average score's regression tasks)	110
Table 61. Predictors present in “IGR” case of study (average score's regression tasks).....	110
Table 62. Predictors present in “COLL1” case of study (average score's regression tasks)	110
Table 63. Predictors present in “COLL2” case of study (average score's regression tasks)	111
Table 64. Predictors present in “COLL1+CLUST” case of study (average score's regression tasks)	111
Table 65. Predictors present in “COLL1+NUM” case of study (average score's regression tasks)	111
Table 66. Predictors present in “COLL2+CLUST” case of study (average score's regression tasks)	112
Table 67. Predictors present in “COLL2+NUM” case of study (average score's regression tasks)	112
Table 68. Predictors present in “IG+COLL” case of study (average score's regression tasks). 112	
Table 69. Predictors present in “IGR+COLL” case of study (average score's regression tasks)	112
Table 70. Results from “RAW” case of study (average score’s regression tasks).....	113
Table 71. Results from “CLUST” case of study (average score’s regression tasks).....	113
Table 72. Results from “NUM” case of study (average score’s regression tasks).....	113
Table 73. Results from “IG” case of study (average score’s regression tasks).....	114
Table 74. Results from “IGR” case of study (average score’s regression tasks)	114

Table 75. Results from “COLL1” case of study (average score’s regression tasks).....	114
table 76. Results from “COLL2” case of study (average score’s regression tasks).....	114
Table 77. Results from “COL1+CLUST” case of study (average score’s regression tasks)	114
Table 78. Results from “COLL1+NUM” case of study (average score’s regression tasks)	114
Table 79. Results from “COLL2+CLUST” case of study (average score’s regression tasks)..	114
Table 80. Results from “COLL2+NUM” case of study (average score’s regression tasks)	115
Table 81. Results from “IG+COLL” case of study (average score’s regression tasks)	115
Table 82. Results from “IGR+COLL” case of study (average score’s regression tasks).....	115
Table 83. Summary of best regression models and algorithms (average score)	116
Table 84. Results from an approach based on mean interaction alone	117
Table 85. Forecasting tasks' cases of study	119
Table 86. Time series' re-arrangement (forecasting tasks).....	144
Table 87. Seasonal periods' selection process (forecasting tasks).....	160
Table 88. Forecasting tasks' cases of study	167
Table 89. Results for regular ARIMA forecasting process	170
Table 90. Results for ARIMA (with regressors) forecasting process	173
Table 91. Results for ARIMA (de-seasonalized data) forecasting process.....	176
Table 92. Results for ARIMA (de-seasonalized data with regressors) forecasting process.....	179
Table 93. Best ARIMA forecasting results	182
Table 94. Results for regular Neural Network forecasting process.....	183
Table 95. Results for Neural Network (with regressors) forecasting process.....	186
Table 96. Results for Neural Network (de-seasonalized data) forecasting process	189
Table 97. Results for Neural Network (de-seasonalized data with regressors) forecasting process	192
Table 98. Best Neural Network forecasting results.....	195
Table 99. Results for ETS forecasting process.....	196
Table 100. Results for TBATS forecasting process	200
Table 101. Combination of best models' forecasting results.....	203
Table 102. Time series' re-arrangement (forecasting tasks).....	206
Table 103. ARIMA forecasting results for course AAA-J (2013-2014).....	206
Table 104. Neural Network forecasting results for course AAA-J (2013-2014)	207
Table 105. Combination forecasting results for course AAA-J (2013-2014)	207
Table 106. ARIMA forecasting results for course BBB-B (2013-2014)	208
Table 107. Neural Network forecasting results for course BBB-B (2013-2014).....	208
Table 108. Combination forecasting results for course BBB-B (2013-2014).....	209
Table 109. ARIMA forecasting results for course BBB-J (2013-2014)	209
Table 110. Neural Network forecasting results for course BBB-J (2013-2014).....	210
Table 111. Combination forecasting results for course BBB-J (2013-2014).....	210
Table 112. ARIMA forecasting results for course CCC-B (2014).....	211
Table 113. Neural Network forecasting results for course CCC-B (2014)	211
Table 114. Combination forecasting results for course CCC-B (2014)	212
Table 115. ARIMA forecasting results for course CCC-J (2014).....	212
Table 116. Neural Network forecasting results for course CCC-J (2014)	213
Table 117. Combination forecasting results for course CCC-J (2014)	213
Table 118. ARIMA forecasting results for course DDD-B (2013-2014).....	214
Table 119. Neural Network forecasting results for course DDD-B (2013-2014)	214
Table 120. Combination forecasting results for course DDD-B (2013-2014)	215
Table 121. ARIMA forecasting results for course DDD-J (2013-2014).....	215

Table 122. Neural Network forecasting results for course DDD-J (2013-2014)	216
Table 123. Combination forecasting results for course DDD-J (2013-2014)	216
Table 124. ARIMA forecasting results for course EEE-B (2014)	217
Table 125. Neural Network forecasting results for course EEE-B (2014).....	217
Table 126. Combination forecasting results for course EEE-B (2014).....	218
Table 127. ARIMA forecasting results for course EEE-J (2013-2014)	218
Table 128. Neural Network forecasting results for course EEE-J (2013-2014).....	219
Table 129. Combination forecasting results for course EEE-J (2013-2014).....	219
Table 130. ARIMA forecasting results for course FFF-B (2013-2014).....	220
Table 131. Neural Network forecasting results for course FFF-B (2013-2014)	220
Table 132. Combination forecasting results for course FFF-B (2013-2014)	221
Table 133. ARIMA forecasting results for course FFF-J (2013-2014).....	221
Table 134. Neural Network forecasting results for course FFF-J (2013-2014)	222
Table 135. Neural Network forecasting results for course FFF-J (2013-2014)	222
Table 136. ARIMA forecasting results for course GGG-B (2014).....	223
Table 137. Neural Network forecasting results for course GGG-B (2014)	223
Table 138. Combination forecasting results for course GGG-B (2014)	224
Table 139. ARIMA forecasting results for course GGG-J (2014).....	224
Table 140. Neural Network forecasting results for course GGG-J (2014).....	225
Table 141. Combination forecasting results for course GGG-J (2014).....	225
Table 142. Summary of forecasting models and algorithms' results.....	226
Table 143. Schedule of the project's tasks.....	228
Table 144. Hardware required for the development of the project.....	228
Table 145. Workforce required for the development of the project.....	228
Table 146. Final project's budget	229

1 INTRODUCTION

1.1 Background and motivation

Although nowadays it is widely assumed that we are already immerse in the so-called Information Age, we are still far away from a complete embracement of what this term is considered to imply.

With increasing insistence in its characterization for the abundance of data available from even more diverse sources, the understanding of the potential discovery of knowledge and opportunities its exploitation may entail has brought with it a paradigm shift with respect to the way we perceive information and its usage.

As a consequence, companies are setting it as a cornerstone for their growth ([1]), with new disciplines emerging with the intention of guiding this force of technological development. However, at a worldwide scale, these organizations are far from an ideal implementation, which would allow them to properly capitalize on its benefits ([2]-[3]).

This goes to show that we are at a stage in which the adhesion to data-driven processes is still nascent, which is also the case of its application to learning environments, namely Learning Analytics. Threshing the concept, it comprises “*the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs*” ([4]).

From this definition, it is important to remark the reference to learning contexts, in which variety resides an important source of value for this field, facet vaguely explored in related literature.

The approach to which most research refers its efforts remains purely academical, ignoring the potential benefits of the intersection of Learning Analytics with other ambits, such as its consideration as part of a Business Intelligence agenda, field from which, in fact, it is acknowledged to drawn on ([5]).

Additionally, it has been pointed out by different institutions involved in the monitoring of this field’s development that, besides the advances made in its related procedural tasks (data models, analytic algorithms, etc), there is a lack of consensus in the approach to its main objectives, resulting in an overall fragmented work which may impair its standardization and maturity ([5]-[6]).

Guidelines have been provided with the purpose of tackling current and upcoming challenges identified for the advancement of Learning Analytics. With special attention on its development and analytics facets, this project aims to take a step towards its feasible adoption.

1.2 Objectives

The objectives stated for this project are the following:

- Assessment of current literature's approach to Learning Analytics in order to tackle its assumptions from a statistical point of view. An analysis of this process' results is intended allow for an objective evaluation of its validity and suitability to accomplish Learning Analytics' objectives.
- Development of scalable predictive and analytics processes as a starting point for its further inclusion as part of a real-world solution.

1.3 Document structure

Apart from the present introduction, this document comprises the following sections:

- State of art: discussion referred to the domain in which Learning Analytics as a discipline is developing and growing.
Guidelines for the improvement of its situation and implementation of solutions are presented.
- Data collection and database setting: presentation of the data-source used for the development of the designed tasks.
A database for its containment is designed and deployed for its operation during the analytics processes to be conducted.
- Feature selection and engineering: the attributes involved in the analytics tasks are reviewed and assessed. Additionally, other features extracted from the data are created.
- Setting an approach: the fundamentals in which current literature's tackle on Learning Analytics is based are discussed. From its assessment, a new basis from which to work is developed.
- Regression models: development and results from the tasks involving regression procedures are presented and discussed.
- Time series models: modelling processes referred to the arrangement and treatment of time series, as well as results from their correspondent forecast, are discussed.
- Conclusions: final assessment of the extent to which the project's objectives have been accomplished.
Further work from which this project is intended to scale is discussed.
- Appendix - Organization: the main tasks involved in the development of this project, in conjunction with its distribution over time, are detailed in this section.

1.4 Socio-Economic Environment

This project's objectives are aimed to exploit the potential existing in the data generated during formative processes in order to improve both its development and outcomes. With this purpose, the designed processes are intended to redound in the embracement of pure analytical tasks (e.g. diagnosis, prediction, etc).

These processes are designed to be included in an operative Business Intelligence application and operate on its available data, significantly reducing any cost of implementation.

Moreover, the inclusion of Learning Analytics in any institution (academical or business-oriented) allows for the improvement of the qualifications of its formative processes' target public, which may lead to benefits referred to efficacy of procedures and institutional leadership.

1.5 Regulatory framework

The tasks to be designed in this project are intended to operate in an already deployed and operative Business Intelligence system, which redounds in the need to address treatment of personal data in order to guarantee the correspondent safety and privacy requirements.

For this purpose, regulations from the Spanish Data Protection Law [7] are considered for the development of this project's objectives in a real-world scenario.

The final system is not intended to transmit any personal information to third-parties since its scope of application implies particular institution's personal.

2 STATE OF ART

2.1 Brief history of Learning Analytics and current challenges

It is important to recall the previous mention to the connection between Business Intelligence and this field when referring to its origins. In fact, its emergence can be linked to the need for understanding behaviour at both internal organizational and consumer scopes ([8]).

The appearance of modern Learning Management Systems in the early 2000s is considered to have played an important role in the discernment of utilities associated to the potential information that could be extracted from them, thus being the germ of the main concerns associated with Learning Analytics today ([8]).

Along the way to 2011, when Learning Analytics as a discipline emerged ([5]), many other fields contributed to its advancement, from which the most influencing, in terms of dependence on its features, are eLearning (constitutes the theoretical basis behind performance enhancement in learning environments through the use of technological resources) and the already mentioned LMS (platforms providing educational resources and activity-tracking services) ([9]).

In 2013, when the first usable data models were finally made up, few early adopters were involved in the development and testing of algorithms on real student data, which roughly comprised a time span of two years since then. It wasn't until 2015's academic year when first trials using those algorithms were conducted, which related findings started to be reported in 2016.

Currently, as Learning Analytics broadens its scope towards consideration of a more detailed model from which to extract conclusions, more disciplines are being taken into account, with Social Network Analysis drawing more attention as the will to include other facets apart from those purely performative in the current data model increases ([9]-[10]).

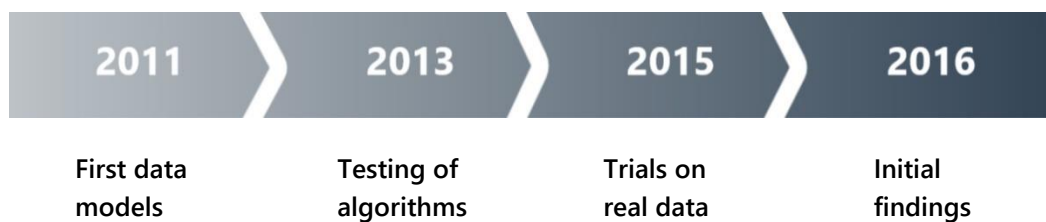


Figure 1. Learning Analytics emergence timeline

Such a recent development history showcases the short maturation process that Learning Analytics has undergone as a formal field, which is often cited by specialized literature as the current main barrier to overcome for the development of the field ([5]-[6]), as it implies challenging, and often connected, situations:

- **Fragmented work** among the different institutions concerned with its development. It may act as a contributor to the field’s stagnation if no efforts are directed towards the alignment of projects and their scopes.

In fact, the most recent and widespread initiative concerned with this factor, Learning Analytics Community Exchange (LACE), ended in summer 2016 ([5]), which reveals the need for a new point of convergence for Learning Analytics research.

- **Impaired validation**: as it has been mentioned, Learning Analytics reduced background doesn’t allow for the presentation of strong evidence of its benefits (“the path from initial pilot studies to validated analytics takes years” ([5]), and thus, most findings supporting its premises are either based on short-term studies or belonging to approaches previous to its formalization ([5]).

- **Low investment** on its deployment at an institutional level, where its consideration has been reported to be that of an “interest rather than a major priority” ([6]). In fact, this can be seen as the concurrent factor of the previous points, which contribute to a poor perception from potential investors in which respects to the field’s maturity and its capabilities.

With this set of circumstances defining the current state of Learning Analytics as a field, the need for understanding the nascent stage at which it is seems clear, focusing efforts on the necessities it involves. Even if potential promising uses and benefits can be formulated, it is first necessary to build a stable and standardized environment for them to be feasible and, more importantly, adopted by target institutions.

This is widely recognized to be a long-term process comprising multiple ambits from which to build up, with literature ([5]-[6]-[11]-[12]) sharing a common understanding of the needs for standardization (aimed towards the solving of issues related to fragmented work by capitalizing on the advancements made by previous projects concerned with this situation) and attraction of investors (centred around the alignment of both validation processes and objectives with available resources in order to improve given priority among stakeholders and investors’ perception of worthiness).

With summarizing purposes, the following pyramid diagram has been elaborated to illustrate the main steps to take towards the generalized adoption of Learning Analytics:



Figure 2. Pyramid graph: main factors influencing Learning Analytics adoption

2.2 Defining a domain: an approach to the adoption of Learning Analytics

After the previous review of the general picture of the state of Learning Analytics, it is important to inspect its adoption opportunities with respect to current trends. This will set a recognizable domain from which to better detect implementation possibilities, establish links with current market and demand, and also justify decisions with respect to the development of this and other projects.

For this purpose, Gartner Inc.'s research towards the identification of technological trends and its evolution has been taken into account, with special attention to the model for the development of a Hype Cycle ([13]).

In summary, Gartner's Hype Cycle sets a relationship between the attention a field is drawing into its potential benefits and the real stage of its implantation (addressing the evolution of expectations over time). It sets a reference for interpreting the state at which a rising technology is, how is it perceived and the environment in which its development is taking place.

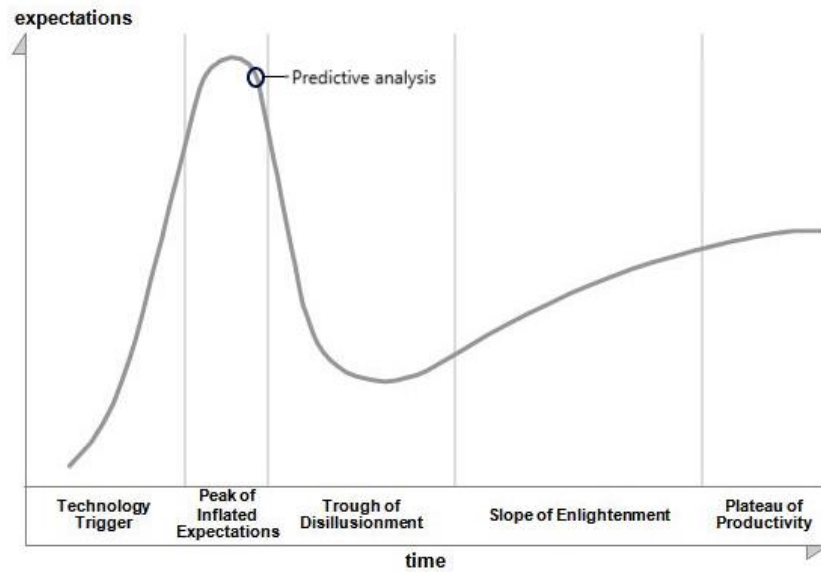


Figure 3. Hype Cycle showing the placement of Predictive Analysis

Coherently with Gartner’s analysis ([14]), the placement shown in the above figure for Predictive analysis responds to its recognition as one of today’s most disruptive forces with respect to the conception of decision-making procedures.

The fact that the discipline is considered to be leaving its “expectations’ peak” matches a widespread statement present in research literature, which points out that major decisions with respect to its total embracement are still to be made and that there is room for the maturation of the processes involved ([1]-[3]-[15]).

This also links with the prognosed entrance in the “through of disillusionment” stage, which will most likely not happen in the aggressive manner depicted in *figure 3* (according to Hype Cycle’s documentation ([13]), it implies a generalized failure of development processes, surmountable by the improvement of already deployed implementations). This assumption of a “calmed path” towards the development of the field has its fundamentals on the huge quantity of early implementations providing substantial benefits to investors ([1]-[3]), seemingly regardless of its recognition as not totally mature.

Although a direct parallelism between the stage of adoption and general perception of Predictive analysis and Learning Analytics can’t be established, it is important to note that the former is a general concept involving any discipline related to data exploitation with prognostic purposes. Thus, implications referred to current likelihood of investment and adoption of predictive technologies should not be dismissed in which regards to Learning Analytics.

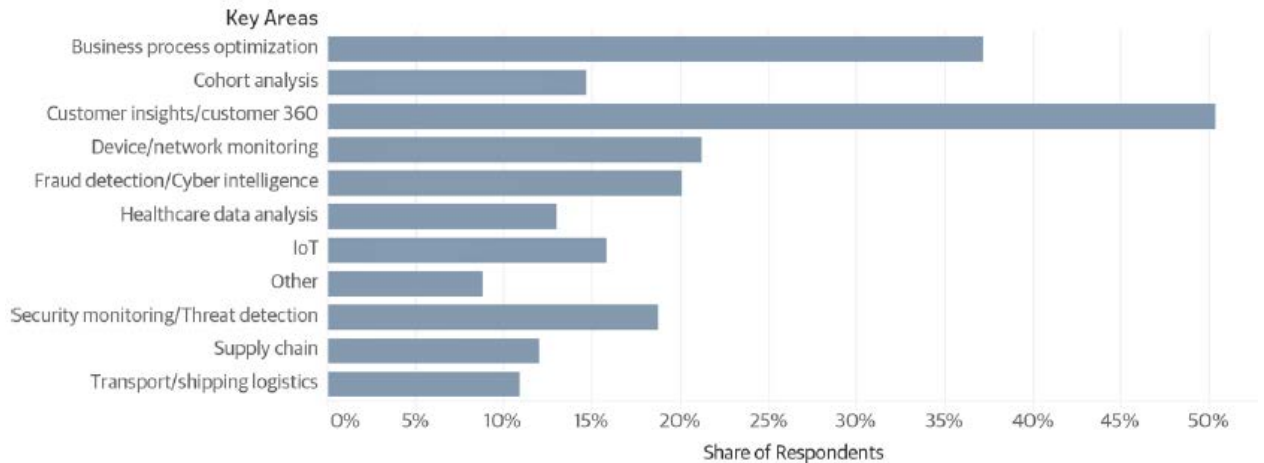


Figure 4. Key areas using analytics within organizations [14]

In consonance with the previous statement, it is important to assess the trends related to how are current analytics implementations approached and their ambit of application. Figure 4 serves as a summary for identifying the main ambits to which data-driven processes are being aimed, with business processes optimization and customer-oriented analytics receiving most efforts ([1]-[16]-[17]). This information can help setting a bridge between applications present in the current market, its demands and Learning Analytics. In fact, recalling the historical background previously given, Learning Analytics' core foundations respond to the organizational purpose of understanding people's behaviour within an institution, thus establishing the following correspondence:

- Customer → Learner
- Business process to optimize → Learning

Any environment at which formation takes place (academical, enterprise, on-line...) is then subject to the possibility of including its assessment in the correspondent institution's performative agenda, frequently approached from a Business Intelligence point of view ([18]).

Finally, it is important to remark that, although a common approach has been established for Learning Analytics at an institutional level, distinct forms of instruction take place at each type of organization. This means that assumptions with respect to the suitability of designed tools to multiple ambits should not be made since significant differences responding to the specific characteristics and necessities of each domain of application may appear ([5]).

2.3 Theoretical reference for designing a tool

After proposing the inclusion of Learning Analytics in an enterprise's Business Intelligence (BI from here on) agenda, problems of current BI tools should be addressed, with the main concern being the low usage among already deployed BI systems ([19]), which seems paradoxical according to the importance they're considered to have among its target public.

It is then coherent to address this event as a target public's engagement/involvement problem, with the cause being discussed to be related with the lack of consideration given to social factors in organizations when designing these tools.

The ongoing paradigm shift in the use of technology (embodied by the term Web 2.0), which fosters the role of users and information sharing as the primary sources of material for a site ([19]), can be thought of as indicative in assessing and tackling these problems. In fact, no current BI tool explicitly facilitates/takes into consideration user interaction and contribution, lacking correspondence with the mentioned Web 2.0 model ([20]).

Coherently with the placement of value in both the platform itself and its community of users, contributions adopt a main role in developing responsiveness to user needs. Attending to O'Reilly's classification of contribution types ([20]), those fostering social networks within the tool's environment should be thought of as the main challenge to tackle in order to fill in the gap between current underused BI tools and a model aimed to meet the requirements of the current technological environment and its usage.

Contributions to the platform itself and its content should also be taken into consideration, primarily with the purpose of guaranteeing the platform's adaptation to user needs, which is more likely to lead to a lasting usage over time. The following practices regarding this subject are to be remarked ([19]):

- Allowance for personalized interaction with the platform, to the extent possible (e.g. plugins, customized reports for a BI tool, etc).
- Increased and perceptible prioritization of communication-related functionalities, including content sharing.
- Analytics functionalities to aid navigation through the platform's content (e.g. tag search, filtering, etc).

The described approach acknowledges the communicative nature of deciding and its social implications ([19]). Thus, it is based and aimed towards the addressment of both social and logical-empirical factors (e.g. KPIs) in the development of tools oriented to aid decision-making processes.

This last point should be intersected with Learning Analytics' concerns about the lack of exploitation of the potential utilities of its processes (e.g. success prediction, prescription of intervention strategies, etc), in favour of traditional practices limited to reporting ([6]).

Consequently, the guidelines defined in this section for the development of a tool should also be aimed towards an efficient merge of human assessment capabilities with more refined analytics functionalities.

2.4 Previous work considered

Reference has been drawn from several ongoing (and ended) projects, especially from those addressing Learning Analytics from an institutional approach. Assessment of their specific and common characteristics allowed for a proper definition and comprehension of the current basis on which these projects are sustained so that the operations and processes designed provide the most possible value.

Among the applications reviewed (which references were found in [5]) with this purpose we find:

- **Academical environment:**
 - OU Analyse [21]: provides a software for the detection of students at risk, as well as a dashboard describing the reasoning of behind its predictions.
A sample dataset provided by the institution behind this project (Open University) has been used in the development of this project.
 - Tribal's Student Insights [22]: aimed to the detection of students at risk through the application of performance-based models (e.g. estimation of the likelihood of a student to pass an assessment).
 - Loop [23]: reporting tool for providing information about courses and its components, as well as interaction data from students.
- **Workplace:**
 - Skillaware [24]: software oriented to the measurement of employee's effectiveness and the detection of knowledge-areas which may need training (operates by capturing and analysing users' behaviours).
- **Informal learning:**
 - Khan Academy analytics [25]: tool for reporting information with respect to estimated effort, engagement and mastery of skills.

As it can be extracted from the summaries of these tool's functionalities, and in accordance to the points exposed in the previous section, most of the current effort is centred around the usage of performative measures, as well as on reporting outputs (with certain exceptions to this last point).

Consequently, and recalling the guidelines exposed for the development of a usable and perdurable tool previously shown, the operations designed along this project's conduction should consider an aim to generate innovative procedures which add valuable information from pure analytics and machine learning tasks.

3 DATA COLLECTION AND DATABASE SETTING

Data was obtained from OU Analyse project's website, which provides a collection containing "data about courses, students and their interactions with Virtual Learning Environment (VLE) for seven selected courses" ([26]).

The records are distributed in seven .csv files, each of which represents a table with its correspondent attributes and unique identifiers.

The .csv file format has each cell surrounded by quotes and separated from each other by comas.

```
"AAA", "2013J", "1752", "TMA", "19", "10"
```

Figure 5. .csv file format (viewed as text)

AAA	2013J	1752	TMA	19	10
-----	-------	------	-----	----	----

Figure 6. .csv file format (Microsoft Excel visualization)

The website also proposes a database architecture based on these tables. It will be discussed in the following sections.

It is relevant to advance that modifications to the original model and data pre-processing tasks (apart from those belonging to the analytics phase) were needed in order to keep feasibility and data coherence.

3.1 Original Database Model

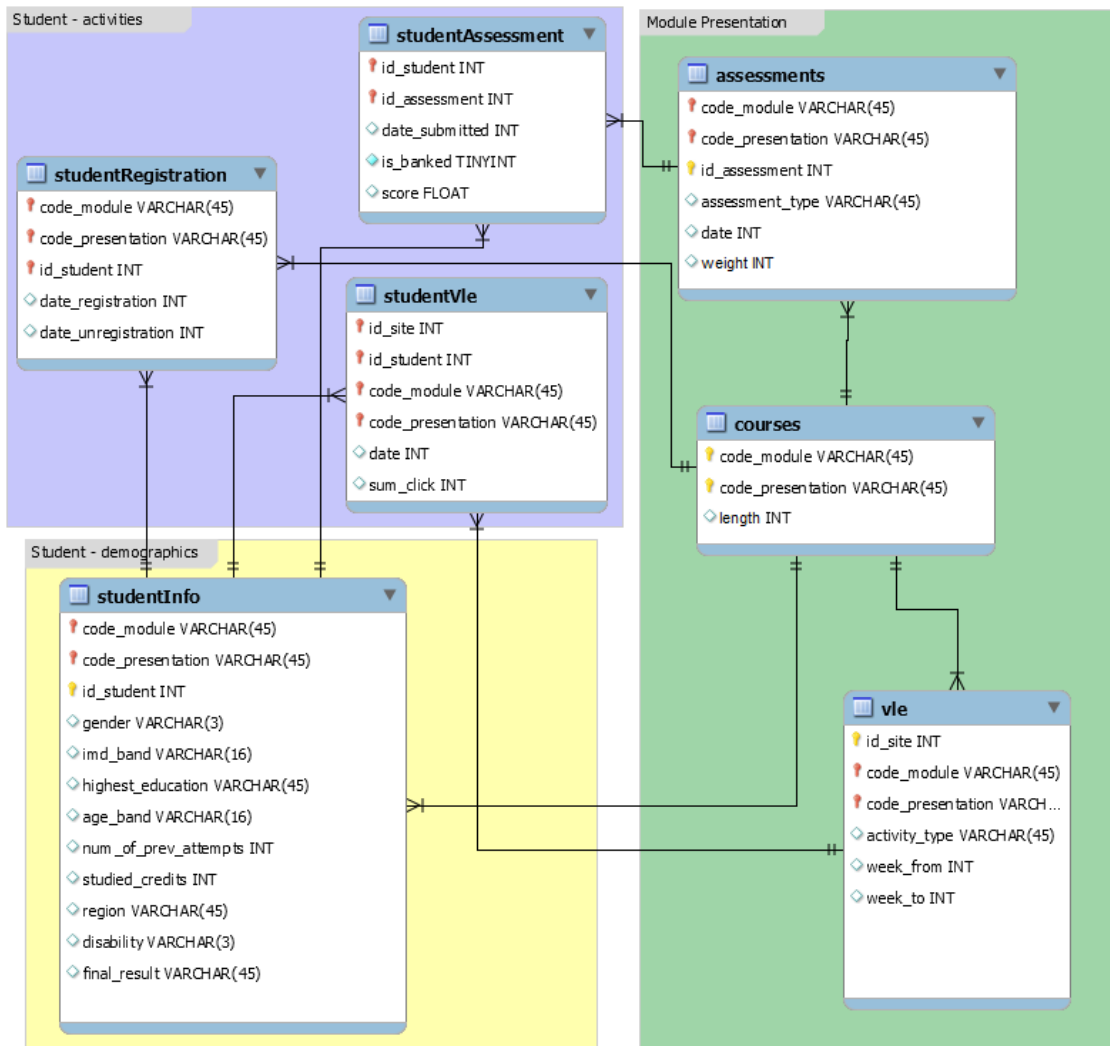


Figure 7. Database schema as shown in OU Analyse project's website

A separation of the tables into three distinct colour sections can be observed:

- Yellow for student demographics, involving information referred to the individuals taking the courses.
- Green for the module presentation, refers to the different elements which compose a course.
- Purple for student activities, relative to the interactions between the students and the course elements.

This, seen as a logical model proposal, identifies the three main entities from which data has been obtained (which, presumably for normalization purposes, was spread across the tables shown in the image).

As so, and for a better understanding of the dataset's domain, we can identify "module presentation" and "student demographics" as noun entities (identify specific elements: courses and students, respectively) and "student activities" as the verb entity (defines the interaction between noun entities).

This information will be relevant when defining the logical model for our proposed database model in further sections.

3.1.1 Domain

As a significant milestone in the assessment and posterior re-design of this project’s database, an explicit interpretation of its domain needs to be set.

The observation of the stored data’s nature reveals that this database is intended to contain historic data. This is, the behaviour and results obtained by a set of students throughout a certain time span.

One of the main implications to note is the way in which user data is stored: records related to each user’s information during a certain course are stored, meaning that records related to the registration of the same user (“student_id”) to a certain set of courses may be stored and, also, they may contain different user information (e.g. change of “age_band” from one year’s registration to another).

This clarifies the differences between a database of the mentioned type and a common user-registry database, in which changes in user data records would be stored as an update to the original record, and not as an additional one (as it happens in this case).

3.1.2 Table and columns detail

It is remarkable how the database schema shown in the webpage lacks critical information, such as primary keys for some tables and a clearer definition of elements of this kind, which will need to be set according to an assessment of the data contained in the table.

As a consequence, the table description detailed in this section has been made taking the following assumption into account:




-  - Primary key
-  - Foreign key
-  - Regular column

Figure 8. Database element’s type legend

To organize each table’s definition the following template has been used:

- **<Table name>: <description>**

Key	Column name	Data type	Description

Table 1. Table’s description template

- **Number of records:** <No.>
- **Observations:** <text>

Despite its attributes are self-descriptive, it is necessary to clarify the characteristics of some of them:

- **Key:** distinguishes whether a column will be a primary key (“PK”), a foreign key (“FK”) or none of them (“-”).
- **Observations:** details relevant information elicited through inspection of the data files and advances any modification that may be performed as a consequence.

Additional information relative to each table’s columns can be seen in OULAD’s dataset documentation ([46]).

- **Courses:** list of available modules and their presentations.

Key	Column name	Data type	Description
PK	code_module	VARCHAR	Module’s identifier.
PK	code_presentation	VARCHAR	Module’s presentation code.
-	length	INT	Length of the module-presentation in days.

Table 2. Course’s description table

- **Number of records:** 22
- **Observations:** none.

- **Assessments:** information about the assessments conducted in each course (module-presentation).

Key	Column name	Data type	Description
FK	code_module	VARCHAR	Module to which the assessment belongs.
FK	code_presentation	VARCHAR	Presentation to which the assessment belongs.
PK	id_assessment	INT	Identification number of the assessment.
-	assessment_type	VARCHAR	Type of assessment.
-	date	INT	Number of days since the start of the module-presentation for the final submission of the assessment.
-	weight	INT	Weight of the assessment in %.

Table 3. Assessment’s description table

- **Number of records:** 206
- **Observations:**
 - Presence of odd values: “date” column had occurrences in which this attribute contained the character “?”.
 - Although this may be considered as critical for data coherence, it is stated in the documentation provided by OULAD’s website

([45]) that <<If the information about the final exam date is missing, it is at the end of the last presentation week>>.

Hence, treatment related to this occurrence must take this into account.

It is present in 11 records ($\approx 5\%$).

- **Vle:** information about the available materials in the VLE.

Key	Column name	Data type	Description
PK	id_site	INT	Identification number for the material.
FK	code_module	VARCHAR	Module to which the Vle content belongs.
FK	code_presentation	VARCHAR	Presentation to which the Vle belongs.
-	activity_type	VARCHAR	Type of module material.
-	week_from	INT	Week from which the material is planned to be used
-	week_to	INT	week until which the material is planned to be used.

Table 4. Vle's description table

- **Number of records:** 6364
- **Observations:**
 - Presence of odd values: “week_from” and “week_to” columns had occurrences in which these attributes contained the character “?” (always simultaneously). Since their defined data type is INT, this event compromises data coherence.

Additionally, a further analysis of this occurrence revealed that it happened to 5243 rows, which is an 82% of the total amount. This elicits a considerable loss of relevance for these two attributes.

- **StudentInfo:** demographic information about the students and their collected results through courses.

Key	Column name	Data type	Description
FK	code_module	VARCHAR	Module on which the student is registered.
FK	code_presentation	VARCHAR	Presentation during which the student is registered to the module.
PK	id_student	INT	Identification number of the student.
-	gender	VARCHAR	The student's gender.
-	region	VARCHAR	Geographic region where the student took the course.
-	highest_education	VARCHAR	Highest student's education level.
-	imd_band	VARCHAR	Index of Multiple Deprivation band of the place where the student took the course.
-	age_band	VARCHAR	Band to which the student's age belongs.
-	num_of_prev_attempts	INT	Number of times a student has attempted a module.
-	studied_credits	INT	Total number of credits of the modules the student is currently studying.
-	disability	VARCHAR	Indicates whether the student has a declared disability.
-	final_result	VARCHAR	Student's final result in the course.

Table 5. Student Info's description table

- **Number of records:** 32593
- **Observations:**
 - Presence of odd values: "imd_band" column had occurrences in which this attribute contained the character "?". This only occurs in 1111 records ($\approx 3\%$), so "imd_band" column relevance is not compromised.
 - Duplicate keys: there were rows which presented duplicated values for the table's primary key "id_student". An in-depth analysis of the reasons for these occurrences concluded that the cause was the (reasonable) possibility for the same student to be registered in more than one course (same "id_student" for different "code_module" and "code_presentation" attributes). The need for a new primary key definition is extracted from this assessment.

- **StudentRegistration:** information about when each student registered to a certain course.

Key	Column name	Data type	Description
FK	code_module	VARCHAR	Identification code for the module to which the student registered.
FK	code_presentation	VARCHAR	Module's presentation to which the student registered.
FK	id_student	INT	Identification number of the student.
-	date_registration	INT	Date in which the student registered to the course.
-	date_unregistration	INT	Date in which the student abandoned the course (if he/she did).

Table 6. Student Registration's description table

- **Number of records:** 32593
- **Observations:**
 - **Presence of odd values:** “date_unregistration” column had occurrences in which this attribute contained the character “?”. Since its defined data type is INT, this event compromises data coherence.

Despite this happens to 22668 of the rows in this table ($\approx 70\%$) and may be considered as a reason to question the relevance of this attribute, it is importance to take context into account for this assessment: this column not having information for a given row doesn't mean the information is missing, it means that the correspondent student didn't unregister.

It can be concluded the need for this column to be interpreted as dependant on a binary context (to have unregistered or not), to which adds information.

Additionally, “date_registration” presented 45 occurrences in which it contained “?” as value.

This roughly represents 0.15% of the rows in this table, so column relevance is not compromised.

- **Lack of primary key definition:** in this case, records for “date_registration” and “date_unregistration” are collected for a given individual enrolled in a course. Thus, “code_module”, “code_presentation” and “id_student” would act as the primary key for this table.

- **StudentAssessment:** contains the student’s results for each assessment they were involved in.

Key	Column name	Data type	Description
FK	id_assessment	INT	Identification number of the assessment a student is assigned to.
FK	id_student	INT	Identification number of the student.
-	date_submitted	INT	Date in which the student submitted the assessment.
-	is_banked	INT	Status flag indicating that the assessment result has been transferred from a previous presentation.
-	score	INT	The student’s score in this assessment.

Table 7. Student Assessments' description table

- **Number of records:** 173912
- **Observations:**
 - **Presence of odd values:** “score” column had occurrences in which this attribute contained the character “?”. Since its defined data type is INT, this event compromises data coherence.

This event’s occurrence is minimal, happening to only 173 records ($\approx 0.01\%$), so “score” column’s relevance is not compromised.
 - **Data coherence:** it was observed that some student IDs present in “StudentInfo” table had no record in “StudentAssessment”.

Although this may lead to a hypothesis regarding lack of data coherence, an evaluation of the context related to these records reveals that this only happens to students who unregistered from the course before submitting any assessment.

As no assessments from these students were submitted during those courses they unregistered from, it makes sense that there are no records of them in this table.
 - **Lack of primary key definition:** essentially, this table stores information for different (specific) student’s assessment. Consequently, “id_assessment” and “id_student” columns have together the properties of a primary key.
 - **Completeness of foreign key:** given the fact that this table references StudentInfo, this shall be done by using its complete primary key which, apart from “id_student”, includes “code_module” and “code_presentation”. These last two attributes need to be added to StudentAssessment.

This may not be considered as an error from the source’s database schema definition, but as a consequence of the previous primary key definitions we have previously made.

- **StudentVle:** information about each student’s interactions with the materials in the VLE.

Key	Column name	Data type	Description
FK	code_module	VARCHAR	Module to which the VLE content belongs.
FK	code_presentation	VARCHAR	Presentation to which the Vle belongs.
FK	id_student	INT	Identification number of the student.
FK	id_site	INT	Identification number of the VLE material.
-	date	INT	Date of the student’s interaction with the material.
-	sum_click	INT	Number of times a student interacts with the material.

Table 8. Student Vle's description table

- **Number of records:** 10655280
- **Observations:**
 - Data redundancy: this table stores information about the interactions of a student with VLE content within a given day. Nevertheless, the original source contains different records for a student interaction within the same day (same course and VLE content).
This generates a problem with data redundancy, since it would be more efficient to store the sum of all interactions within the same day (for the same student, course, and VLE content) as a single record.
 - Lack of primary key definition: in coherence with what was pointed out in the previous observation regarding redundancy, the information present in this table is considered as the record of the number of interactions (“sum click”) of a student with a certain VLE content from a course on a given date. According to this, “code_module”, “code_presentation”, “id_student”, “id_site” and “date” would compound the primary key for this table.

3.1.3 Additional observations

In addition to the table analysis previously outlined, other factors not strictly related to tables individually have been assessed.

Correspondence between studentInfo and studentRegistration

While inspecting relationships between tables, it was observed that studentInfo and studentRegistration tables were likely to present a one to one correspondence between them, being the exact same number of records for each one of these tables the first clue to intuit this occurrence.

More specifically, the proposed hypothesis states that every student whose details are stored in studentInfo’s table has a corresponding record in studentRegistration’s table.

For this purpose, the following methodology, which was executed by a R script, has been employed:

- Inner join both tables, so that coinciding elements between tables will persist to the result of this operation.

Since both tables are uniquely identified by the attribute set composed by “code_module”, “code_presentation” and “id_student”, they will be the ones by which the join operation will be executed.

- This procedure was conducted by using “merge” R’s function which, used with the following syntax, allows to perform inner joins:

```
merge(<table1>, <table2>)
```

Figure 9. Merge R function

- In case the mentioned result contains the same number of rows as the tables, the hypothesis will be confirmed.

After the mentioned process, the following result was obtained:

innerJoin	32593 obs. of 14 variables
studentInfo	32593 obs. of 12 variables
studentRegistration	32593 obs. of 5 variables

Figure 10. Inner join between studentInfo and studentRegistration’s tables

The conditions for confirming our previous hypothesis are met (same number of observations/rows for the inner join than for the tables separately), meaning that a one to one correspondence between studentInfo and studentRegistration’s tables can be stated.

This conclusion also implies that the one-to-many relationship from StudentInfo to StudentRegistration shown in OULAD’s website logical model (figure 7) should be a one-to-one relationship instead.

** The resulting 14 variables shown for the innerJoin data frame is the result of the sum of the identifying set previously mentioned (“code_module”, “code_presentation” and “id_student”, 3 attributes) and the remaining variables from both tables (9 from studentInfo and 2 from studentRegistration).*

It is relevant to point out the format in which dates are presented in this dataset: they’re all presented as numbers representing the count of days since the start of the correspondent course.

This format is neither descriptive nor useful in which respects to the analytics tasks intended to be performed (such as time series forecasting). Consequently, it is necessary to set a proper date format representing equivalent information.

This, and other pre-processing operations, are described in the following section.

3.2 Data pre-processing and other tasks performed

This first section shows the different procedures developed as a response to the previously note table observations. Its structure will be based on the following template:

<tableName>

<problemObserved>

<solution>

<result>

Previous to the detail of the mentioned procedures, with a purpose of clarity and objectiveness, the main points considered for the treatment of odd values from different columns is presented:

a. Significance of the number of occurrences for the given odd value.

I. Removal may be considered if the rate of occurrence is considered to be high for the specific column.

b. Explicit dependencies (detailed in OULAD’s documentation, [26], or intuitable through simple inspection) of the affected column with other attributes in the database.

I. In case a dependency exists, an in-depth analysis would need to be performed.

Although these points have been taken into account throughout the development of the following assessment, the purpose for their presentation is to make the processes and conclusions involved more understandable, and they are not going to be reviewed in a categorical way.

- **Courses**

No changes were needed.

- **Assessments**

- **Odd values in “date” column**

In coherence with OULAD’s website documentation ([26]) with respect to these occurrences (which says that missing dates correspond to the end of the last presentation week), missing values were substituted by the last day correspondent to the course referenced in that row. This information was retrieved from courses.csv.

code_module	code_presentation	id_assessment	assessment_type	date	weight
AAA	2013J	1757	Exam	-1	100

code_module	code_presentation	module_presentation_length
AAA	2013J	268

Figure 11. Fixing process of assessments table’s odd values

code_module	code_presentation	id_assessment	assessment_type	date	weight
AAA	2013J	1757	Exam	268	100

Figure 12. Fixed assessments table's odd values

- **Vle**
 - **Odd values in “week_from” and “week_to” column**
As mentioned in the previous section, ≈82% of the records for these columns contain odd values.
This elicits an excessive loss of relevance for these columns, resulting in their removal from the data collection.
- **StudentInfo**
 - **Odd values in “imd_band” column**
With this event having such a low rate of occurrence (≈3%), no treatment further from the substitution of “?” values by “Unknown” (on account of descriptiveness) was conducted.
 - **Primary key definition**
As reasoned in the previous section, “code_module” and “code_presentation” columns were added to the table’s primary key. This solves the problem with duplicate keys.
The resulting primary key is defined by: “id_student”, “code_module” and “code_presentation”.
- **StudentRegistration**
 - **Odd values in “date_unregistration” column**
Records containing “?” were substituted by “-999” to keep coherence with the specified data type (INT).
 - **Odd values in “date_registration” column**
Records containing “?” were substituted by “-999” to keep coherence with the specified data type (INT).
 - **Primary key definition**
As reasoned in the previous section, “code_module”, “code_presentation” and “id_student” have been defined as the primary key for this table.
- **StudentAssessment**
 - **Odd values in “score” column**
It is important to remark the direct dependency of this column’s values with that of “final_result” (from studentInfo’s table).

$$\text{assessment1_score} * \text{assessment1_weight} + \dots + \text{assessmentN_score} * \text{assessmentN_weight} = \text{total_score} \rightarrow \text{final_result (Pass, Fail...)}$$

Figure 13. Calculation of “final_result” from student results

However, the lack of presence of odd values for “final_result” column (as seen in the previous section’s observations) elicits that this occurrence doesn’t directly affect its content (the way it affects final marks calculation doesn’t lie within the scope of this analysis). Thus, presence of odd values in “score” column can be considered as exclusively referred to studentAssessment records.

As a result, and considering the minimal rate of appearance of these odd values ($\approx 0.01\%$), records containing “?” were substituted by “-1” to keep coherence with the specified data type (INT).

- **Primary key definition**

As reasoned in the previous section, “id_assessment” and “id_student” have been defined as the primary key for this table.

- **Foreign key completeness**

The addition of the columns “code_module” and “code_presentation” to this table, and its definition as part of its foreign key solves the problem with the incomplete reference to StudentInfo table.

The resulting foreign key is defined by: “id_student”, “code_module” and “code_presentation” (same as StudentInfo’s primary key, formerly defined in this section).

- **StudentVle**

- **Primary key definition**

As reasoned in the previous section, “code_module”, “code_presentation”, “id_student”, “id_site” and “date” have been defined as the primary key for this table.

- **Data redundancy**

To better understanding of the problem, a sample of the unprocessed data collection is shown:

```
"AAA", "2013J", "28400", "546652", "-10", "4"  
"AAA", "2013J", "28400", "546652", "-10", "1"  
"AAA", "2013J", "28400", "546652", "-10", "1"
```

Figure 14. Redundant records (viewed as text)

It can be observed that, for the same primary key (“code_module”, “code_presentation”, “id_student”, “id_site” and “date”), there are multiple records with different “sum_click” values.

The solution consisted in the sum of the “sum_click” values for each repeated primary key and its assignation to a single record of the correspondent key.

3.2.1 Ensuring data consistency

One of the main factors assessed during the inspection process underwent by the database was data consistency.

Data consistency refers to the requirement that any given database transaction must change affected data only in allowed ways (ensuring that no database constraints are violated).

Figure 15. Data consistency definition

Under the scope of this term’s definition, the tasks presented hereunder have had as a main objective the fulfilment of the constraints assumed for this database (primary and foreign keys).

As done in the previous section, a template has been used as an organized display of the involved procedures’ detail:

<Problem>

Tables involved: <table name 1>...<table name N>

<description/text>

Treatment: <text>

- **Multiple correspondence failures between withdrawn students’ records and their un-registration date**

Tables involved: studentInfo and studentRegistration

According to the dataset’s documentation provided by OULAD’s website ([26]) and, more specifically, to that referred to studentRegistration’s “date_unregistration” attribute <<Students who unregistered have Withdrawal as the value of the final_result column in the studentInfo.csv file>> (it’s important to point out that the value present in the dataset was really “Withdrawn”, and not “Withdrawal”).

code_module	code_presentation	id_student	[...]	final_result
AAA	2013J	30268		Withdrawn

code_module	code_presentation	id_student	date_registration	date_unregistration
AAA	2013J	30268	-92	12

Figure 16. Correspondence between withdrawn students and their un-registration date

However, when assessing the accomplishment of this correspondence between tables, multiple critical occurrences were detected:

- 84 students with the “Withdrawn” tag present in their studentInfo’s record were missing a value for their correspondent “date_unregistration” attribute (from studentRegistration).

code_module	code_presentation	id_student	[...]	final_result
BBB	2014B	2512349		Withdrawn
code_module	code_presentation	id_student	date_registration	date_unregistration
BBB	2014B	2512349	-17	-999

Figure 17. Missing un-registration date for a withdrawn student

It's important to remember the change we made with respect to un-registration dates: missing values originally containing “?” changed to “-999” in order to keep coherence with the column's data type.

- 9 students with a specified value for their “date_unregistration” attribute (studentRegistration) weren't listed as “Withdrawn” in studentInfo's record. In fact, they were registered as “Fail” (didn't pass the course).

code_module	code_presentation	id_student	[...]	final_result
BBB	2013J	554243		Fail
code_module	code_presentation	id_student	date_registration	date_unregistration
BBB	2013J	554243	-35	166

Figure 18. Missing withdrawal tag for a student with an un-registration date assigned

Treatment: although there are cases in which it's possible to intuit the missing or inconsistent values from the records, this approach isn't appropriate (specially in terms of objectiveness).

As a result, deletion of these records has been the choice and, since few records (a total of 93) are affected, effect on posterior analytics will be minimal.

It is important to remark that this process also involved the removal of the correspondent records from other tables in the database. The tables affected were:

- studentInfo (origin): went from 32593 records to **32491**.
- studentRegistration (origin): went from 32593 records to **32491**.
- studentVle (consequence): went from 8459320 records to **8454770**.
- studentAssessment (consequence): went from 173912 records to **173852**.

3.2.2 Adding format to date values

As it has been already mentioned when detailing the main observations made on OULAD's dataset, the format used to represent time constraints is not helpful in any way (time spans between two dates may be useful in the future, but this information is easily extractable from a proper date format which would aid more processes).

This affects the following tables and attributes:

Table name	Attributes affected
courses	module_presentation_length
assessments	date_assessment
studentRegistration	date_registration, date_unregistration
studentAssessment	date_submission
studentVle	date_interaction

Table 9. Tables affected by date formatting process

The treatment proposed for this event is based on the following information extracted from OULAD’s dataset documentation ([26]):

Course’s table - code_presentation attribute: code name of the presentation. It consists of the year and “B” for the presentation starting in February and “J” for the presentation starting in October.

Figure 19. Information on courses’ identification format.

With completeness purposes, February 1st and October 1st (in addition to the year specified by “code_presentation”) have been assumed to be the exact dates in which courses start.

Knowing this, the process can be simply summarized to adding the value (number of days) contained in each of the attributes shown in the previous table to the starting date of the correspondent course.

code_module	code_presentation	module_presentation_length
AAA	2013J	268

↳ 2013-10-01 + 268 days

code_module	code_presentation	end_date
AAA	2013J	26/06/2014

Figure 20. Modification of date format

The above figure describes the methodology in which the process, which was conducted through a R script, is based. Additionally, it is important to notice the fact that the columns containing the original format are removed after this process (as it has been already mentioned, obtaining this information is almost trivial and there is no reason to keep it as a column in our tables).

Also, as “code_presentation” attribute already represents the start date for each course, there is no need to store an additional column referred to this information (for none of the tables affected).

In this specific case (referred to Course’s table), it can be seen that the name of the column changed, but this is an exception which obeys to descriptiveness purposes.

3.3 Result and first logical model

As a result of the modifications previously detailed, an initial logical model (to use as a starting point for further analysis and modifications) is proposed.

In regard to entity definition, it's important to recall the observations made when we first analysed OULAD's proposed database model. With descriptiveness purposes, the following identifiers have been changed:

- Module presentation -> Course
- Student activities -> Interaction
- Student demographics -> Student

Entity's name	Entity type
Course	Noun
Student	Noun
Interaction	Verb

Table 10. Entities' type definition

With this data, the following entity-relationship diagram is elaborated:

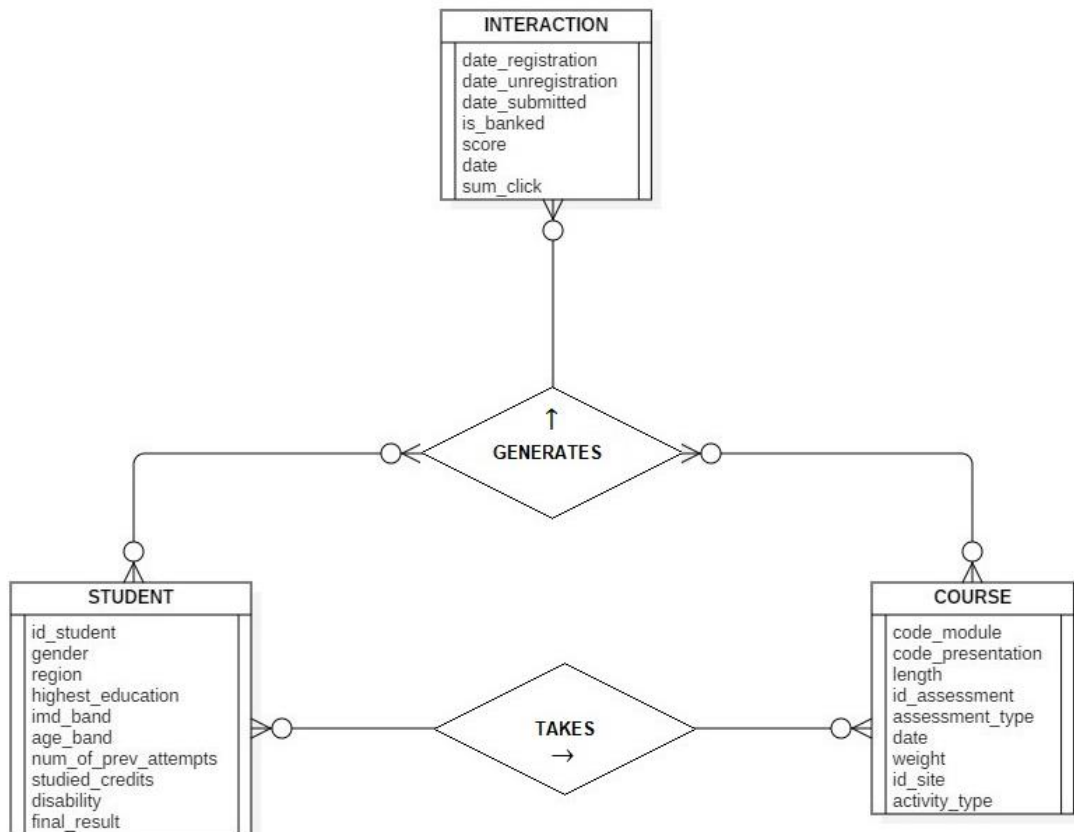


Figure 21. Database's entity-relationship diagram

Attending to it, and taking advantage of the table definition we already know, the following logical model can be elicited:

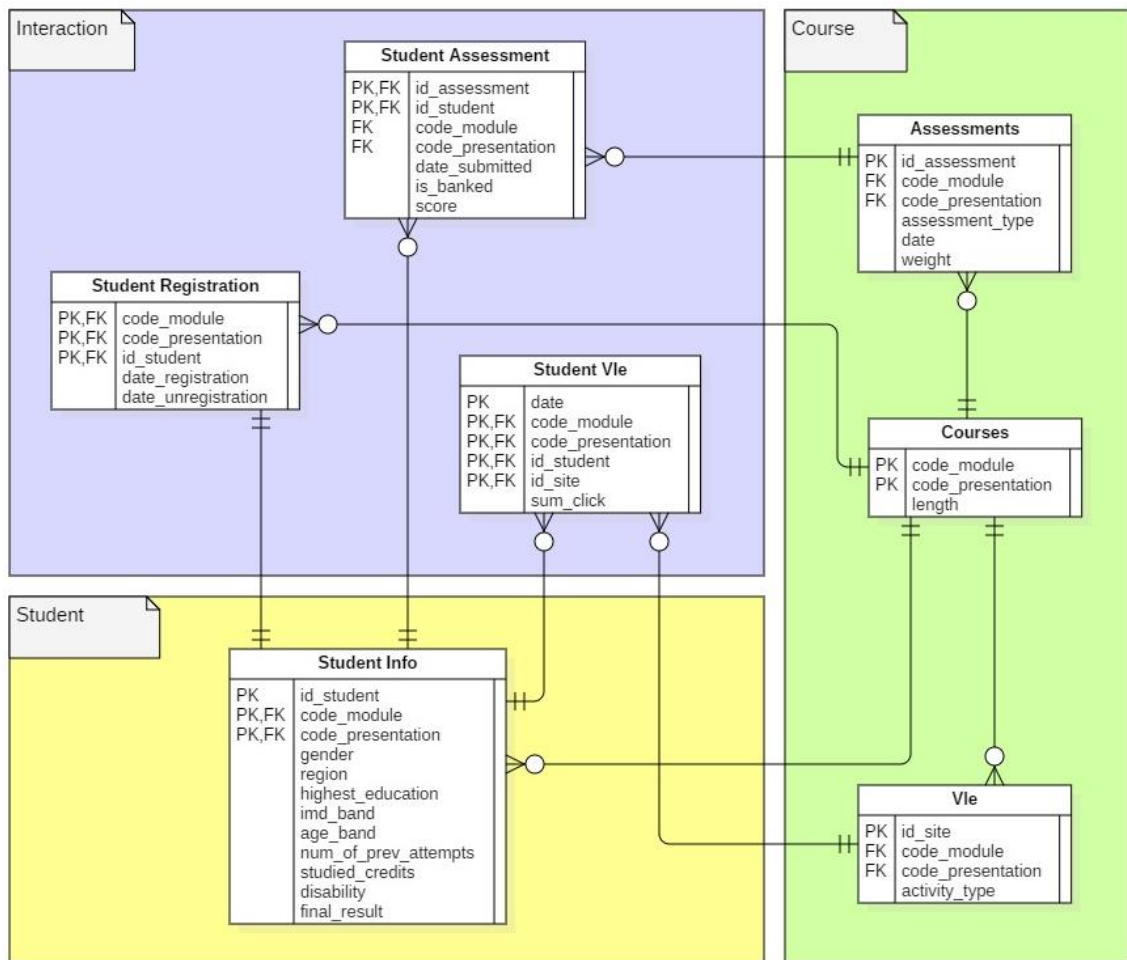


Figure 22. Database's logical model

As it has been already pointed out, having a proposed database model beforehand (with data spread across different tables), permits to save a great amount of time and effort, which otherwise had to be placed into normalizing the dataset from its raw form.

However, in order to guarantee its accomplishment, normalization will still be reviewed hereafter and, as a result, this logical model is subject to change.

3.4 Database normalization

Although database normalization objectives are multiple and aim to meet efficiency metrics of different kind (performance, data consistency, etc.), some of the processes involved in their fulfilment may not fall within the scope of this project, but it does intersect with our interests in which respects to data inspection and retrieval. This is, the ease with which a query can be expressed (or it's expressive power).

In consonance with this, a series of iterative steps have been performed, in which accomplishment of the three main levels of database's normal forms are checked:

- 1NF
- 2NF

- 3NF

The attainment of the properties correspondent to each of these levels results in a more efficient (logical) arrangement of the data, being then possible to objectively guarantee:

- Minimization of data redundancy, which leads to improved data consistency.
- Robustness in which regards to possible insertion, deletion or modification anomalies.
- Descriptiveness of the model.

This will favour the treatment (querying) of the data stored in our database, as well as the elicitation of information from it.

3.4.1 First Normal Form (1NF)

Apparent lack of consensus with respect to the criteria referred to this level of normalization was observed while inspecting different web resources ([27]). An in-depth search and analysis of literature ([28]) allowed, firstly, to clarify the definition assumed for some terms widely (mis)used in 1NF reviews:

- Repeating groups: a column that can accommodate multiple values ([29]).
- Atomicity constraint: any row and column intersection in a table contains exactly one value of the applicable type (which can be arbitrarily complex) ([30]).
- Null values violate 1NF since there's a constraint indicating that columns must be typed (and a null value has no defined type) ([29]).
 - In this respect, and related to the database reviewed in this project, we are not directly treating null values, but odd values instead (which differ with the former in that they're typed).
Despite they're generally considered as part of any SQL-DB, we'll discuss their occurrences and treatment in order to minimize their presence and avoid the consequent logical implications.

In addition, this research process allowed to elicit the following main points to fulfil:

1. A table must have no duplicates (rows or columns).
 - Column uniqueness can be assumed, since the previous review of each table's composition reveals how each contains different attributes
2. There must be no significance in the order of either rows or columns.
 - The occurrence of this situation would elicit a critical lack of data consistency. Previous database review reveals independence between attributes, and no dependency between rows of the same table could exist (which would cause the need for a specific order).
Thus, this property will also be assumed.
3. In every row, each column must have a single value (with columns named and typed).

- Since the dataset provided by OULAD’s website is SQL-based, and the columns in a SQL table are explicitly named and typed, only a single value of that type can be accommodated in a single cell ([29]).

This information can be used to infer the accomplishment of this point beforehand.

Following, the remaining condition regarding row uniqueness will be checked on each of the tables from the database’s model described in *Figure 22*. Additionally, and as previously indicated, presence of odd values will be reviewed.

Row uniqueness checking

The methodology for this task has consisted in a two-step process:

1. Counting of unique rows within a table.
 - For this purpose, a R script has been used, which relies on the complement of the *duplicated* function’s result:

```
<table_name>[!duplicated(<table columns to take into account>), ]
```

Figure 23. R function: "duplicated"

This operation will retrieve a data frame containing those values which are not duplicated from the specified table.

2. Comparison of the value obtained with the total number of rows from that table.
 - If the values coincide, then the condition of uniqueness is fulfilled.

After loading the different datasets (each corresponding to one of our database’s tables) into our R environment, an initial count for the number of rows contained in each of them (number of observations) can be seen:

assessments	206 obs. of 6 variables
courses	22 obs. of 3 variables
studentAssessment	173852 obs. of 7 variables
studentInfo	32491 obs. of 12 variables
studentRegistration	32491 obs. of 5 variables
studentVle	8454770 obs. of 6 variables
vle	6364 obs. of 6 variables

Figure 24. Loaded tables prior to row uniqueness checking

Then, the negated *duplicated* function is applied to each of the previously loaded data frames, generating a set of them with all possible duplicated row removed:

UNIQUEROWS_assessments	206 obs. of 6 variables
UNIQUEROWS_courses	22 obs. of 3 variables
UNIQUEROWS_studentAssessment	173852 obs. of 7 variables
UNIQUEROWS_studentInfo	32491 obs. of 12 variables
UNIQUEROWS_studentRegistration	32491 obs. of 5 variables
UNIQUEROWS_studentVle	8454770 obs. of 6 variables
UNIQUEROWS_vle	6364 obs. of 6 variables

Figure 25. Result of the application of R’s “duplicated” function for row uniqueness checking

As it can be seen, the number of observations for each data frame doesn't vary from its initial state to the application of the *duplicated* function.

Therefore, accomplishment of the assessed condition has been successfully verified.

Odd values assessment

In accordance with the initial analysis of odd values made in this document and its posterior treatment (mainly based on replacement to fulfil data-type constraints), the following tables will be reviewed:

- studentInfo and its "imd_band" attribute
- studentRegistration and its "date_registration" and "date_unregistration" attributes

None of the remaining tables presented, at this point, odd values in their cells.

For structuring this section, the following template has been used:

<table's name>

Attribute involved: <attribute's name>

Number of occurrences: <occurrences>/<total number of rows in the table>
(<percentage>)

Treatment: <text>

** "Attribute involved" and "Treatment" sections will appear repeated for each table's attribute being assessed.*

- **studentInfo**

Attribute involved: imd_band

Number of occurrences: 1110/32491 (≈ 3%)

Treatment: being an attribute which represents a property of every row in the table and not having any information to re-assign or infer a value for these cells led us to preserve this column as it is. Also, no practical re-distribution of the data across a new logical structure would solve this situation.

A significant advantage is that the portion of data affected is almost irrelevant, so no impact on subsequent analytics is expected.

- **studentRegistration**

Attribute involved: date_registration

Number of occurrences: 45/32491 (≈ 0.15%)

Treatment: the same reasoning applied for studentInfo's "imd_band" attribute was used in this case to preserve the column with no treatment.

The relevance of this situation is even lower than that of "imd_band".

Attribute involved: date_unregistration

Number of occurrences: 10068/32491 ($\approx 30\%$)

Treatment: in this case, the occurrence's high rate of appearance, summed to the fact that this attribute does not represent every row in the table (not all students unregistered, so not all students should have an un-registration date assigned) are the main reasons behind the decision to re-distribute this data into a new logical structure.

The new data arrangement for this table will be the following:

- studentRegistration table, with all its original attributes except date_unregistration:

Column name	Data type
code_module	VARCHAR
code_presentation	VARCHAR
id_student	INT
date_registration	INT

Table 11. Modified StudentRegistration table

The table will also contain the same records as before (no reduction in number occurred).

- studentUnregistration table, which will contain the un-registration date referred to those students from studentInfo's table who unregistered (and not every student, as before):

Column name	Data type
code_module	VARCHAR
code_presentation	VARCHAR
id_student	INT
date_unregistration	INT

Table 12. StudentUnregistration table

According to its description, this dataset will contain a total of 10068 records (the number of rows from the original studentRegistration's table referred to an unregistered student).

To grant descriptiveness to the result of this process, the logical representation of the changes made is presented:

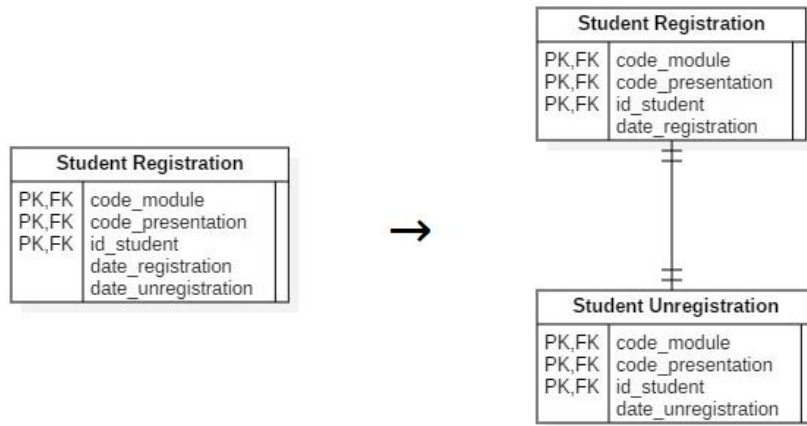


Figure 26. Change from the initial studentRegistration data arrangement (left) to its modification (right).

This new distribution, apart from solving the problem with the high presence of odd values in “date_unregistration” attribute, represents a more logically distributed model and eases the treatment of the information related to unregistered students (which involves a critical source of knowledge for this project, such as course abandoning and thus, it will be oftenly queried).

The re-arrangement of data described was processed by an R script.

At this point, the fulfilment of the conditions required to guarantee the attainment of 1NF normalization level has been checked.

3.4.2 Second Normal Form (2NF)

In this case, with 1NF verified, it’s needed to remove any partial dependency from the model’s tables (if there’s any).

Partial dependencies occur when an attribute in a table depends on only part of the primary key (and not the whole primary key).

Figure 27. Partial dependency definition

On account of simplicity and relevance, only the assessment of tables in which such events were detected will be shown in this section. A similar template as the one in 1NF section has been used:

<table’s name>

<logical structure/image>

Attributes involved: <attribute’s name 1>...<attribute’s name N>

Type of occurrence: <text>

Treatment: <text>

- **studentInfo**

Student Info	
PK	id_student
PK,FK	code_module
PK,FK	code_presentation
	gender
	region
	highest_education
	imd_band
	age_band
	num_of_prev_attempts
	studied_credits
	disability
	final_result

Figure 28. StudentInfo's table.

Attributes involved: gender, region, highest_education, imd_band, age_band and disability.

Type of occurrence: the detailed attributes only logically relate to “id_student”, since they are properties which belong only to the student, and there is no possible dependency or relation to establish with the course itself (identified by “code_module” and “code_presentation”, the other part of the primary key for this table).

Treatment: with the purpose of re-distributing the data into a new arrangement that solves this occurrence, a re-interpretation of user data's nature was needed.

In coherence with this database's domain (previously detailed in the correspondent section) and recalling the fact that the same user could be registered with different personal data (which comprises the attributes mentioned in this case), it can be understood that student data belongs to a certain user-registration process, and not directly to the user (meaning that these registrations, or sign-ups, would need their own id).

This proposal was conducted with studentInfo's table as its origin and, since it's concerned with students' registration to a course, it directly involved studentRegistration. Thus, the following distribution of the database's structure was implied:

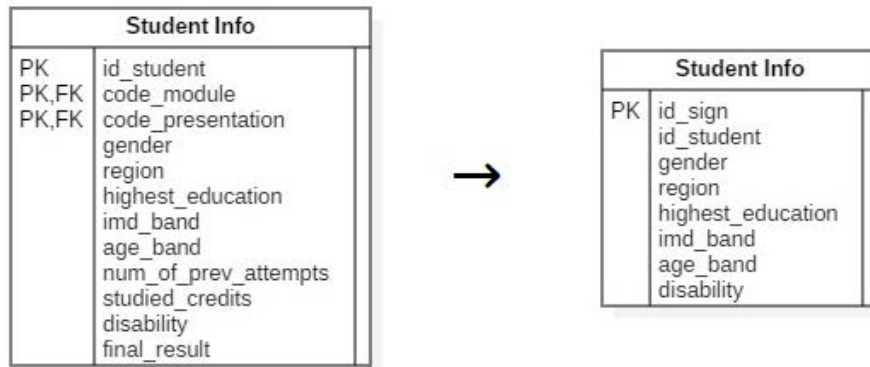


Figure 29. Change from the initial studentInfo data arrangement (left) to its modification (right).

At this point, the main changes that can be observed are the following:

- Inclusion of “id_sign” attribute, which will identify each unique student data entry in the table.

To adequately explain the main direct implication this involves, it is important to recall the one to one correspondence (exposed in previous additional observations’ section) existing between studentInfo and studentRegistration’s tables.

Existing the possibility that some studentRegistration records represent the enrolment of a single student to multiple courses, it is also probable that the correspondent records stored in studentInfo contain repeated user data.

This can be proven by making use of the already mentioned “duplicated” R function (negated to obtain distinct rows as a result) on studentInfo’s table, specifically on those attributes representing student information (“gender”, “region”, “highest_education”, “imd_band”, “age_band” and “disability”, apart from “id_student” to identify each information set).

After the execution of the correspondent script, the following results were obtained:

studentInfo	32491 obs. of 13 variables
uniqueUserData_studentInfo	<u>28782</u> obs. of 13 variables

Figure 30. Unique user information from studentInfo

It can be observed that, indeed, studentInfo’s table stores repeated data for certain users (when they are registered to more than one course with the same personal data).

With this confirmed, it is now important to note that new StudentInfo’s records will be unique in which respects to student’s information, not as in the original table, which contained a row for each student’s registration to a course, regardless if that information was already stored or not.

- Removal of “studied_credits” and “num_of_prev_attempts” as advanced in the previous analysis. They have been re-interpreted as referring to a student’s registration to a course, not directly to the student.

- Removal of course information (“code_module” and “code_presentation”), which allows this table to get rid of the partial dependencies previously described.
- Removal of “final_result”, which has also been re-interpreted as a consequence of the interaction of a student with a course, and not as a student’s property.

Now the implications of this changes will be shown, as well as their relationship with studentRegistration’s table:

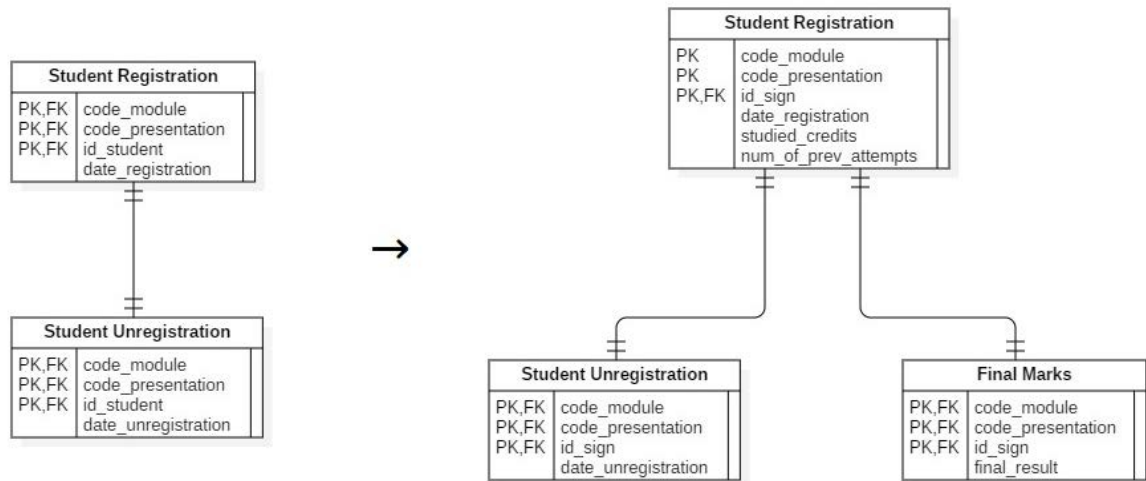


Figure 31. Second change from studentRegistration’s data arrangement (left) to its modification (right)

The changes shown in this figure respond to a series of requirements generated by the new data arrangement:

- Replacement of “id_student” by “id_sign” (foreign key from studentInfo’s table), as registry to a course will now reference a unique student information record.

This is a consequence of the changes detailed for studentInfo’s table referred to students’ information uniqueness.

- Inclusion of “studied_credits” and “num_of previous_attempts” attributes, which were previously removed from studentInfo’s table. They will now relate to a student enrolled in a course (and his/her personal information) via “id_sign”.

- Addition of a new table, finalMarks, which will contain the “final_result” for each student enrolled in a course, with the slight difference that this attribute will now only be present for those students who didn’t unregister from a course (in other words, those who completed the course they were enrolled in).

This implies that those records for “final_result” containing the “Withdrawn” tag were removed, as they related to unregistered students (which information is adequately stored in studentUnregistration’s table).

- Change of each table’s primary key, now including “id_sign” instead of “id_student”, accordingly to the unique identification needed for each row.

Finally, it is important to note that the described changes also modified the way the involved tables related to other. Specifically:

- **studentVle**

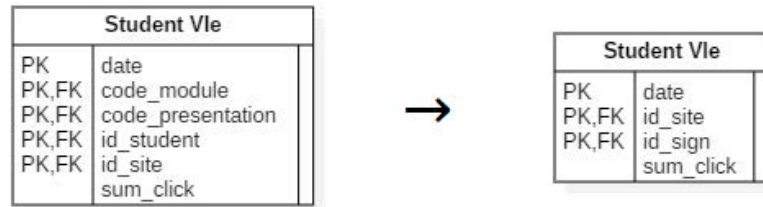


Figure 32. Change from studentVle’s data arrangement (left) to its modification (right)

- **StudentAssessment**

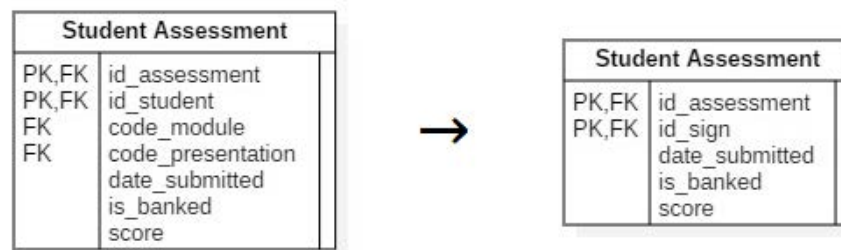


Figure 33. Change from studentAssessment’s data arrangement (left) to its modification (right)

It can be seen, in both cases, that “id_student”, “code_module” and “code_presentation” attributes were removed in favour of “id_sign”. This is due to studentInfo’s new primary key definition, which involved exactly the same change and, since both studentAssessment and studentVle refer to studentInfo, the foreign key through which they do consequently changed.

With the detailed partial dependency solved as shown, and with no more detected occurrences of this type throughout the database, it can be stated that 2NF normalization level for our database has been attained.

3.4.3 Third Normal Form (3NF)

After ensuring that 2NF for our model is fulfilled, absence of transitive dependencies needs to be checked.

Transitive dependencies occur when a non-prime attribute (not part of the primary key) depends on other non-prime attributes rather than depending on the primary key.

Figure 34. Transitive dependency definition

Inspection of the database’s structure (2NF) and its attribute’s relationships concluded that no transitive dependencies are present in it.

Only a possible case was observed in studentInfo’s table for “imd_band” and “region” attributes, which descriptions (according to OULAD’s website, [26]) are the following:

- region: geographic region where the student took the course.
- imd_band: Index of Multiple Deprivation band of the place where the student took the course.

These definitions may lead to think that “imd_band” is set according to the same zone specified in the “region” attribute, and thus depends on it, eliciting a transitive dependency.

However, further research regarding the terms in which “imd_band” is defined revealed that the areas to which it refers doesn’t strictly match those defined by “region”. More specifically, <<it is the official measure of relative deprivation for small areas in England >> ([31]), with these areas varying in in extension (from neighbourhoods to local councils).

This added to the presence of only 13 different regions in studentInfo’s table (as shown in below figure) and the fact that no detail about the local council to which each “imd_band” value belongs makes it impossible to establish a correspondence between these two attributes, eliminating the possibility that a transitive dependency is involved.

studentInfo	32491 obs. of 13 variables
uniqueRegions_studentInfo	<u>13 obs. of 13 variables</u>

Figure 35. Unique regions present in studentInfo’s table

As a conclusion, it can be concluded that our database meets the conditions needed to be considered as structured in 3NF normalization level.

3.4.4 Additional observations

After the previously specified re-arrangement of the database and its implementation as a SQL model, access and evaluation of its content became easier, which, given the significant amount of data with which we count, aided processes referred to this type of tasks.

The detail of these procedures is shown in this section, previous to the setting of the final database’s logical and physical models, to avoid the unnecessary replication of information its exposition before and after the modifications here described would entail (no structural changes took place).

3.4.4.1 Validating assessments' weights

Two of the most sensitive items from this dataset are assessments and its referred weights, since the final result label assigned to each student depends on them. As so, annotations from OULAD's dataset documentation ([26]) with respect to these elements are of special importance:

Assessment's table – assessment_type attribute: three types of assessments exist: Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Final Exam (Exam).

Assessment's table – weight attribute: weight of the assessment in %. Typically, Exams are treated separately and have the weight 100%: the sum of all other assessments is 100%.

Figure 36. Information on assessments' types and weights.

These observations need to be assessed in order to guarantee proper data integrity, purpose for which Assessment's table was examined, in conjunction with its correspondent information stored in StudentAssessment (score of specific assessments) and StudentInfo's (student's final results) records.

Along this process, critical events compromising the reliability of the previously shown documentation were detected:

- Courses from GGG module ("code_module") had 0 % assigned as the weight for every assessment other than the final Exam (100 %) for all its presentations ("code_presentation": 2013J, 2014B and 2014J).

code_module	code_presentation	id_assessment	assessment_type	weight	date_assessment
GGG	2013J	37418	CMA	0	18/05/2014
GGG	2013J	37419	CMA	0	18/05/2014
GGG	2013J	37420	CMA	0	18/05/2014
GGG	2013J	37421	CMA	0	18/05/2014
GGG	2013J	37422	CMA	0	18/05/2014
GGG	2013J	37423	CMA	0	18/05/2014
GGG	2013J	37415	TMA	0	01/12/2013
GGG	2013J	37416	TMA	0	02/02/2014
GGG	2013J	37417	TMA	0	23/03/2014
GGG	2013J	37424	Exam	100	18/05/2014

Figure 37. Example of the occurrence of assessments with unassigned weight

Although it would make sense that these courses based their results solely on exam's results, further inspection revealed that no assessment other than CMA and TMA were done during those courses. This fact, added to the presence of both students who failed and passed these courses, deprives the effects of both final exams and weights (from any type of assessment) from interpretability and leads to the conclusion that only CMA and TMA assessments should be the only assessments taken into account as a general rule.

```

SQL> SELECT count(*) FROM studentAssessment
2  INNER JOIN assessments
3  ON studentAssessment.id_assessment = assessments.id_assessment
4  WHERE code_module='GGG'
5  AND assessment_type != 'CMA' AND assessment_type != 'TMA'
6  ;

```

COUNT(*)
0

Figure 38. SQL query showing that no assessments distinct from CMAs and TMAs were conducted for GGG modules.

```

SQL> SELECT count(*) FROM studentAssessment
2  INNER JOIN assessments
3  ON studentAssessment.id_assessment = assessments.id_assessment
4  WHERE code_module='GGG'
5  AND assessment_type != 'Exam'
6  ;

```

COUNT(*)
15211

Figure 39. SQL query showing the number of assessments distinct from Exams conducted for GGG modules.

```

SQL> SELECT * FROM finalMarks
2  WHERE code_module='GGG'
3  AND ROWNUM <= 10
4  ;

```

ID_SIG	COD	CODE_	FINAL_RESUL
S16332	GGG	2013J	Pass
S16009	GGG	2013J	Fail
S15852	GGG	2013J	Pass
S15987	GGG	2013J	Fail
S9434	GGG	2013J	Distinction
S20161	GGG	2013J	Pass
S25385	GGG	2013J	Pass
S20838	GGG	2013J	Pass
S6728	GGG	2013J	Fail
S4113	GGG	2013J	Pass

10 rows selected.

Figure 40. SQL query showing the variety of results for GGG modules' students (limited to first 10 occurrences).

- Reinforcing the previous statement, 4 students who only did an Exam (assessment type) during the whole course they were registered to were identified, with different final results assigned, apparently regardless of these exam's scores.

Their detection came along during the comparison of the number of students who finished their course (which means they have records in finalMarks' table) and did any type of assessment and those who did any type of assessment but Exams.

```

SQL> SELECT COUNT(*) FROM
2 (SELECT distinct Q.code_module, Q.code_presentation, finalmarks.id_sign FROM finalmarks
3 INNER JOIN
4 (SELECT * FROM studentassessment
5 INNER JOIN assessments
6 ON studentassessment.id_assessment=assessments.id_assessment) Q
7 ON Q.id_sign=finalmarks.id_sign AND Q.code_module=finalmarks.code_module AND Q.code_presentation=finalmarks.code_presentation)
8 ;

COUNT(*)
-----
21148

```

Figure 41. SQL query showing the number of students who finished a course performing any type of assessment.

```

SQL> SELECT COUNT(*) FROM
2 (SELECT distinct Q.code_module, Q.code_presentation, finalmarks.id_sign FROM finalmarks
3 INNER JOIN
4 (SELECT * FROM studentassessment
5 INNER JOIN assessments
6 ON studentassessment.id_assessment=assessments.id_assessment) Q
7 ON Q.id_sign=finalmarks.id_sign AND Q.code_module=finalmarks.code_module AND Q.code_presentation=finalmarks.code_presentation
8 WHERE Q.assessment_type!='Exam')
9 ;

COUNT(*)
-----
21144

```

Figure 42. SQL query showing the number of students who finished a course performing any type of assessment but exams

Further investigating their related records from finalMarks' table led to the mentioned conclusion.

CODE_MODULE	CODE_PRESENTATION	ID_SIGN
DDD	2013J	S11453
DDD	2014B	S17690
DDD	2014B	S8624
DDD	2014J	S9381

Figure 43. Capture from the result of and R anti-join between the results of querying for the students who finished a course performing any type of assessment and those who did any assessments but Exams


```

SQL> SELECT * FROM finalmarks
  2 WHERE code_module='DDD' AND code_presentation='2013J'
  3 AND id_sign='S11453';

ID_SIG COD CODE_ FINAL_RESUL
-----
S11453 DDD 2013J Pass

SQL> SELECT * FROM finalmarks
  2 WHERE code_module='DDD' AND code_presentation='2014B'
  3 AND id_sign='S17690';

ID_SIG COD CODE_ FINAL_RESUL
-----
S17690 DDD 2014B Pass

SQL> SELECT * FROM finalmarks
  2 WHERE code_module='DDD' AND code_presentation='2014B'
  3 AND id_sign='S8624';

ID_SIG COD CODE_ FINAL_RESUL
-----
S8624 DDD 2014B Pass

SQL> SELECT * FROM finalmarks
  2 WHERE code_module='DDD' AND code_presentation='2014J'
  3 AND id_sign='S9381';

ID_SIG COD CODE_ FINAL_RESUL
-----
S9381 DDD 2014J Fail

```

Figure 44. Individual query for each final result of the four students who only did an Exam (assessment type) during the course they were registered to.

```

SQL> SELECT * FROM studentassessment
  2 INNER JOIN assessments
  3 ON assessments.id_assessment = studentassessment.id_assessment
  4 WHERE code_module = 'DDD' AND code_presentation = '2013J'
  5 AND id_sign = 'S11453';

ID_SIG ID_ASSESSMENT IS_BANKED SCORE DATE_SUBM COD CODE_ ID_ASSESSMENT ASSE WEIGHT DATE_ASSE
-----
S11453 25354 0 84 30-MAY-14 DDD 2013J 25354 Exam 100 19-JUN-14

SQL> SELECT * FROM studentassessment
  2 INNER JOIN assessments
  3 ON assessments.id_assessment = studentassessment.id_assessment
  4 WHERE code_module = 'DDD' AND code_presentation = '2014B'
  5 AND id_sign = 'S17690';

ID_SIG ID_ASSESSMENT IS_BANKED SCORE DATE_SUBM COD CODE_ ID_ASSESSMENT ASSE WEIGHT DATE_ASSE
-----
S17690 25361 0 60 25-SEP-14 DDD 2014B 25361 Exam 100 30-SEP-14

SQL> SELECT * FROM studentassessment
  2 INNER JOIN assessments
  3 ON assessments.id_assessment = studentassessment.id_assessment
  4 WHERE code_module = 'DDD' AND code_presentation = '2014B'
  5 AND id_sign = 'S8624';

ID_SIG ID_ASSESSMENT IS_BANKED SCORE DATE_SUBM COD CODE_ ID_ASSESSMENT ASSE WEIGHT DATE_ASSE
-----
S8624 25361 0 82 23-SEP-14 DDD 2014B 25361 Exam 100 30-SEP-14

SQL> SELECT * FROM studentassessment
  2 INNER JOIN assessments
  3 ON assessments.id_assessment = studentassessment.id_assessment
  4 WHERE code_module = 'DDD' AND code_presentation = '2014J'
  5 AND id_sign = 'S9381';

ID_SIG ID_ASSESSMENT IS_BANKED SCORE DATE_SUBM COD CODE_ ID_ASSESSMENT ASSE WEIGHT DATE_ASSE
-----
S9381 25368 0 51 01-JUN-15 DDD 2014J 25368 Exam 100 20-JUN-15

```

Figure 45. Individual query for each assessment score of the four students who only did an Exam (assessment type) during the course they were registered to

It can be clearly seen that there exists no logical correspondence between the score from student S9381 and the final result he or she has assigned:

- Student S11453
 - Exam's score: **84**
 - Final result: **Pass**
- Student S17690
 - Exam's score: **60**
 - Final result: **Pass**
- Student S8624
 - Exam's score: **82**
 - Final result: **Pass**
- Student S9381
 - Exam's score: **51**
 - Final result: **Fail**

Also, there is no clear definition of how a Distinction final result is achieved (there are students with mean scores near 80 points whose final results are labelled as so).

As a result of these observations, it was decided to remove the “weight” attribute from Assessment’s table, as well as its assessments identified as “Exam”. Consequently, any record of an Exam registered in StudentAssessment’s table (4959 from a total of 173852 assessments, a 2.85 %) would also be removed.

Additionally, any record referred to the four aforementioned students who only did an Exam during their correspondent courses would also need to be removed. This includes:

- Their assigned final results from finalMarks’ table.
- Their registration to the correspondent course from studentRegistration’s table.
- Their records of activity along those courses from studentVle’s table.

**Their Exam’s scores referred to the course they were registered to would have already been deleted from Assessments’ table as part of the process detailed prior to this one.*

To describe the effect of the previous processes on the data volume contained in each affected table, the following summary is presented:

- Assessments: went from 206 records to **182**.
 - Additionally, got its “weight” attribute dropped.
- StudentAssessment: went from 173852 records to **168893**.
- FinalMarks: went from 22428 records to **22424**.
- StudentRegistration: went from 32491 records to **32487**.
- StudentVle: went from 8454770 records to **8454354**.

3.5 Resulting Database Model

At this point, as stated earlier in the objectives set for the normalization process previously conducted, our proposed schema satisfies the main efficiency requirements for a database model.

Further normalization tasks can still be performed and, although their dependency on a more in-depth analysis of the schema and its particularities (likelihood of certain queries, performance measures, etc.) doesn't match the intendments of this project, it is planned that the system scales up in these terms.

3.5.1 Logical model

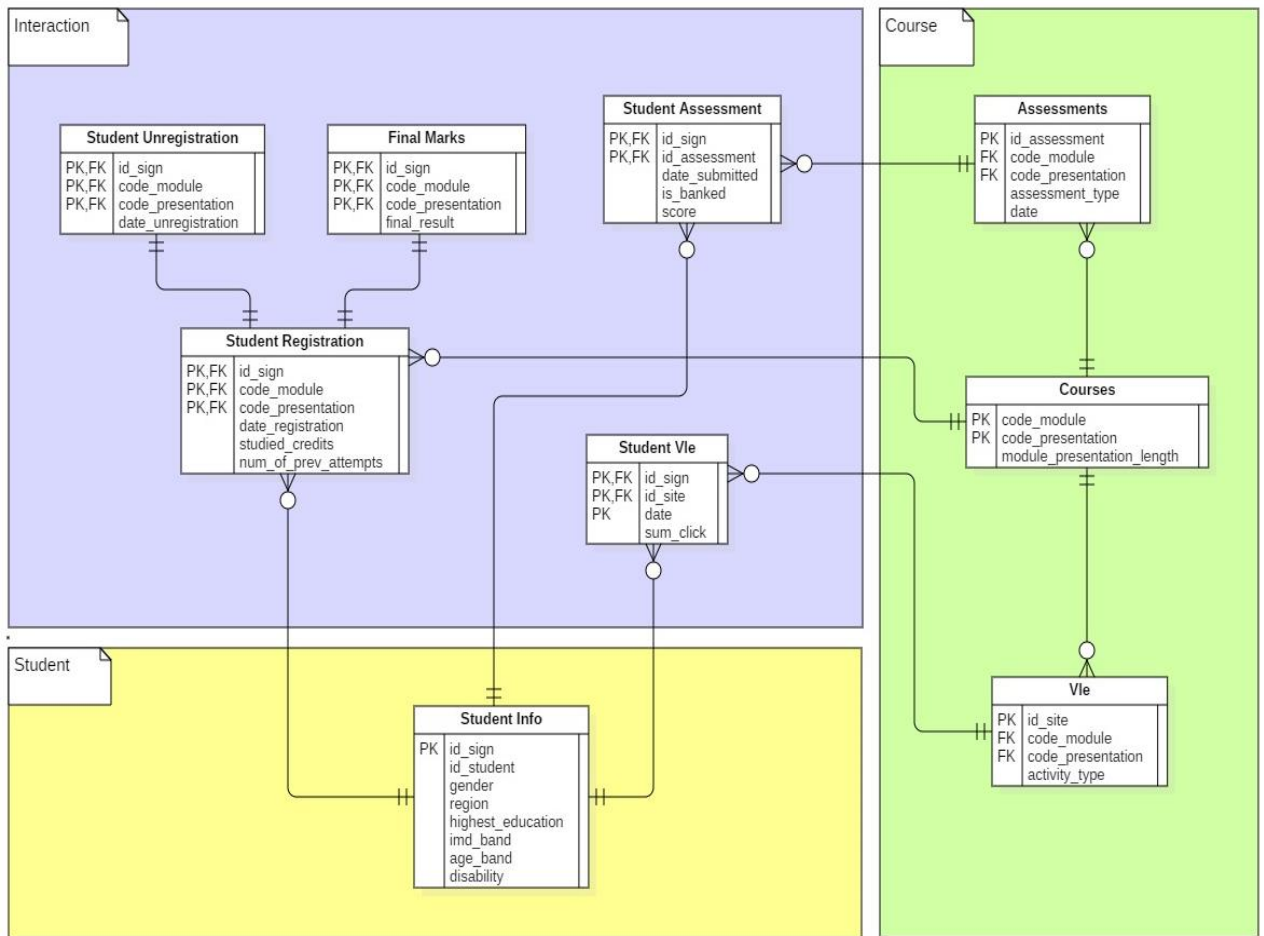


Figure 46. Final database's logical model

3.5.2 Physical model

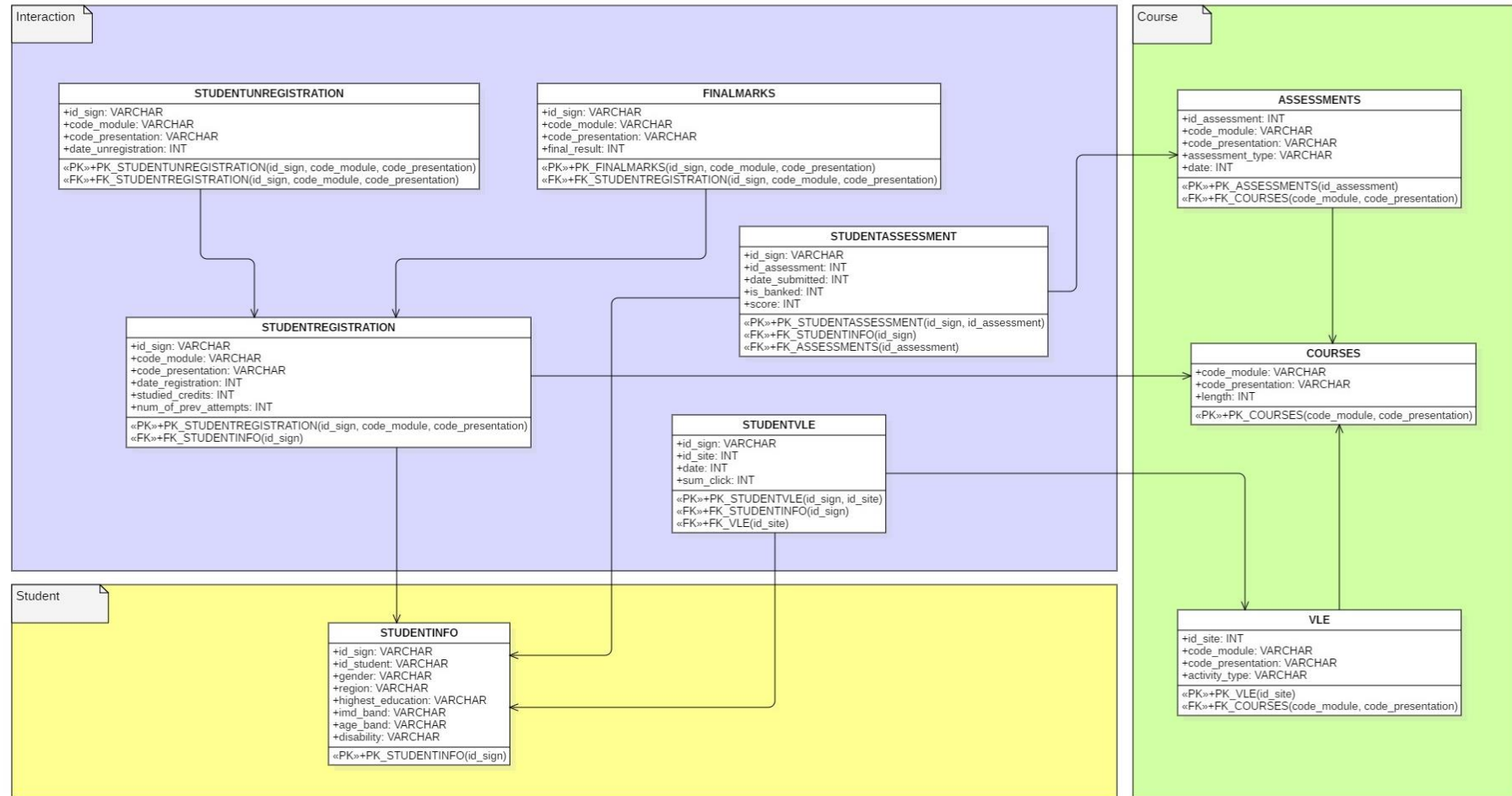


Figure 47. Final database's physical model

4 FEATURE SELECTION AND ENGINEERING

Once available data and its logical arrangement has been discussed along with the model towards which further analytics processes will be aimed, it is necessary to present the set of variables considered for these purposes.

Thus, with the objectives of describing and predicting students' results and interaction with their learning environment (which role as both a predictor and a dependent variable to predict has already been highlighted), the constituent attributes of the initial model under study will be presented using the following template:

Name	Description	Source
<text>	<text>	<Crafted/ -> Table: <text> Column: <text>

Table 13. Features' description template

Column names are self-descriptive, with "Source" varying in accordance to whether an attribute was directly extracted from the database raw content or some processing took place in order to elicit such information. In both cases, origin's table and column's name will be provided, along with the tag "Crafted" if it happens to be of the latter type.

Additional information justifying the presence (or elaboration) of these attributes is provided after the table's presentation.

Name	Type	Description	Source
Gender	Categorical	Indicates whether the student is a Female (F) or a Male (M).	Table: StudentInfo Column: gender
Region	Categorical	Student's region of residence.	Table: StudentInfo Column: region
Highest_education	Categorical	Student's highest level of education achieved.	Table: StudentInfo Column: highest_education
Imd_band	Categorical	Student's council's Index of Multiple Deprivation.	Table: StudentInfo Column: imd_band
Age_band	Categorical	Student's range of age.	Table: StudentInfo Column: age_band
Disability	Categorical	Indicates whether the student has a disability (Y) or not (N).	Table: StudentInfo Column: disability
Previous_attempts	Numerical	Number of times the student has attempted to pass a module.	Table: StudentRegistration Column: num of prev attempts
Studied_credits	Numerical	Number of credits the student is studying at the time of registration.	Table: StudentRegistration Column: studied_credits
Assessments_PerCourse	Numerical	Number of assessments which take place at the course the student is registered in.	<u>Crafted</u> Table: Assessments Column: code_module and code_presentation (unique occurrences)

Resource_variety	Numerical	Number of different resources the student's course offers.	<u>Crafted</u> Table: Vle Column: activity_type (unique occurrences within each course)
CmaTma_rate	Numerical	Number of CMA assessments in the student's course compared to TMA ones.	<u>Crafted</u> Table: Assessments Column: assessment_type (occurrences of CMA with respect to TMA within each course)
Course_EndingPeriod	Categorical	Indicates whether a course has its ending on June or September.	<u>Crafted</u> Table: Courses Column: end_date (month)
Registration_DayRef	Numerical	Difference in number of days between the beginning of the course and the student's registration.	<u>Crafted</u> Table: StudentRegistration Column: date_registration (difference with respect to the beginning of the course)
Transferred_scores	Categorical	Indicates whether a student has some of his/her scores transferred from previous courses (Y) or not (N).	<u>Crafted</u> Table: StudentAssessment Column: is_banked (a unique occurrence for a student makes this flag turn to positive)
MaxDiff	Numerical	Indicates the greatest time span (as number of days) between two continuous VLE interactions for a student in a course.	<u>Crafted</u> Table: StudentVle Column: date_interaction (difference in days between a unique date and the immediate next)

Table 14. Features' description

The first 6 attributes (from “Gender” to “Disability”) refer to student demographics. Their inclusion in the initial model pretends to test the influence these factors may have on both interaction with the platform and final results. This would help design further analytics with respect to certain subsets of students and understand how these characteristics may affect learning.

“Previous_attempts”, “Studied_credits”, “Registration_dayRef”, “MaxDiff” and “Transferred_scores” represent the intersecting characteristics between students and the course they are enrolled in. As so, they are intended to showcase the impact the initial scenario in which students tackle their registration to a course may have on results and interaction with the platform (e.g. how excessive workload from other courses may impair overall learning, success expectations when registering on an already started course, etc).

It is needed to point out Disability and Transferred_scores' low rate of positive occurrences (only 522 have scores transferred and 1919 have an acknowledged disability,

both out of 23404 distinct students), which minimizes its chance of making a real impact on the general model’s fit. Coherently with this, and as it refers to a specific subset of students (e.g. one will only have scores transferred from a course which failed to pass, thus specifically referring to students who are re-trying to pass a failed course), its potential influence should also be studied for those specific cases.

Finally, “Assessments_PerCourse”, “Resource_variety”, “CmaTma_rate” and “Course_EndingPeriod” describe relevant characteristics of the courses the students have registered to. Their presence addresses the need for assessing how learning takes place under each specific set of circumstances, with would aid the development of further courses as a response to student’s needs (e.g. workload management, calendar design, etc).

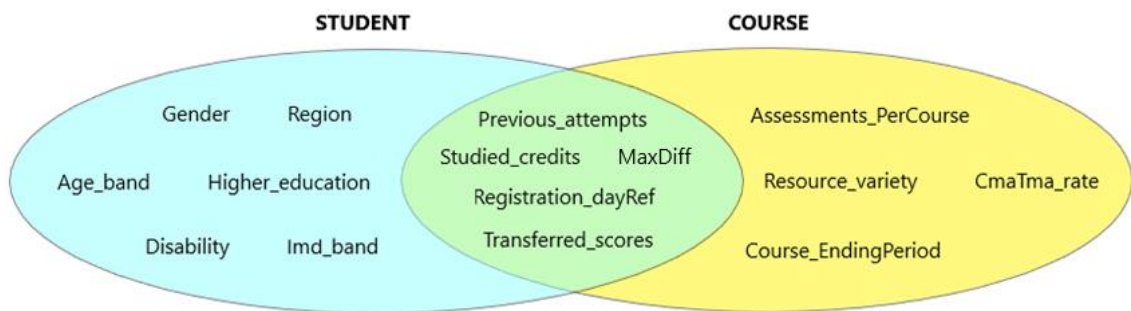


Figure 48. Initial attributes: domain visualization

The above figure represents the extent to which we have been able to depict the elements involved in learning for our particular case. However, as it will be seen in the following sections, it is subject to change, especially in which respects to its refinement, aimed towards the improvement of the model’s predictive capabilities.

Additionally, it is important to acknowledge the limitations of the data collection with which we pretend to conduct our analytics processes, specially the lack of qualitative information from students and evaluators, which will surely complement our model’s information and improve its performance in both assessment and predictive tasks ([32]).

5 SETTING AN APPROACH

As a first step into the development of further data-driven processes, it is fundamental to define the approach from which our application's analytics and predictive tasks will be developed.

For this purpose, it will be important to take current state of art into account and review the potentiality of adopting a model based on today's widespread methodologies, which aim to define student's engagement through the measurement of purely performative (and more specifically, quantitative) attributes.

Adapting this take on performance to the possibilities provided by the information present on our dataset, interaction with the platform's content (VLE) would be interpreted as a measure of engagement with the learning process, thus having a main role in explaining student's results.



Figure 49. Diagram illustrating the traditional approach to student engagement

The above figure intends to illustrate the understanding of data extracted from interaction with content (generally clickstream data, as in our own dataset) as a cornerstone for making inferences about learner behaviours, frequently held by current research and practice in the field of Learning Analytics ([32]).

5.1 Assessing our case

In which respects to our particular case, we have assessed the relationship between the mean daily interaction of each student along the correspondent course's duration and the labelling of their course completion as "Fail" or "Success" (includes both students who simply passed and those who did it with distinction).

The information needed for the elaboration of a measure for the mean interaction of a student along a course was extracted from StudentVle's table, which "sum_click" attribute contains the track of the clickstream data for each day an interaction with the Virtual Learning Environment (VLE) took place.

It is important to remark that students who withdrew from a course were not taken into account in this particular analysis as that group is understood to be disjoint from the scope of this case (which focuses on results from completed courses).

Data from 22428 students was involved in the study of this case.

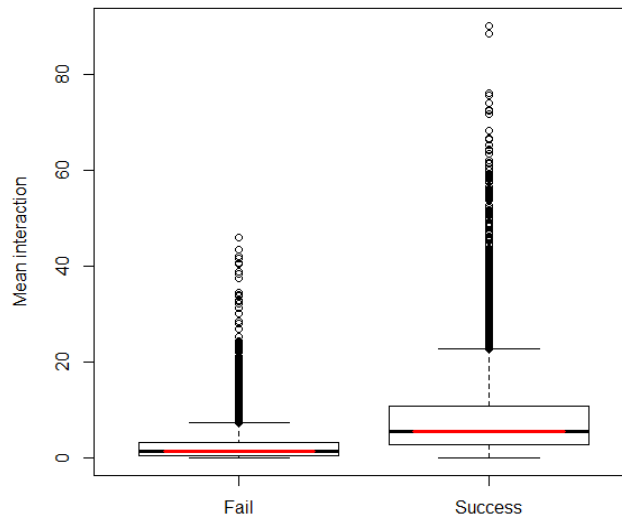


Figure 50. Boxplot for mean interaction and final result labelling

The boxplot presented allows to catch a glimpse of a clear behavioural distinction between students who failed and those who succeeded, with the latter group presenting a higher mean of interaction.

The red line plotted represents the median of the distribution for both groups, which matches their means. This, in conjunction to the consideration of the Central Limit Theorem ([33]) and the substantial size of our samples, supports the assumption of normality for both distributions, which will aid further observations.

Attention was also paid to outliers, which detection and treatment consisted on the removal of those cases lying out of the Interquartile Range (IQR) from each group. Apart from the distorting effects on certain statistics (e.g. variance, mean...), it was decided to remove them due to the nature of the domain being treated, in which cases such as these need to be assessed separately (e.g. why is a student with a high interaction rate failing the course?), making the worthiness of a modification process to fit them in the distribution questionable.

However, results of will be presented for both cases, with and without outliers.

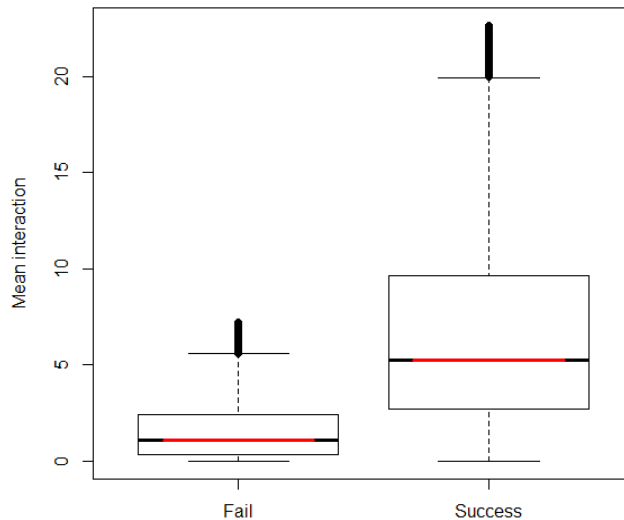


Figure 51. Resulting boxplot for mean interaction and final result labelling after outlier removal

In this case, the difference in mean and range of values covered by each group becomes clearer.

To test the validity of this assumption, a simple methodology has been designed:

1. Assessment of the results from a Welch's t-test to justify further investigation of the case.

The criticality and controversy that has recently emerged with respect to the use of p-values and its implications (specially in terms of proper methodologies and replicability of studies ([34]) makes it necessary to clarify the use given in the scope of this evaluation:

Recalling Fisher's purposes while designing this tool ([35]) and given the understanding of the null hypothesis as an expectation of a case towards which to direct efforts or not, a high p-value for the event under study would be interpreted as the existence of enough information to consider the development of further experiments in that direction as potentially fruitful.

Consequently, with the null hypothesis for our particular case being the equality between each group's mean, a low p-value for the Welch's test would aid us to support the decision of further investigating the difference between the means of "Fail" and "Success" groups (alternative hypothesis).

Additionally, it is important to mention the choice of this test as an alternative to ANOVA's (Analysis of Variance) due to its assumption of homoscedasticity for the variables undergoing the test, which wasn't satisfied by our data (as revealed by the Bartlett's test ([36]) conducted, which revealed a significant difference between each group variance). Welch's t-test aims its usage towards data groups with unequal variances and sample sizes, as they are in our case.

2. Evaluation of Cohen's D (standardized difference between two means) as a measure of the effect size of mean interaction values with respect to final results.

It is necessary to remark the necessity for normative references from which to extract the standardizer to use when calculating this statistic ([37]), in order to avoid a potential bias towards the overestimation of its result.

As for this case we do not have a formal reference for the mean interaction of students with the particular environment (VLE) referred to the dataset, we will rely on the “unbiased version” of this measure, known as Hedge’s D, which corrects its calculation to avoid undesirable biases.

In summary, this measure will give us an estimation of the difference in the mean interaction between “Success” and “Fail” groups.

The results obtained from these processes are summarised and discussed hereafter:

Data	Process	Result
With outliers	Welch’s p-value	< 2e-16
	Cohen’s D (Hedge’s correction)	1.017497
Without outliers	Welch’s p-value	< 2e-16
	Cohen’s D (Hedge’s correction)	1.60421

Table 15. Welch’s t-test and Cohen’s D results (influence of interaction in final results)

As previously advanced, the significantly low Welch’s p-values help aim our assumptions and analysis towards the previously advanced difference in the mean interaction between groups. Additionally, Cohen’s D values reveal a considerable significance for this difference (Cohen’s guidelines ([38]) refer to values of 0.8 as relative to large effects, and educational literature ([39]) identifies ranges spanning from -0.5 to 1.75 for this particular domain), of particular strength for the case not considering outliers.

For better understanding the meaning and implications of Cohen’s D as a measure of the effect size of a particular feature, the following graphs are proposed (with the acknowledgement that, despite assuming normality for both “Fail” and “Success” distributions, their different variances are not portrayed in these charts, thus being necessary to remark their illustrative purposes).

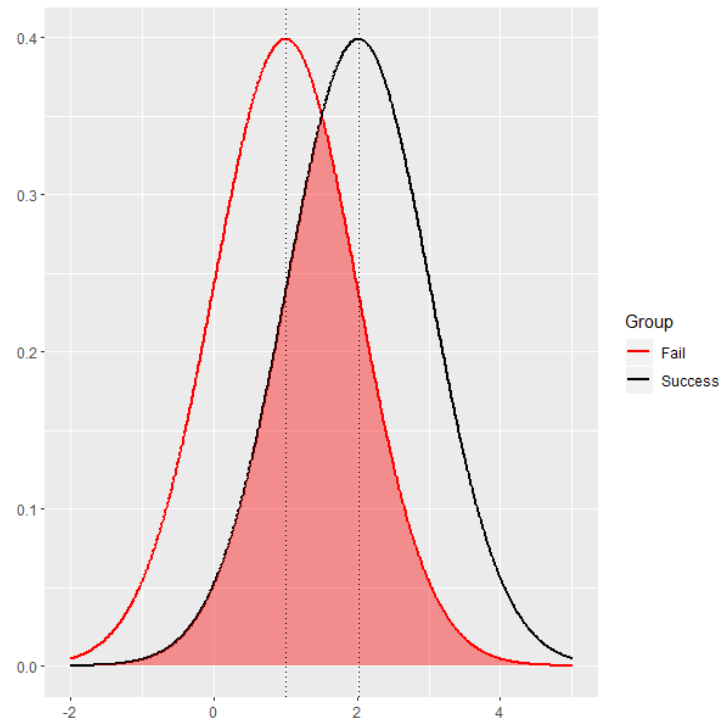


Figure 52. Visualization of Cohen's D meaning (value of 1.017497)

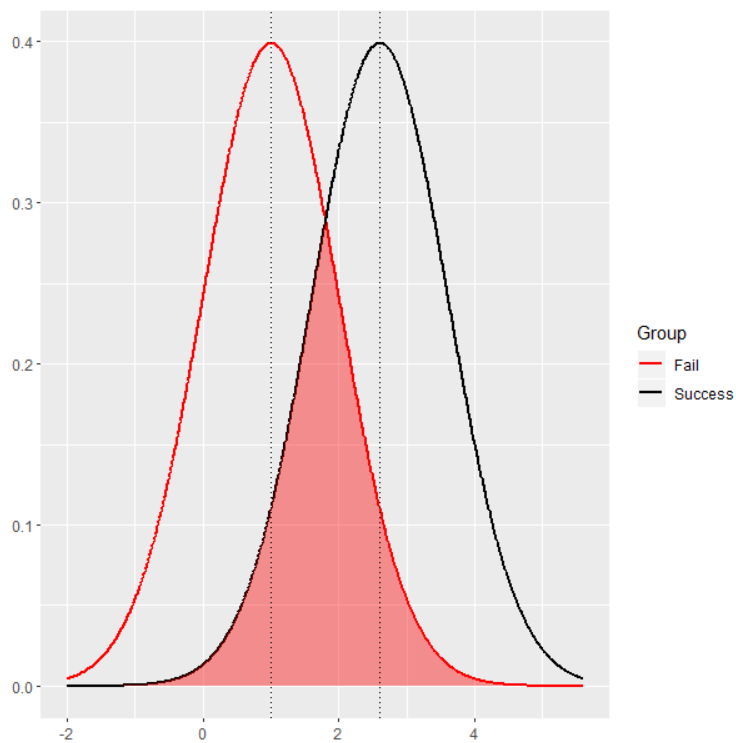


Figure 53. Visualization of Cohen's D meaning (value of 1.60421)

It can be observed that, as Cohen's D increases, so does the "separation" between distributions, with their overlapping region consequently dwarfing (they become independent, and thus better described by the measured variable). For our particular case, it may be interpreted as a measure of the impact of mean interaction in final results.

However, it is important to address its interpretation cautiously, taking into account the nature and purpose of the statistic (measure the effect of a particular feature of each group: “Fail” and “Success”; on a certain parameter: “mean interaction”) and its domain of application.

Consequently, and considering the fact that the group identification for this case comes after the occurrence of the measured event (interaction), which discards any causality inference possible, a valid interpretation would be that having failed or passed a course may explain certain patterns of interaction.

Nonetheless, the already mentioned overlapping region should not be dismissed, especially when aiming to construct an accurate predictive model for the results (intuitively said, this area should be minimal to be able to perform a classification with enough precision). In this case, although mean interaction may aid this task, it appears to not be enough to be considered as an appropriate nor complete model.

As complementary to this, the Intraclass Correlation coefficient (ICC, how well values from each group describe them) was computed with the following results:

Data	ICC
With outliers	0.2
Without outliers	0.4

Table 16. Intra-Class correlation results (final results described by mean interaction)

Despite a considerable increase can be observed for the dataset not considering outliers, descriptiveness of each group, although existent, is scarce (matching the assessment made with respect to Cohen’s D). This reinforces the hypothesis that the conception of interaction as central for estimating results is biased and should be re-addressed for completeness.

5.2 Reinforcing statements

Further reviewing this case, with the purpose of broadening the scope of the premises with respect to the approach given to our work, a complementary analysis inspecting the relationship of mean interaction with mean scores along the course was conducted.

It is needed to note that mean scores were computed not taking into account assigned weights to each assessment scored, since we are trying to measure the impact of interaction on pure performance along the course, sense which may be distorted by weights (considered as a subjective or external factor penalizing certain scores).

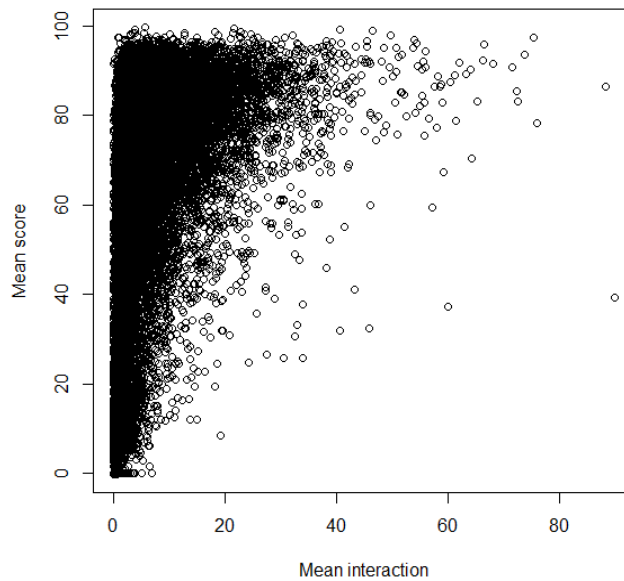


Figure 54. Visualization of mean interaction against mean score values

At first sight, it may appear that no pattern exists in which respects to this relationship. However, as did in the previous case, it is important to consider the impact of outliers on the distribution and its statistics. Its treatment consists of the same reasoning and procedure as before (IQR for detection and removal of occurrences).

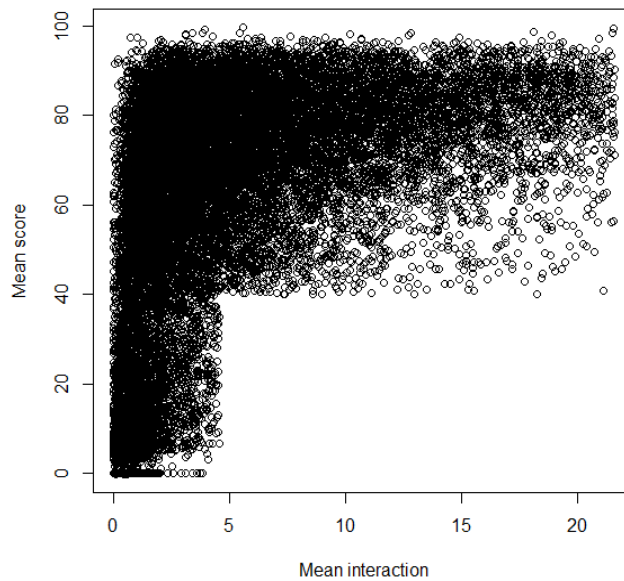


Figure 55. Visualization of mean interaction against mean score values (outliers removed)

After removing outliers, a clearer arrangement of the data appears at sight. It is remarkable how, despite comprising practically the complete range of scores in the mean interaction's interval $[0,5]$, no values for mean score below 40 (which, according to the dataset's documentation is considered as passed) appear for mean interactions superior to 5.

For tackling the possibilities of modelling mean scores obtained with respect to mean interaction, a simplistic approach consisting on assessing both a traditional linear model and a logarithmic model (at sight of the distribution's shape) has been conducted.

Although graphs for visualizing the models for the data containing outliers are not present, results are equally shown.

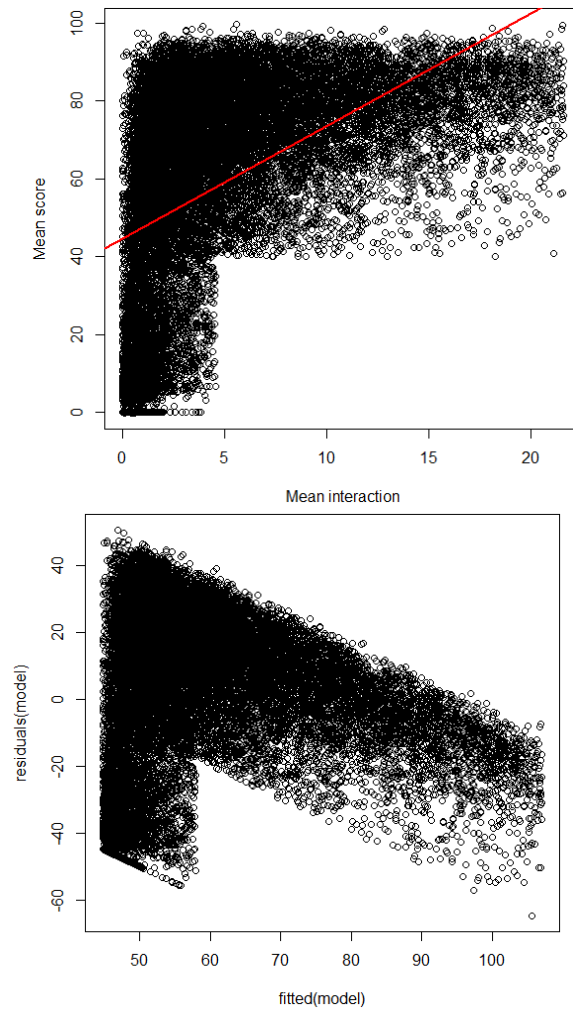


Figure 56. Mean interaction against mean score values: fitness of a linear model (left) and its correspondent residuals (right)

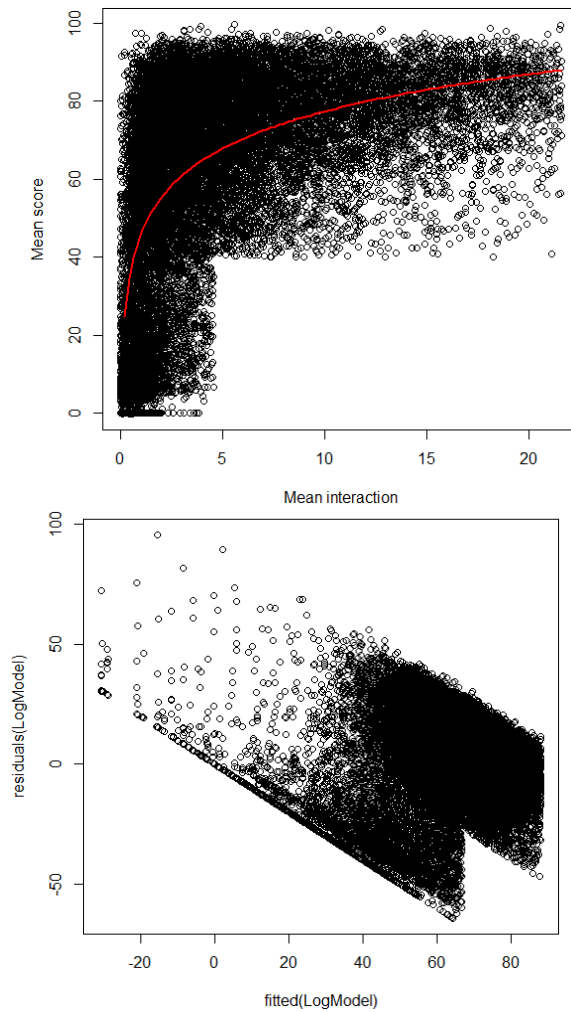


Figure 57. Mean interaction against mean score values: fitness of a logarithmic model (left) and its correspondent residuals (right)

Data	Model	Correlation coefficient	R-squared
With outliers	Linear	0.43	0.18
	Logarithmic	0.65	0.42
Without outliers	Linear	0.51	0.26
	Logarithmic	0.67	0.45

Table 17. Appropriateness of linear and logarithmic models to describe final results based on mean interaction

Although a noticeable improvement in the correlation coefficient for the logarithmic model can be observed, its joint interpretation with the correspondent R-squared value (proportion of the mean score's variance that is predictable from mean interaction's data) elicits a low significance for this model (and any other), thus not being able to consider it as appropriate, neither for our purposes nor from a generic point of view.

A visual interpretation of these inferences can be made from the observation of the residuals' plots, which can be understood as a depiction of R-squared. Ideally, for a model considered to fit a distribution and be suitable for prediction, this plot shouldn't show a patterned arrangement such as those of the ones presented.

Summing up our findings, it can be concluded that a descriptive and/or predictive model for results solely based on the tracked interaction of students with the content offered for the completion of the course (a purely performative measure) is scarce in which respects to addressing a much broader picture (point shared by most literature focusing on this point ([5]-[6]-[10]-[11]-[32])).

Acknowledging the potential limitations of the database we are working with in which respects to the completeness of its scope, and coherently with the previous statements, efforts will be directed towards a multivariate approach taking into account multiple possibilities for describing both interaction and outcomes.

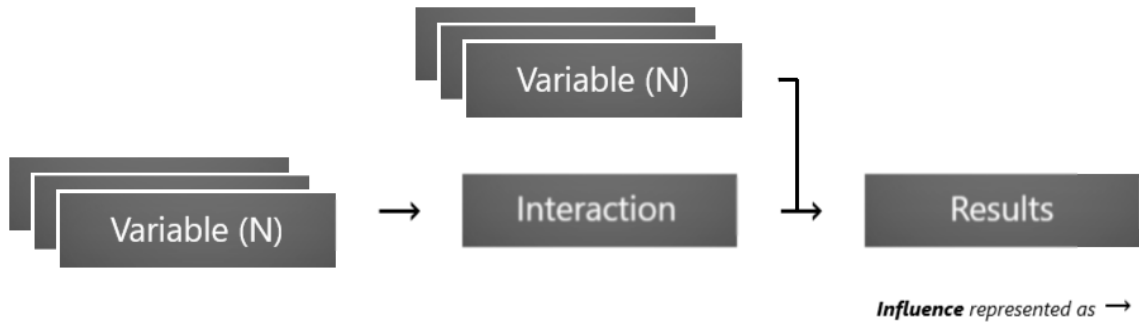


Figure 58. Diagram illustrating the approach decided for our project analytics tasks

Remarkably, the mean interaction variable developed for the previously detailed processes, will act as both a variable to forecast over time and a predictor for students' results. This case will be showcased and discussed in the following sections, when treating its related tasks.

6 REGRESSION MODELS

Once defined the set of features considered for the development of our conceptual cornerstones' (interaction and results) models, the process of its confection and refinement towards an optimal performance is discussed in this section.

Firstly, it is necessary to point out the varying nature of some of the previously exposed features over time, namely:

- Mean interaction
- Average score
- MaxDiff

This “varying nature” concept refers to the fact that their assigned values change in accordance to the time at which its being measured, and its inclusion on a model is subject to the characteristics of the scenario being analysed (e.g. predictions with respect to data from an ongoing or a finished course).

It can be said that the workflow follows a top-down methodology for the development of our analytics tasks. As so, general models involving all the information available (i.e. from the beginning to the end of every course) are set and assessed first so they serve as a reference for the construction of predictive models referred to the prognosis of certain features from an early date.

Following, a general framework for the development of each model is presented:

- Random separation of the data:
 - **Training set** (60 %) with the unique purpose of training the model under study.
 - **Validation set** (20 %) to ensure the avoidance of any bias that may arise during processes referred to data pre-processing or parameter tuning/setting of the algorithm being used (if needed).
 - **Test set** (20 %) with respect to which predictive performance of the model is assessed.
- Assessment and treatment of categorical data to better fit the capabilities of those algorithms which may be sensitive to this type of variable. Two independent processes are conducted:
 - **Clustering of categories** via agglomerative clustering: aims to reduce the cardinality of this variables to aid data distinction by clustering them in accordance to a specific measure of similarity referred to the dependent variable (e.g. rate of positive occurrences within each group if the variable to predict is categorical).
 - **Transformation to numerical** variables using weight of evidence as measure ([40]). Given its formula:

$$WOE_i = \ln\left(\frac{P_i/TP}{N_i/TN}\right)$$

i → index of each group's distinct factor
P(i) → number of positive occurrences referred to the group's *i*th factor
N(i) → number of negative occurrences referred to the group's *i*th factor
TP → total number of positive occurrences
TN → total number of negative occurrences

Figure 59. Weight of Evidence's formula

The main task involved in this process is to identify a label with which differentiate positive occurrences from negative ones. Consequently, for this conversion to result in a predictor potentially describing the dependent variable, this labelling factor should vary according it (the variable to predict; e.g. defining a threshold over which its referred values are considered to be positive, and negative if they are below, if the dependent variable is numerical).

- Selection of an appropriate set of attributes for the model, taking the following processes into account:
 - Measures of **attributes' value** (effect on the predicted variable) by applying a criterion based on:
 - Information gain as indicative of each attribute's potential to describe the dependent variable ([41]).
 - Information gain ratio to compensate any possible bias arising from the overestimation of variables with high cardinality by the information gain estimation ([41]).

For both cases, attributes measured to not have a significant impact on the predicted variable (criteria for this distinction will be specified) will be removed.

- Treatment and assessment of **collinearity**: with this purpose, and attending to the distinct types of our defined set of features, this process will be split in three parts, each one attending to:
 - Numerical variables, using its correlation matrix as an indicative of whether or not collinearity exists (≥ 0.75).
 - Categorical variables, for which another matrix, this time computing Cramer's V, is generated. Values ≥ 0.75 are considered to be indicative of collinearity ([42]).
 - Categorical vs. numerical variables, for which each pair of the type categorical-numerical the intra-class correlation coefficient has been calculated. Values ≥ 0.75 are considered to be indicative of collinearity ([43]).

Those cases marked as collinear by the specified criteria will have their least relevant feature (according to its measured importance) removed.

The effect of each of these processes will be tested both separately and in conjunction for every algorithm defined for the model and task (classification, regression, etc) under study, which results in the following list of cases to study:

Cases of study		
Raw variable set	Category clustering	Numerical transformation of categories
Collinearity treatment	Information gain treatment	Information gain ratio treatment
Collinearity + information gain treatment	Collinearity + information gain ratio treatment	Collinearity + Category clustering
Collinearity + Numerical transformation of categories	Information gain treatment + Category clustering	Information gain treatment + Numerical transformation of categories
Information gain ratio treatment + Category clustering	Information gain ratio treatment + Numerical transformation of categories	Collinearity + Information gain treatment + Category clustering
Collinearity + Information gain treatment + Numerical transformation of categories	Collinearity + Information gain ratio treatment + Category clustering	Collinearity + Information gain ratio treatment + Numerical transformation of categories

Table 18. Regression tasks' cases of study

It is important to remark the conception of information gain and information gain ratio as distinct criteria to evaluate features' importance, thus not being assessed jointly.

Additionally, this leads to the possibility of the collinearity assessment to define two distinct sets of features for its removal: one attending to information gain's criteria and the other based on information gain ratio.

As a conclusion, coherently with the previous explanation, the list of study cases presented is subject to change of its cardinality, although, if occurring, this increase will be due to an already defined case including more than one set of features, and not because of the appearance of a new study case.

6.1 Mean Interaction (prediction task)

The first prediction model was built for mean interaction, the main feature with which we count to prognose student performance.

Predictors			
avgScore	maxDiff	rate cmaTma	resourceVariety
numberAssessments	gender	studied credits	highest education
imd band	num of prev attempts	bankedFlag	age band
region	regFrom iniDate	disability	course endPeriod

Table 19. Predictors employed for mean interaction's regression tasks

All processes following detailed have been conducted on the validation set.

6.1.1 Treatment of categorical data

Two different datasets were produced as the result of this procedure, one from the transformation of the selected variables to numerical and other from the clustering of these variables' factors.

- Selected variables

As specified in table 14, our defined dataset counts with categorical variables, from which factor-cardinality is detailed hereafter:

Predictor	Number of factors
gender	2
highest education	5
imd band	10
bankedFlag	2
age band	3
region	13
course endPeriod	2
disability	2

Table 20. Categorical predictors and their cardinality (mean interaction's regression tasks)

Given the purpose to reduce categorical variables' cardinality to (potentially) improve the performance of the selected algorithms on the dataset, and observing the previous table's description, it can be elicited that only 4 out of the total 8 categorical variables may be subject to this process (no reduction nor transformation would be of significant relevance for dichotomous variables):

- Highest education
- IMD band
- Age band
- Region
- Transformation to numerical

As previously detailed, the numerical-conversion process was performed attending to the weight of evidence referred to each occurrence of a category:

$$WOE_i = \ln\left(\frac{P_i/TP}{N_i/TN}\right)$$

i → index of each group's distinct factor
P(i) → number of positive occurrences referred to the group's *ith* factor
N(i) → number of negative occurrences referred to the group's *ith* factor
TP → total number of positive occurrences
TN → total number of negative occurrences

Figure 60. Weight of Evidence's expression

For each of the variables to transform, positive and negative occurrences were distinguished taking into account the labelling of a student's final result as "Success" or "Fail", respectively.

Thus, the variables involved in the calculation of this measure would be explained as follows:

- TP: total number of "Success" occurrences
- TN: total number of "Fail" occurrences
- P(i): number of "Success" occurrences referred to the group's *ith* factor
- N(i): number of "Fail" occurrences referred to the group's *ith* factor

With the structure of the procedure clearly defined, the calculation of the new variables is practically straight-forward.

REGION	REGION_WOE
East Anglian Region	0.016642507
West Midlands Region	-0.156222309
South West Region	0.210240642
East Midlands Region	0.166579705
South West Region	0.016642507
London Region	0.471535959
North Western Region	0.166579705
East Anglian Region	0.166579705
East Anglian Region	0.166579705
South East Region	0.471535959
North Western Region	0.016642507
South West Region	0.016642507
East Anglian Region	0.292972039
Yorkshire Region	-0.141463290

Figure 61. Sample of the transformation "Region" variable from categorical to numerical

As a result, a new dataset is generated with these categorical variables converted to numerical, consequently being part of the case studies to assess.

- Clustering of factors

One of the main questions to answer before initiating a grouping process refers to the number of clusters to define and, more fundamentally, if clustering is in fact needed or worth being performed.

To respond to these issues, and after defining the criteria of similarity between groups as the success rate of each group (difference in the rate of success-labelled students with respect to the total for that group), the average silhouette method ([44]) was considered.

The silhouette statistic measures the appropriateness of a clustering process by jointly assessing cluster cohesion (similarity of instances within a cluster) and cluster separation (distance between clusters).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$a(i)$ → average distance between the point i and the other points in the cluster
 $b(i)$ → average distance between the point i and all points in the nearest cluster
 $s(i)$ → silhouette of point i

Figure 62. Formula for the silhouette of a cluster's point

In summary, the premise behind its usage states that the optimal number of clusters “k” is the one that maximizes the average silhouette over a range of possible values for “k” ([45]).

Following, results for each of the variables considered are shown:

Age band:

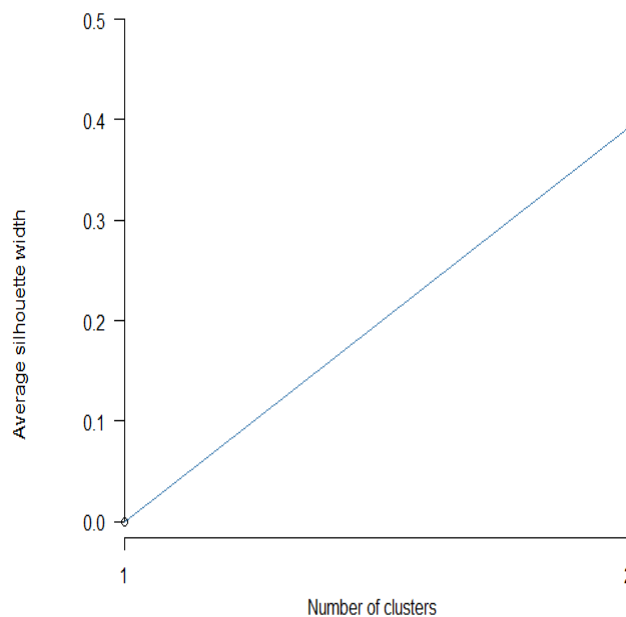


Figure 63. Average silhouette's width for Age's band categories' clustering

As a 3-factor variable, little margin for its optimization through clustering of its categories was available. In fact, results show that little cohesion between groups (average silhouette < 0.5) would be derived from this process.

Consequently, this variable is discarded as re-categorizable (although it will be included in the resulting model).

Highest education:

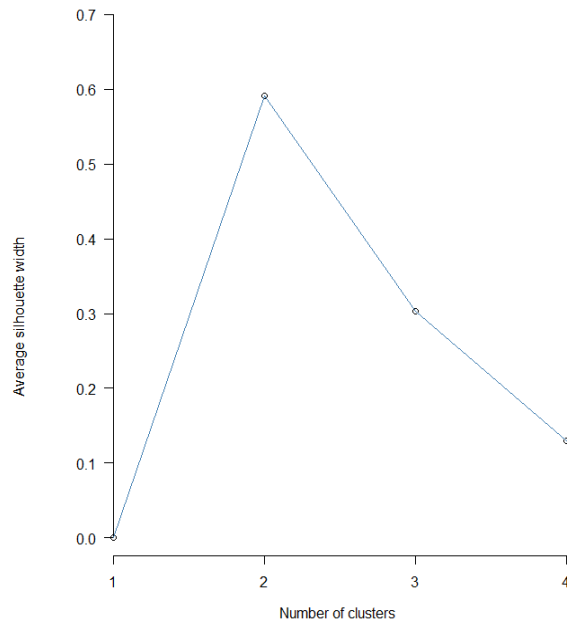


Figure 64. Average silhouette's width for Highest education categories' clustering

For this variable, only the case for the re-arrangement of its factors along 2 clusters produced significant outcomes (average silhouette ≈ 0.7).

A detailed observation of this scenario is presented:

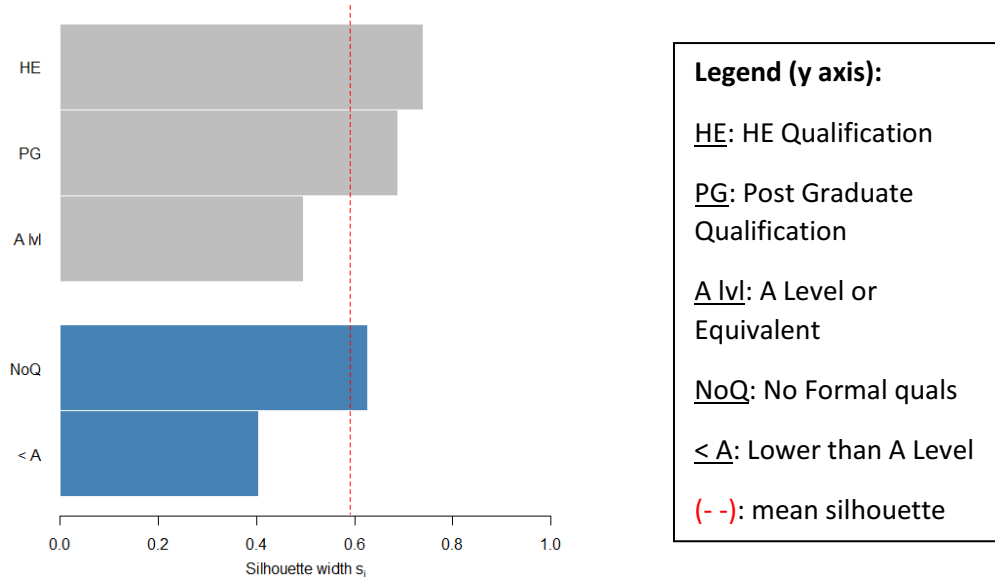


Figure 65. Clustering of Highest education level

It can be observed how the output for this process distinguishes two groups with an implicit educational rank according to its level which, from lowest to highest, would be:

1. No Formal quals
2. Lower than A level
3. A level or equivalent
4. HE Qualification
5. Post Graduate Qualification

HIGHEST_EDUCATION	successRate
No Formal quals	0.4285714
Lower Than A Level	0.5853199
A Level or Equivalent	0.7431034
HE Qualification	0.7791096
Post Graduate Qualification	0.8125000

Figure 66. Explicit clustering of Highest education level

This difference becomes explicitly clear attending to the success rates of each group, which elicits a direct relationship between the previously defined rank and the success rate of its students.

In addition to the definition of the new arrangement for Highest education's categories, a relationship worth studying in further analysis has been identified.

Region:

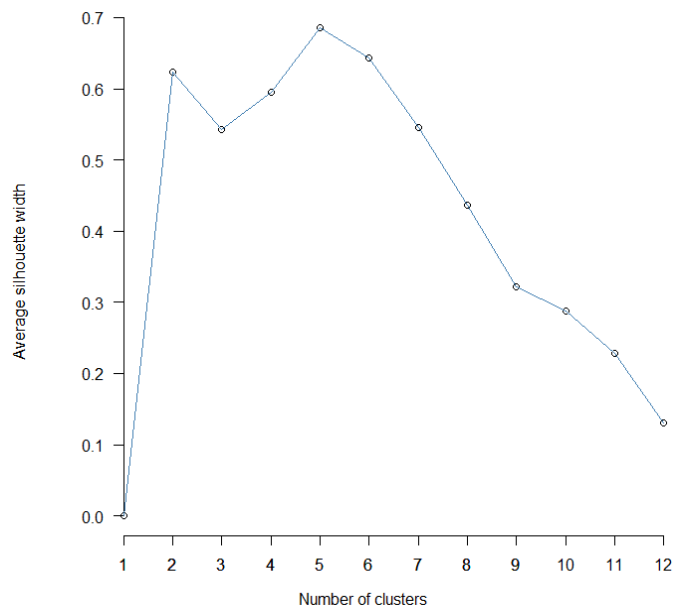


Figure 67. Average silhouette's width for Region categories' clustering

A wide variety of possibilities for grouping this variable's factors are observed, ranging from 2 to 7 clusters, from which average silhouette's width starts to decrease to non-significant levels (< 0.5).

Legend (y axis):

<u>EMR</u> : East Midlands Region	<u>SER</u> : South East Region	<u>SWR</u> : South West Region
<u>SR</u> : South Region	<u>EAR</u> : East Anglian Region	<u>YR</u> : Yorkshire Region
<u>LR</u> : London Region	<u>NR</u> : North Region	<u>NWR</u> : North Western Region
<u>SC</u> : Scotland	<u>W</u> : Wales	<u>WMR</u> : West Midlands Region
<u>IR</u> : Ireland		

(-): mean silhouette

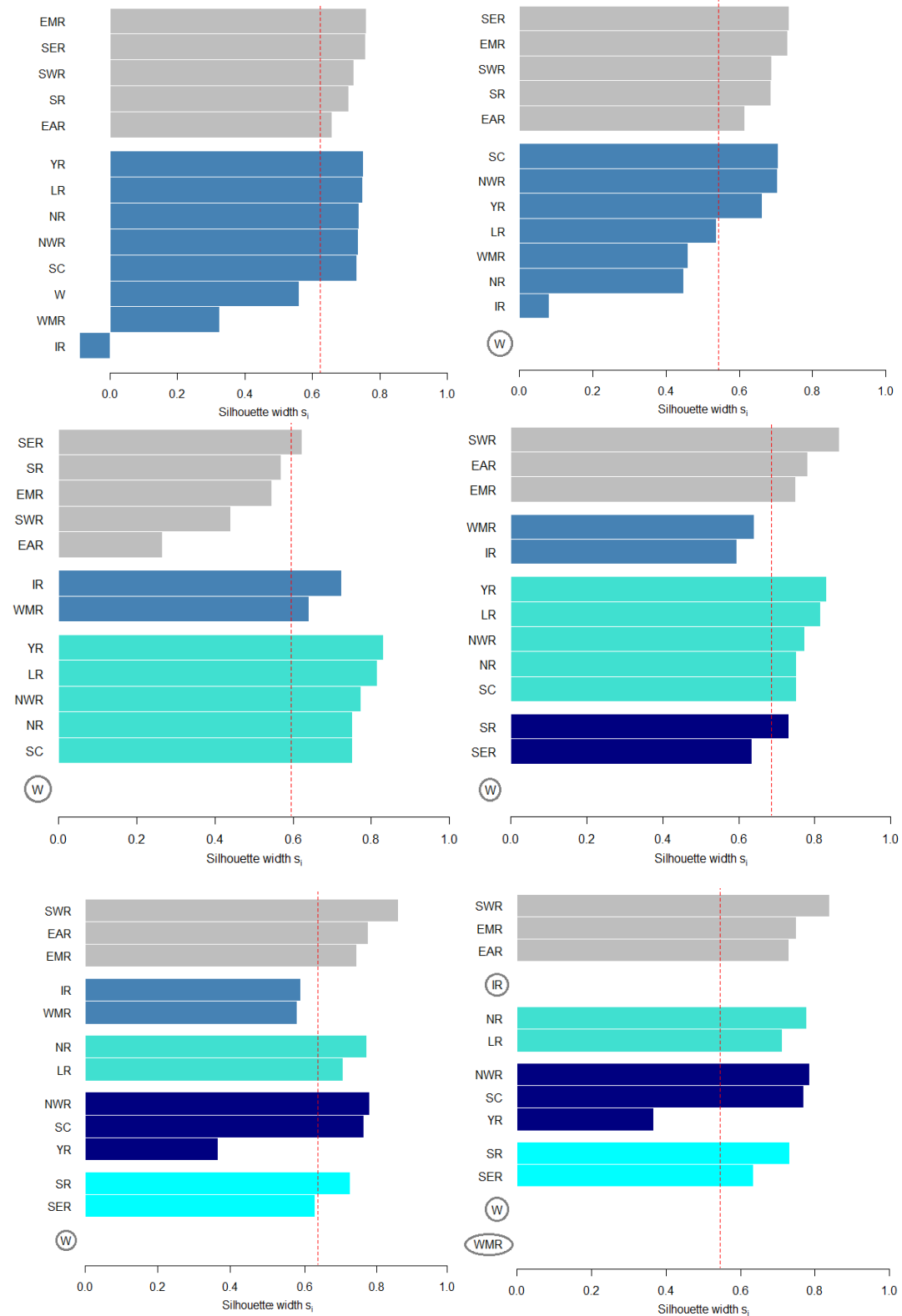


Figure 68. Clustering of Region

During the assessment of the illustrated cases, it is important to notice the evolution of each factor's silhouette along each of them to generate an appropriate basis for our decision.

The IR's negative value (possibly indicative of low cohesion to its assigned cluster) for the first case (2 clusters) is "fixed" in the second (3 clusters) by isolating W, and at expense of a decrease of the average silhouette value. Although these two cases are undesirable, the observation of their differences aids the comprehension of the entire scenario.

Following, the third case (4 clusters) generates a new cluster for IR and WNR, which increases the average silhouette of the whole, but significantly impairs that of the grey group. This is solved in the fourth (5 clusters) case by separating SR and SER which, apart from increasing both average and grey's silhouette, maintains the values for the other clusters almost invariable with respect to the previous case.

Both last two cases (6 and 7 clusters) have worse average silhouette values (with cluster-individual silhouette of previous groups remaining almost invariable), in coherence with the observed decrease in figure 67. The 7-cluster case, specifically, displays three different clusters corresponding to only one factor. This, apart from differing from this process objectives, doesn't even imply a significant improvement of cluster-individual silhouette values, which leads to discard these two cases.

As a coherent conclusion with the evaluation performed, this variable's categories are to be re-assembled according to the 5-cluster case.

REGION	successRate
Wales	0.5767918
North Western Region	0.6495957
London Region	0.6525000
Scotland	0.6604555
Yorkshire Region	0.6639676
North Region	0.6747967
Ireland	0.6909091
West Midlands Region	0.6953642
East Anglian Region	0.6955556
East Midlands Region	0.7106109
South West Region	0.7108434
South Region	0.7707809
South East Region	0.7786260

Figure 69. Explicit clustering of Region

Further study of the differences in success rates between regions may be valuable.

IMD_band:

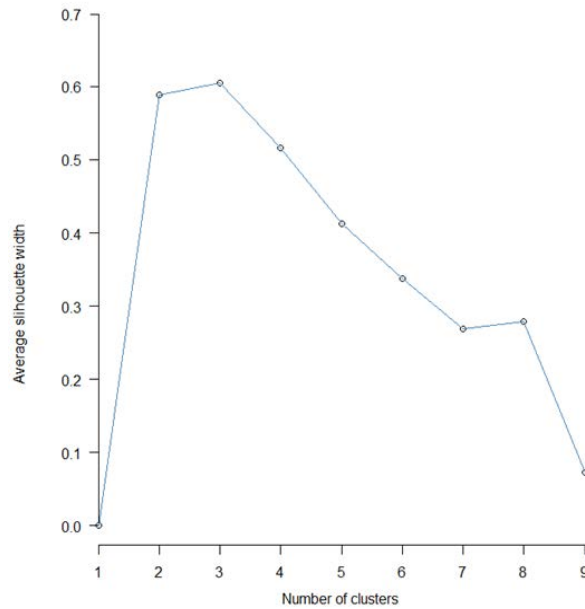


Figure 70. Average silhouette's width for IMD's band categories' clustering

The three cases for which a significant average silhouette value (> 0.5) has been observed are those referred to 2, 3 and 4 clusters.

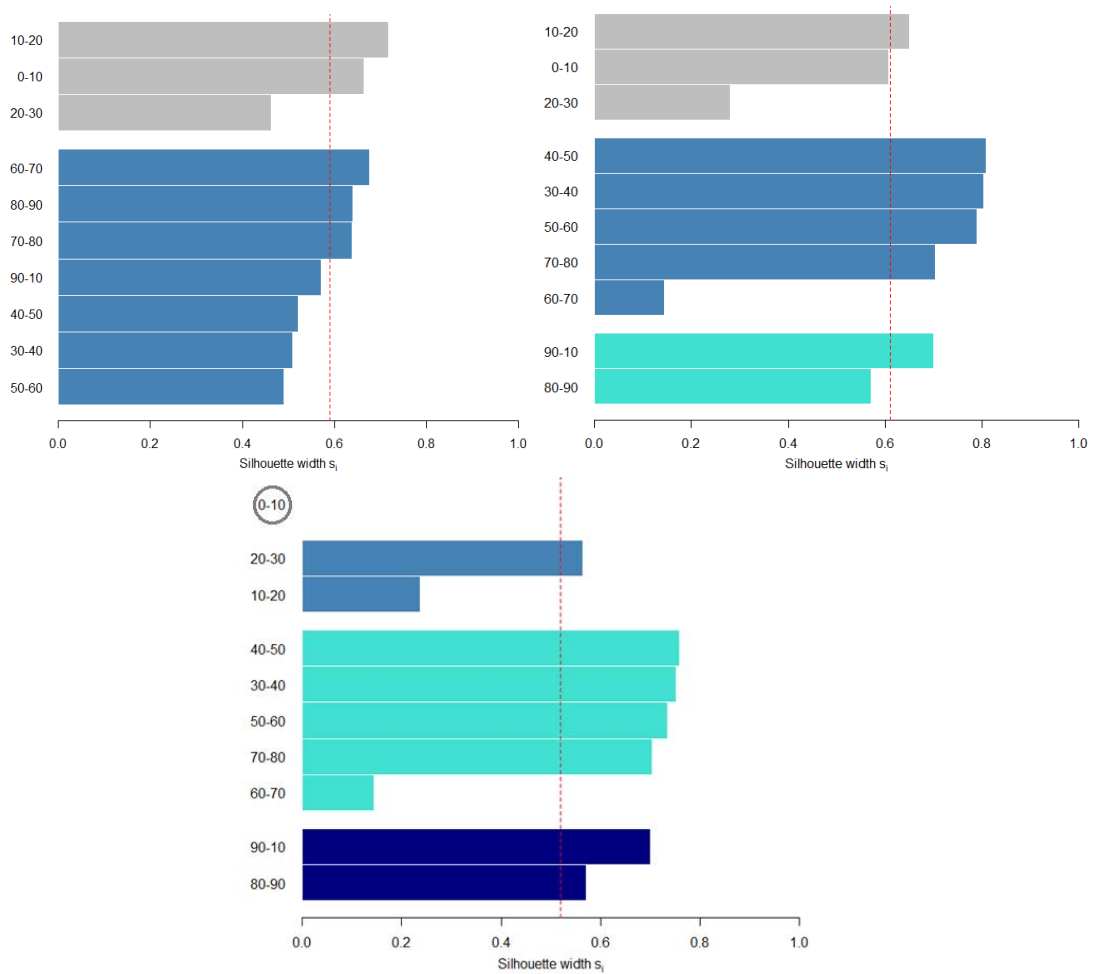


Figure 71. Clustering of IMD's band

A beneficial separation is performed from the first case to the second: 80-90 and 90-100 are grouped into an own cluster, resulting in a significant increase of the first blue cluster's individual silhouette.

Similar to the situation with Region cases, the separation taking place for the following case not only impairs whole average silhouette and individual silhouette for the second blue cluster, but also includes a group of cardinalities 1.

In accordance with the previous observations, the clustering selected for this variable's categories is that of the second case (3 clusters).

IMD_BAND	mean_successRate
0-10%	0.5721374
10-20%	0.6108541
20-30%	0.6456357
30-40%	0.6879975
40-50%	0.6956686
50-60%	0.6912920
60-70%	0.7502002
70-80%	0.7125606
80-90%	0.7601539
90-100%	0.7811295

Figure 72. Explicit clustering of IMD band

As observed for Highest education's groups, this variable's factors also imply a rank worth studying, with an indirect relationship between the success rate and the IMD band to which a student belongs, which would be ranked as follows, from lowest at-risk bands to the highest:

1. 90-100%
2. 80-90%
3. 70-80%
4. 60-70%
5. 50-60%
6. 40-50%
7. 30-40%
8. 20-30%
9. 10-20%
10. 0-10%

It is important to remind that these ranges would translate to belonging to the "x and y percentage of the most deprived councils", meaning that, as the percentage rises, there is less risk of deprivation.

6.1.2 Feature selection

As different procedures were applied for the identification of valuable features for our analysis, different sets of them were produced as an outcome of each of these processes.

- Feature importance

Different variable sets were produced in accordance to the measurements applied for filtering them, as detailed in the methodology description section.

A simple criterion was applied for this selection process: those with information gain/gain ratio value below the mean will be discarded. Although

concerns about the possible bias arising from the use of the mean, it will be lately exposed how the presence of very small values leads to the “selection threshold” decreasing, thus loosening the strictness with which a variable is discarded. Although this fact doesn’t end up playing any important role in the process, it was interpreted as a desirable approach to avoid excessive trimming of the original set, which would lead to the possibility of losing valuable information.

Information gain:

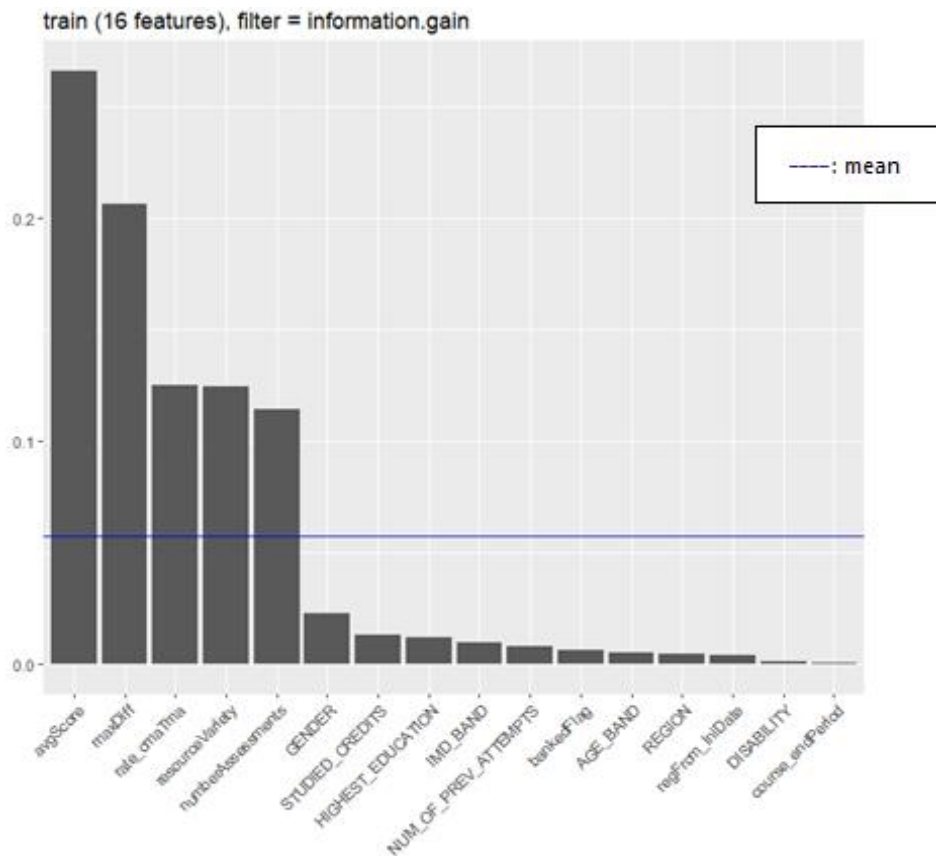


Figure 73. Features' importance with respect to mean interaction values (information gain)

The simplicity of the selection criterion aids the interpretation of this graph. The following features will compose a new set for the cases of study:

- avgScore
- maxDiff
- rate_cmaTma
- resourceVariety
- numberAssessments

Some hypothesis can be extracted out of it, specially about the influence of the type of content available for study (resourceVariety) and the presence or absence of computer monitored assessments in a course (rate_cmaTma).

Additionally, other expectable events become explicit, like the effect of the number of assessments (presumably, as it increases the interaction of students will

also do as a response) and the maximum difference between a student’s interaction days (assumption states that lower values of this variable are linked to a sustained interaction with the platform along time and, thus, higher values of it).

Finally, average score’s high value invites to infer that students achieving higher grades along the course are more prone to keep on interacting with the platform, although this cannot be objectively stated from the scenario we are working with (our average score and interaction values belong to ended courses, not in progress) and needs to be studied in detail.

Predictors	
avgScore	maxDiff
numberAssessments	rate cmaTma
resourceVariety	

Table 21. Predictors resulting from applying an Information Gain filter (mean interaction's regression tasks)

Information gain ratio:

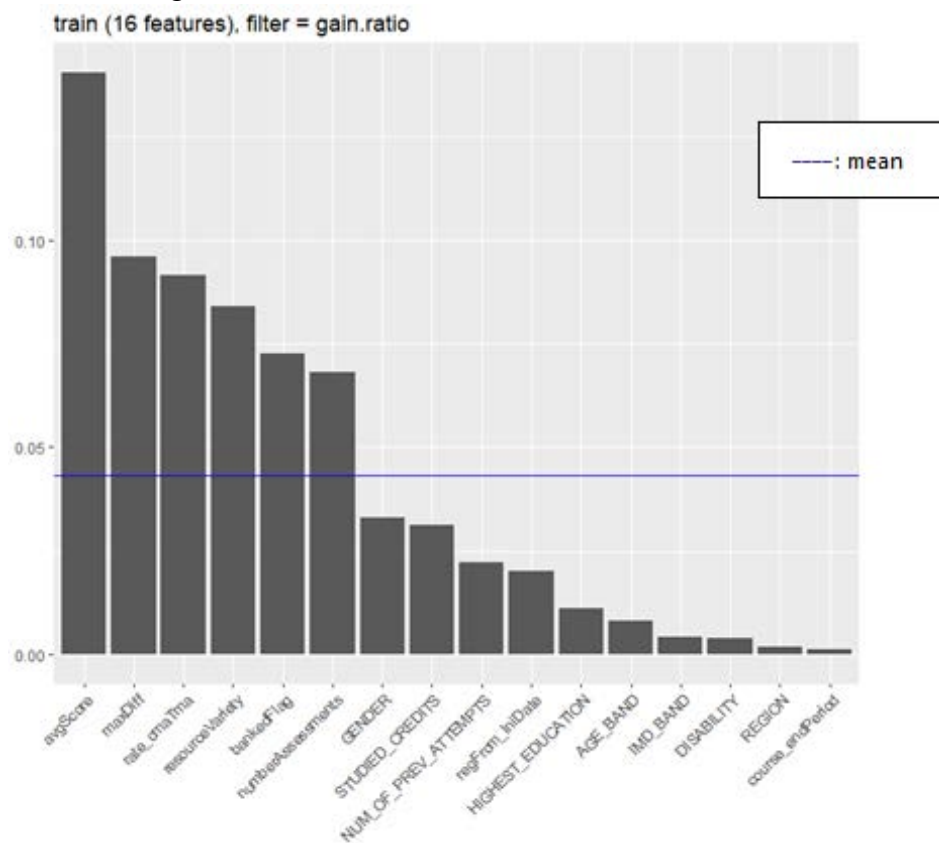


Figure 74. Features’ importance with respect to mean interaction values (information gain ratio)

The consideration of intrinsic information (i.e. potential information an attribute generates by splitting data according to its values, in which information gain ratio is based to smoothen regular information gain’s bias towards multivalued attributes) keeps the same set of valuable features as information gain with the addition of bankedFlag.

Thus, the new set will take into account the influence of having transferred scores from a previous course which, hypothetically, would imply a relief from current course's workload, although the way it would affect interaction cannot be assumed (it may lead to its decrease due to reduced workload as well as to its increase as the student would have more time to focus on few assessments). Placing effort in the assessment of this feature's effect may be valuable.

Predictors		
avgScore	maxDiff	rate_cmaTma
numberAssessments	bankedFlag	resourceVariety

Table 22. Predictors resulting from applying an Information Gain Ratio filter (mean interaction's regression tasks)

- Collinearity

In accordance with the methodology stated in its description's section, the three relationship measurements employed will be assessed separately.

Additionally, in account of avoiding redundancy, results from the analysis of this measures for the cases involving clustered or numerical-transformed variables are not included, as results do not vary (and they should not, as a transformation does not imply a change in the way the transformed variable relates to the predicted one).

Correlation of numerical features:

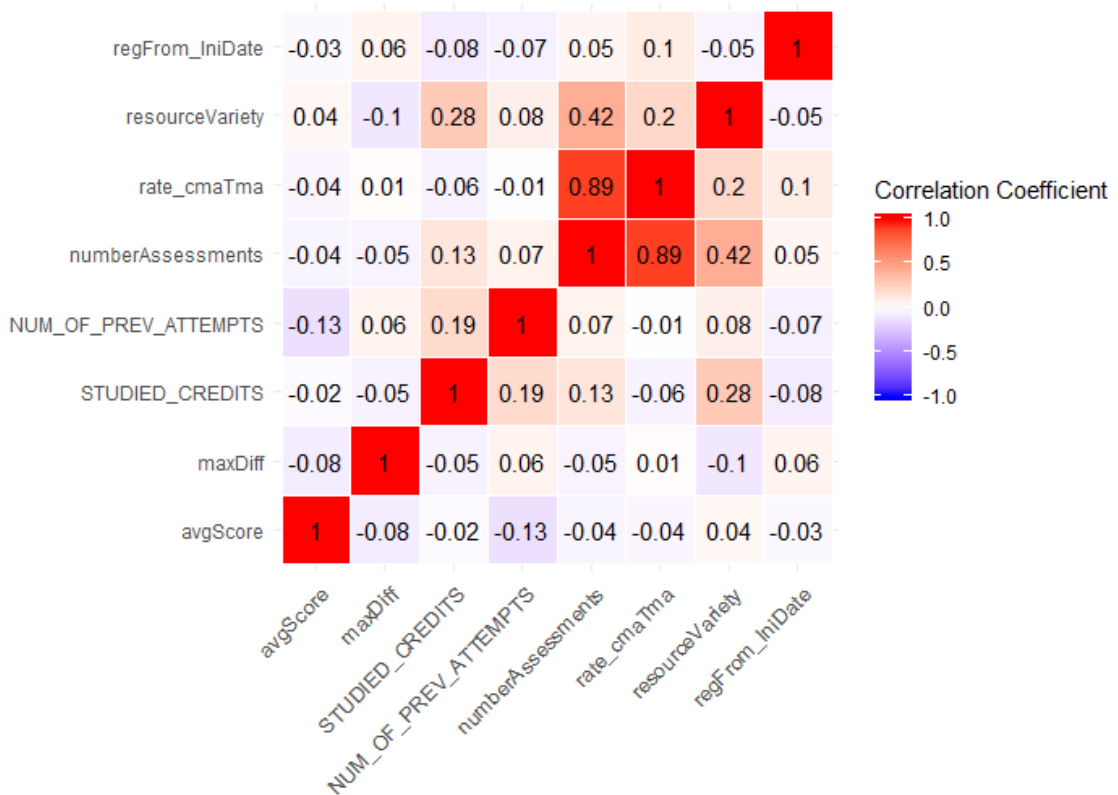


Figure 75. Correlation coefficient matrix of variable set's (numerical)

From the results exposed, it can be concluded that a high positive correlation exists between the number of assessments in a course (numberAssessments) and the rate of presence of computer monitored assessments (rate_cmaTma). Although this can be partially explained by the fact that the latter is computed from the former's data (along with each assessment's type-label), it explains how CMAs are an optional addition to a course's basic structure, composed of TMAs (teacher monitored assessments), fact that was formulated when it was observed that few courses had CMAs, whereas all of them had TMAs.

Following the criterion established for this case, although these variables are considered to contribute with different information (which in fact, cannot be directly inferred from one or the other alone), and observing the higher information gain and information gain ratio values for rate_cmaTma, numberAssessments is discarded from the final dataset.

Although not significant, it is also noticeable the relationship between the number of assessments in a course and the variety of its resources (resourceVariety). It may be interpreted as a response to student needs, which may be specific to certain assessments.

Cramer's V for categorical data:

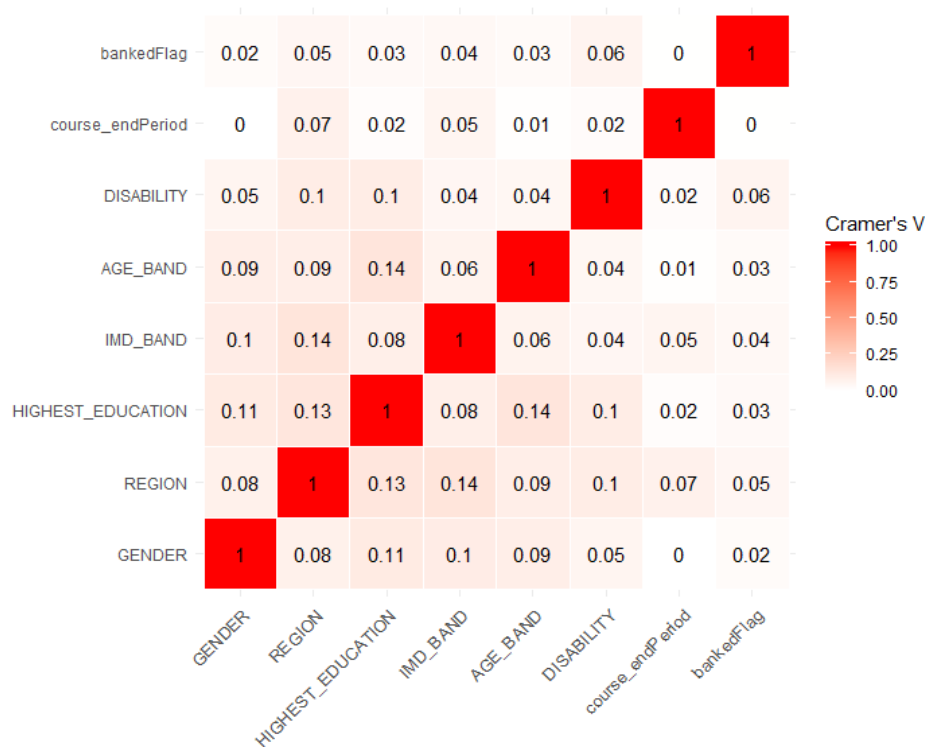


Figure 76. Cramer's V matrix of variable set's (categorical)

No relationship between categorical data can be elicited from the results presented in above's figure. Thus, no removal of categorical data is inferred from it.

Intra-Class Correlation – Categorical vs. Numerical data:



Figure 77. Intra-Class Correlation matrix of variable set's (categorical vs. numerical)

A significant relationship can be observed between the presence or absence of transferred scores (bankedFlag) and the number of attempts a student has undergone in a course (num_of_prev_attempts). This is expectable, since only students with at least one previous attempt can have scores transferred from a previous course. However, the fact that this relationship is not total (= 1) can be explained by the fact that not every student re-attempting a course has transferred scores.

Two different sets are extracted attending to the different rank these two variables have for information gain (NUM_OF_PREV_ATTEMPTS > bankedFlag) and information gain ratio (bankedFlag > NUM_OF_PREV_ATTEMPTS) values.

The following variable sets results from the collinearity assessment conducted:

Predictors		
avgScore	maxDiff	rate_cmaTma
course_endPeriod	gender	studied_credits
imd_band	disability	bankedFlag
region	regFrom_iniDate	resourceVariety
highest_education	age_band	

Table 23. Predictors resulting from applying a Collinearity filter based on Information Gain (mean interaction's regression tasks)

Predictors		
avgScore	maxDiff	rate_cmaTma
course_endPeriod	gender	studied_credits
imd_band	disability	num_of_prev_attempts
region	regFrom_iniDate	resourceVariety
highest_education	age_band	

Table 24. Predictors resulting from applying a Collinearity filter based on Information Gain Ratio (mean interaction's regression tasks)

6.1.3 Cases of study

After the processes previously described, a total of 13 cases of study were identified. This section will serve as a depiction of the set of features included in each of them, as well as to specify the source of each variable (the process from which it has been extracted, blank if no processing took place) and the IDs with which we will refer to in the results assessment section.

- Original variable set
 - ID: RAW

Predictors			
Name	Source	Name	Source
avgScore	--	rate_cmaTma	--
numberAssessments	--	studied_credits	--
imd_band	--	bankedFlag	--
region	--	disability	--
maxDiff	--	resourceVariety	--
gender	--	highest_education	--
num_of_prev_attempts	--	age_band	--
regFrom_iniDate	--	course_endPeriod	--

Table 25. Predictors present in "RAW" case of study (mean interaction's regression tasks)

- Clustered factors of categorical variables
 - **ID: CLUST**

Predictors			
Name	Source	Name	Source
avgScore	--	rate_cmaTma	--
numberAssessments	--	studied_credits	--
imd_band	Category clustering	bankedFlag	--
region	Category clustering	disability	--
maxDiff	--	resourceVariety	--
gender	--	highest_education	Category clustering
num of prev attempts	--	age_band	--
regFrom iniDate	--	course_endPeriod	--

Table 26. Predictors present in “CLUST” case of study (mean interaction’s regression tasks)

- Transformation of categorical variables to numerical
 - **ID: NUM**

Predictors			
Name	Source	Name	Source
avgScore	--	rate_cmaTma	--
numberAssessments	--	studied_credits	--
imd_band	Numerical transformation	bankedFlag	--
region	Numerical transformation	disability	--
maxDiff	--	resourceVariety	--
gender	--	highest_education	Numerical transformation
num of prev attempts	--	age_band	--
regFrom iniDate	--	course_endPeriod	--

Table 27. Predictors present in “NUM” case of study (mean interaction’s regression tasks)

- Information gain filtering
 - **ID: IG**

Predictors			
Name	Source	Name	Source
avgScore	--	maxDiff	--
numberAssessments	--	rate_cmaTma	--
resourceVariety	--		

Table 28. Predictors present in “IG” case of study (mean interaction’s regression tasks)

- Information gain ratio filtering
 - **ID: IGR**

Predictors			
Name	Source	Name	Source
avgScore	--	maxDiff	--
numberAssessments	--	rate_cmaTma	--
resourceVariety	--	bankedFlag	--

Table 29. Predictors present in “IGR” case of study (mean interaction’s regression tasks)

- Collinearity set #1
 - ID: COLL1

Predictors			
Name	Source	Name	Source
avgScore	--	rate cmaTma	--
regFrom iniDate	--	studied credits	--
imd band	--	disability	--
region	--	resourceVariety	--
maxDiff	--	highest education	--
gender	--	age band	--
num of prev attempts	--	course endPeriod	--

Table 30. Predictors present in “COLL1” case of study (mean interaction’s regression tasks)

- Collinearity set #2
 - ID: COLL2

Predictors			
Name	Source	Name	Source
avgScore	--	rate cmaTma	--
course endPeriod	--	studied credits	--
imd band	--	bankedFlag	--
region	--	disability	--
maxDiff	--	resourceVariety	--
gender	--	highest education	--
regFrom iniDate	--	age band	--

Table 31. Predictors present in “COLL2” case of study (mean interaction’s regression tasks)

- Collinearity set #1 and categorical factor clustering
 - ID: COLL1+CLUST

Predictors			
Name	Source	Name	Source
avgScore	--	rate cmaTma	--
numberAssessments	--	studied credits	--
imd band	Category clustering	disability	--
region	Category clustering	resourceVariety	--
maxDiff	--	highest_education	Category clustering
gender	--	age band	--
num of prev attempts	--	course endPeriod	--
regFrom iniDate	--		

Table 32. Predictors present in “COLL1+CLUST” case of study (mean interaction’s regression tasks)

- Collinearity set #1 and numerical transformation of categorical variables
 - **ID:** COLL1+NUM

Predictors			
Name	Source	Name	Source
avgScore	--	rate_cmaTma	--
numberAssessments	--	studied_credits	--
imd_band	Numerical transformation	disability	--
region	Numerical transformation	resourceVariety	--
maxDiff	--	highest_education	Numerical transformation
gender	--	age_band	--
num_of_prev_attempts	--	course_endPeriod	--
regFrom_iniDate	--		

Table 33. Predictors present in “COLL1+NUM” case of study (mean interaction’s regression tasks)

- Collinearity set #2 and categorical factor clustering
 - **ID:** COLL2+CLUST

Predictors			
Name	Source	Name	Source
avgScore	--	rate_cmaTma	--
numberAssessments	--	studied_credits	--
imd_band	Category clustering	bankedFlag	--
region	Category clustering	disability	--
maxDiff	--	resourceVariety	--
gender	--	highest_education	Category clustering
regFrom_iniDate	--	age_band	--
course_endPeriod	--		

Table 34. Predictors present in “COLL2+CLUST” case of study (mean interaction’s regression tasks)

- Collinearity set #2 and numerical transformation of categorical variables
 - **ID:** COLL2+NUM

Predictors			
Name	Source	Name	Source
avgScore	--	rate_cmaTma	--
numberAssessments	--	studied_credits	--
imd_band	Numerical transformation	bankedFlag	--
region	Numerical transformation	disability	--
maxDiff	--	resourceVariety	--
gender	--	highest_education	Numerical transformation
regFrom_iniDate	--	age_band	--
course_endPeriod	--		

Table 35. Predictors present in “COLL2+NUM” case of study (mean interaction’s regression tasks)

- Information gain filtering and collinearity
 - **ID: IG+COLL**

Predictors			
Name	Source	Name	Source
avgScore	--	maxDiff	--
resourceVariety	--	rate_cmaTma	--

Table 36. Predictors present in "IG+COLL" case of study (mean interaction's regression tasks)

- Information gain ratio filtering and collinearity
 - **ID: IGR+COLL**

Predictors			
Name	Source	Name	Source
avgScore	--	maxDiff	--
numberAssessments	--	rate_cmaTma	--
bankedFlag	--		

Table 37. Predictors present in "IGR+COLL" case of study (mean interaction's regression tasks)

There are no occurrences of cases jointly assessing information gain/gain ratio with factor clustering or numerical transformation since the former filtering removes the categorical variables involved in the latter processes.

Additionally, when evaluating the information gain/gain ratio filtering along with collinearity, there is no need for distinction between the two different types of collinearity defined, since both former filtering process discard NUM_OF_PREV_ATTEMPTS, which comparison with bankedFlag is the cause of this distinction.

Standardization of variables was performed for every case of study shown in this section.

6.1.4 Selected algorithms

The selection of algorithms to with which the performance of this regression task is going to be evaluated follows the guidelines of ([46]) for an appropriate choose of algorithms in real-world problem solving (although the reference discusses classification problems, its conclusions can also be extended to a regression scenario).

The following algorithms have been selected, aiming to the maintenance of diversity in which respects to the family to which each classifier belongs (based on statistic approaches, artificial intelligence, connectionist approaches ([46]):

- Penalized regression
 - **ID:** PEN
- Neural network
 - **ID:** NNET
- Random Forest
 - **ID:** RF
- Support Vector Machine
 - **ID:** SVM
- Gradient Boosting Machine
 - **ID:** GBM

6.1.5 Results

RAW	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	3.666353	3.024514	3.024514	3.112313	2.772323
TsRMSE	3.668253	2.909256	1.304475	2.94046	2.490461
TrRRSE	0.739959	0.610281	0.610281	0.628031	0.559522
TsRRSE	0.740224	0.587069	0.263233	0.593374	0.502549
ResVar	13.46134	8.466744	1.701951	8.336236	6.214712

Table 38. Results from “RAW” case of study (mean interaction’s regression tasks)

CLUST	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	3.671501	2.967157	2.769279	3.110683	2.757348
TsRMSE	3.657359	2.902504	1.405686	2.991015	2.591294
TrRRSE	0.739711	0.596836	0.55714	0.625427	0.554811
TsRRSE	0.736807	0.584718	0.283194	0.602574	0.522038
ResVar	13.37947	8.428141	1.97579	8.624339	6.722325

Table 39. Results from “CLUST” case of study (mean interaction’s regression tasks)

NUM	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	3.615617	2.933197	2.766695	3.099019	2.761398
TsRMSE	3.596317	2.871304	1.360082	2.93329	2.581995
TrRRSE	0.728076	0.588807	0.55542	0.621925	0.554477
TsRRSE	0.72422	0.578201	0.273879	0.590682	0.519927
ResVar	12.93488	8.245495	1.849375	8.319833	6.668566

Table 40. Results from “NUM” case of study (mean interaction’s regression tasks)

IG	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	3.645509	2.892823	2.819164	3.049207	2.825442
TsRMSE	3.644003	2.955483	3.101907	3.00105	2.74923
TrRRSE	0.731765	0.579871	0.565424	0.61134	0.566522
TsRRSE	0.731399	0.593203	0.622596	0.60236	0.551778
ResVar	13.27992	8.736046	9.622743	8.754003	7.560981

Table 41. Results from "IG" case of study (mean interaction's regression tasks)

IGR	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	3.621705	2.875305	2.807703	3.009568	2.810583
TsRMSE	3.61954	2.832483	2.643892	2.959965	2.706478
TrRRSE	0.730459	0.578894	0.56521	0.606017	0.565999
TsRRSE	0.730005	0.571257	0.533232	0.596982	0.545859
ResVar	13.10225	8.027471	6.990747	8.523806	7.32684

Table 42. Results from "IGR" case of study (mean interaction's regression tasks)

COLL1	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	3.671985	3.092626	2.80121	3.295181	2.779274
TsRMSE	3.667084	2.911839	1.33432	3.210714	2.653382
TrRRSE	0.739352	0.623369	0.564753	0.664166	0.560251
TsRRSE	0.738395	0.586314	0.268683	0.646439	0.534326
ResVar	13.45007	8.479764	1.780587	9.903921	7.045732

Table 43. Results from "COLL1" case of study (mean interaction's regression tasks)

COLL2	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	3.641207	3.085838	2.790165	3.161279	2.763652
TsRMSE	3.648498	2.871011	1.340905	3.051373	2.578556
TrRRSE	0.733544	0.62259	0.56289	0.637778	0.557429
TsRRSE	0.735013	0.578351	0.270131	0.61473	0.519458
ResVar	13.31511	8.245658	1.798255	8.965913	6.649942

Table 44. Results from "COLL2" case of study (mean interaction's regression tasks)

COLL1+CLUST	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	3.673552	2.994119	2.764016	3.178057	2.761989
TsRMSE	3.687141	2.92005	1.447119	3.053626	2.674239
TrRRSE	0.735815	0.598112	0.552041	0.634736	0.551687
TsRRSE	0.738417	0.584796	0.289812	0.611541	0.535554
ResVar	13.5963	8.527827	2.094026	8.996706	7.154446

Table 45. Results from "COLL1+CLUST" case of study (mean interaction's regression tasks)

COLL1+NUM	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	3.612147	2.960078	2.781693	3.171643	2.768066
TsRMSE	3.601613	2.859174	1.385721	3.021711	2.615588
TrRRSE	0.728105	0.594289	0.558449	0.63685	0.555771
TsRRSE	0.725932	0.576304	0.279306	0.609058	0.527216
ResVar	12.97337	8.175835	1.919692	8.815064	6.845729

Table 46. Results from "COLL1+NUM" case of study (mean interaction's regression tasks)

COLL2+CLUST	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	3.631538	2.985563	2.769972	3.142185	2.758297
TsRMSE	3.633173	2.867869	1.454881	3.035213	2.63734
TrRRSE	0.732154	0.602893	0.559162	0.63417	0.556795
TsRRSE	0.732368	0.578111	0.293267	0.611826	0.531616
ResVar	13.2025	8.225663	2.11665	8.892342	6.956809

Table 47. Results from "COLL2+CLUST" case of study (mean interaction's regression tasks)

COLL2+NUM	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	3.606814	2.983944	2.775548	3.212571	2.778785
TsRMSE	3.62934	2.912633	1.399139	3.097191	2.620825
TrRRSE	0.727081	0.600736	0.558907	0.646873	0.559612
TsRRSE	0.731682	0.587186	0.28207	0.624393	0.528357
ResVar	13.17347	8.484838	1.957141	9.231473	6.871608

Table 48. Results from “COLL2+NUM” case of study (mean interaction’s regression tasks)

IG+COLL	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	3.621714	2.861452	2.822561	3.039863	2.815776
TsRMSE	3.620365	2.858228	2.889301	2.864239	2.704179
TrRRSE	0.730259	0.577419	0.569598	0.613334	0.568088
TsRRSE	0.729925	0.576255	0.582532	0.577486	0.545211
ResVar	13.10858	8.178662	8.349406	8.018326	7.316466

Table 49. Results from “IG+COLL” case of study (mean interaction’s regression tasks)

IGR+COLL	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	3.643727	2.899167	2.847411	3.039732	2.830237
TsRMSE	3.640638	2.883359	3.188651	2.979762	2.700718
TrRRSE	0.730373	0.580036	0.569792	0.608363	0.566379
TsRRSE	0.729613	0.57781	0.639041	0.597163	0.541238
ResVar	13.25548	8.319996	10.16856	8.646379	7.294671

Table 50. Results from “IGR+COLL” case of study (mean interaction’s regression tasks)

6.1.5.1 Summary and discussion

It can be observed from the results shown that a clear rank can be established in which respects to the performative capabilities of each algorithm assessed (from best to worst):

1. Random Forest
2. Gradient Boosting Machine
3. Neural Network
4. Support Vector Machine
5. Penalized regression

Although this ordering is a generalization (some results exist in which GBM slightly outperforms Random Forest), it accurately captures the main pattern existing in the results presented.

It also aids to elicit the fact that there is a clear performance difference between algorithms based on decision trees (Random Forest and GBM) and the others (Neural Network, SVM and Penalized regression, which are based on parameters’ coefficient adjustment).

The first hypothesis that come to mind when aiming to explain the cause for this issue are related to both the ability of each algorithm to treat categorical data and to deal with high dimensionality data (which is our case). These two possibilities are discarded since it can be extracted from the results how cases implying a numerical transformation of categorical data and/or lower dimensionality didn’t imply any significant change in the performance difference between algorithms addressed.

A more plausible explanation is related to the fact that the three worst-performing algorithms are the ones with higher reliability in a precise parameter tuning process (Neural Networks' performance is heavily dependent on an adequate number of hidden layers and neurons per layer, as SVM is on the setting of an appropriate kernel and Penalized regression is on the selection of an appropriate model to attempt to describe the data).

As a more in-depth approach to this scenario, the fact that Neural Network is superior than SVM and Penalized regression for all cases may imply the fact that data distribution is complex enough to not be adequately captured by a simple linear/non-linear type of description (as that of SVM and Penalized regression), thus needing a more refined process for coefficient assignment (as that of Neural Networks).

Consequently, decision trees, and specifically Random Forest and GBM, may be better capturing the mentioned pattern complexity by iteratively sub-setting the data-space.

Finally, as an introductory approach to the selection of the model and algorithm to consider as more appropriate for this case, each case of study's best result and algorithm is shown in the following table:

	RAW	CLUST	NUM	IGR	COLL1
	Random Forest				
TrRMSE	3.024514	2.769279	2.766695	2.807703	2.80121
TsRMSE	1.304475	1.405686	1.360082	2.643892	1.33432
TrRRSE	0.610281	0.55714	0.55542	0.56521	0.564753
TsRRSE	0.263233	0.283194	0.273879	0.533232	0.268683
ResVar	1.701951	1.97579	1.849375	6.990747	1.780587
	COLL2	COLL1+CLUST	COLL1+NUM	COLL2+CLUST	COLL2+NUM
	Random Forest				
TrRMSE	2.790165	2.764016	2.781693	2.769972	2.775548
TsRMSE	1.340905	1.447119	1.385721	1.454881	1.399139
TrRRSE	0.56289	0.552041	0.558449	0.559162	0.558907
TsRRSE	0.270131	0.289812	0.279306	0.293267	0.28207
ResVar	1.798255	2.094026	1.919692	2.11665	1.957141
	IG	IG+COLL	IGR+COLL		
	GBM				
TrRMSE	2.825442	2.815776	2.830237		
TsRMSE	2.74923	2.704179	2.700718		
TrRRSE	0.566522	0.568088	0.566379		
TsRRSE	0.551778	0.545211	0.541238		
ResVar	7.560981	7.316466	7.294671		

Table 51. Summary of best regression models and algorithms (mean interaction)

It is remarkable how all cases based on an information gain/gain ratio filter imply a significant impairment of the performance measures, which leads to think that valuable attributes were removed from those models.

Additionally, it can be concluded that no transformation of categorical data led to any type of improvement. In fact, there is a slight impairment of the performance measures.

The only cases that may set a doubt with respect to their selection against that of the raw model (the best performing model) are the ones related to the collinearity filters alone. However, and although a reasoning based on parsimony (prevalence of the model with

less attributes) may justify their selection, there is no other objective nor performance-based evidence to pick them over the raw model.

Consequently, it can be concluded that the raw model, treated with a Random Forest algorithm, is the most suitable for the conduction of regression tasks on this scenario.

6.2 Average score (prediction task)

The building of a prediction model for the average score of a student (with courses finished) has considered the same initial set of features used for the mean interaction model, with the exception of the substitution of the average score variable (the predicted variable for this model) with mean interaction, which is now a predictor.

Predictors			
mean_interaction	maxDiff	rate_cmaTma	resourceVariety
numberAssessments	gender	studied_credits	highest_education
imd_band	num_of_prev_attempts	bankedFlag	age_band
region	regFrom_iniDate	disability	course_endPeriod

Table 52. Predictors employed for average score's regression tasks

6.2.1 Treatment of categorical data

In account of simplicity and avoid de redundancy of information, it is important to point out the fact that the same categorical variables involved in the modelling of mean interaction are present in this process. Coherently with it, the same factor clustering and numerical transformation took place for these variables.

Following, a summary of these procedures is shown:

Categorical variables transformed to numerical:

- Highest education
- IMD band
- Age band
- Region

Categorical variables which factors were clustered:

- Highest education: from 5 to 2

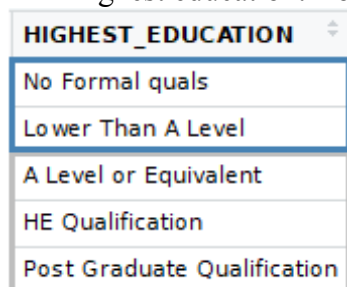


Figure 78. Explicit clustering of Highest education level

- IMD band: from 10 to 3

IMD_BAND
0-10%
10-20%
20-30%
30-40%
40-50%
50-60%
60-70%
70-80%
80-90%
90-100%

Figure 79. Explicit clustering of IMD band

- Region: from 13 to 5

REGION
Wales
North Western Region
London Region
Scotland
Yorkshire Region
North Region
Ireland
West Midlands Region
East Anglian Region
East Midlands Region
South West Region
South Region
South East Region

Figure 80. Explicit clustering of Region

6.2.2 Feature selection

Same processes of information gain/gain ratio feature filtering and collinearity assessment as for mean interaction modelling have been conducted for this case.

- Feature importance

The criterion for the setting of the “removal threshold” at the information gain/gain ratio mean has been maintained.

Information gain:

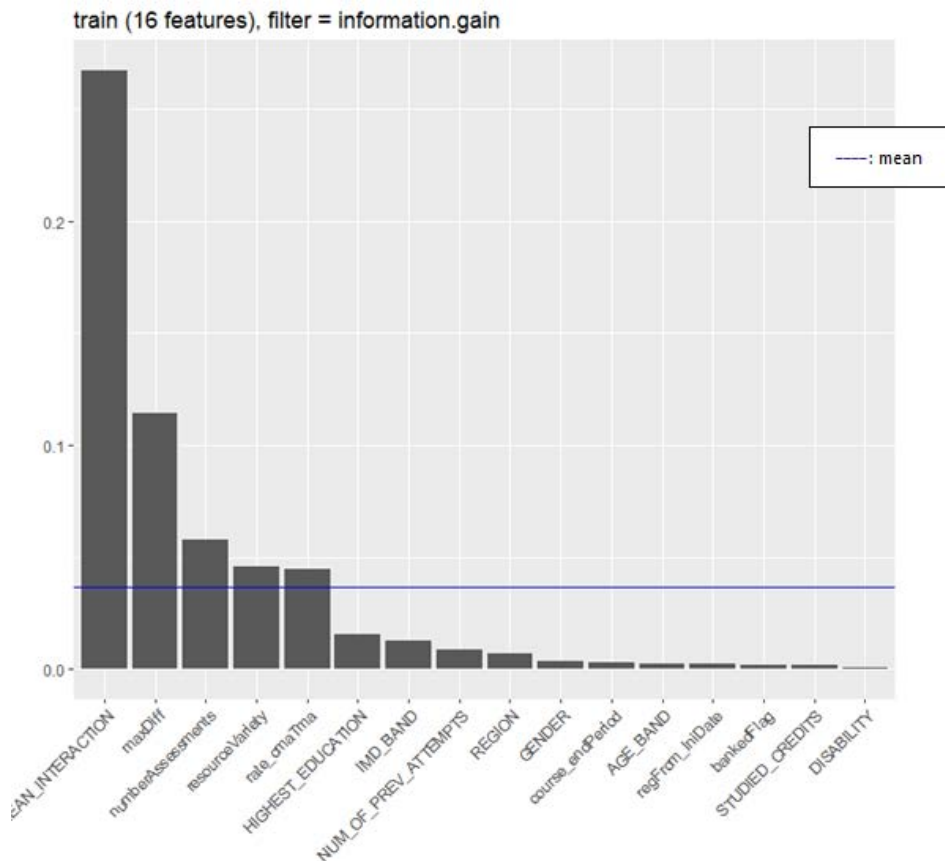


Figure 81. Features' importance with respect to mean interaction values (information gain)

The following features will compose a new set for the cases of study:

- MEAN_INTERACTION
- maxDiff
- numberAssessments
- rate_cmaTma
- resourceVariety

It is remarkable how, apart from mean interaction, the same subset of main attributes identified to be affecting average score are the same as for mean interaction's case with a lower information gain value. This may be partially explained by the fact that both mean interaction and average score play an important role in describing the other, thus being influenced by the same features (presumably not in the same way).

The fact that maxDiff has a significant information gain value reinforces the potential value of analysing the already formulated hypothesis that students with smaller gaps in their interaction with the platform may have better results, produced by the higher values of interaction these smaller gaps imply, among other factors.

The number of assessments, variety of resources and rate of presence of CMAs (rate_cmaTma) lead to think that both workload and content or assessments preference do play an important role in achievement.

Predictors	
mean_interaction	maxDiff
numberAssessments	rate_cmaTma
resourceVariety	

Table 53. Predictors resulting from applying an Information Gain filter (average score's regression tasks)

Information gain ratio:

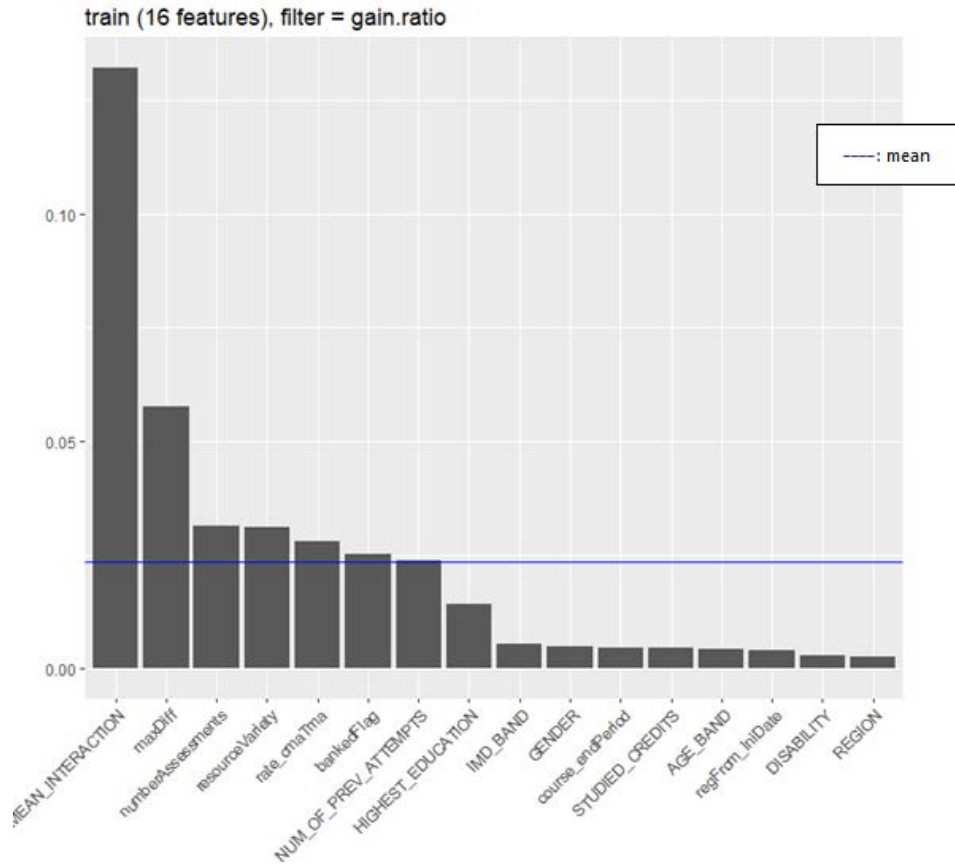


Figure 82. Features' importance with respect to mean interaction values (information gain ratio)

This distinct filtering process results in the addition of bankedFlag and NUM_OF_PREV_ATTEMPTS to the set of features identified by considering information gain.

The addition of these two factors (which correlation was already discussed) supports the hypothesis that having transferred scores from previous courses imply a relief in the workload experienced by the student, which may redound in better overall grades.

Predictors		
mean_interaction	maxDiff	rate_cmaTma
numberAssessments	resourceVariety	bankedFlag
num_of_prev_attempts		

Table 54. Predictors resulting from applying an Information Gain Ratio filter (average score's regression tasks)

- Collinearity

Correlation of numerical features:

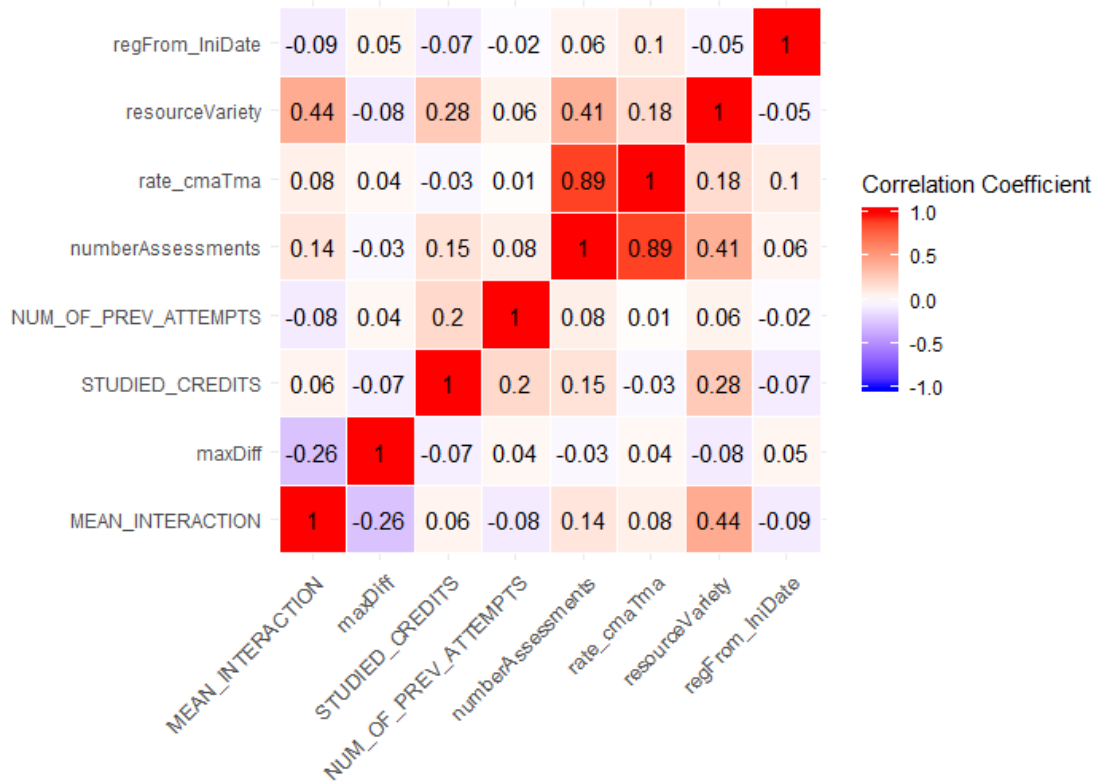


Figure 83. Correlation coefficient matrix of variable set's (numerical)

As expected, the same correlation phenomena observed in the collinearity assessment of mean interaction modelling process can be observed in this case (rate_cmaTma and numberAssessments, which is also shows a certain relationship with resourceVariety).

The remarkable difference for this evaluation resides in the information gain/gain ratio criteria, which places more value on the number of assessments per course than on the rate of CMAs presence for the modelling of the average score of a student.

Additionally, mean interaction shows relation (not enough for removal) with the variety of resources within a course. This could be explained by the influence of the latter on mean interaction, as observed during the previous modelling process. A broader variety of resources may imply interaction values as a students' response to the need of reviewing more content during a course.

Cramer's V for categorical data:

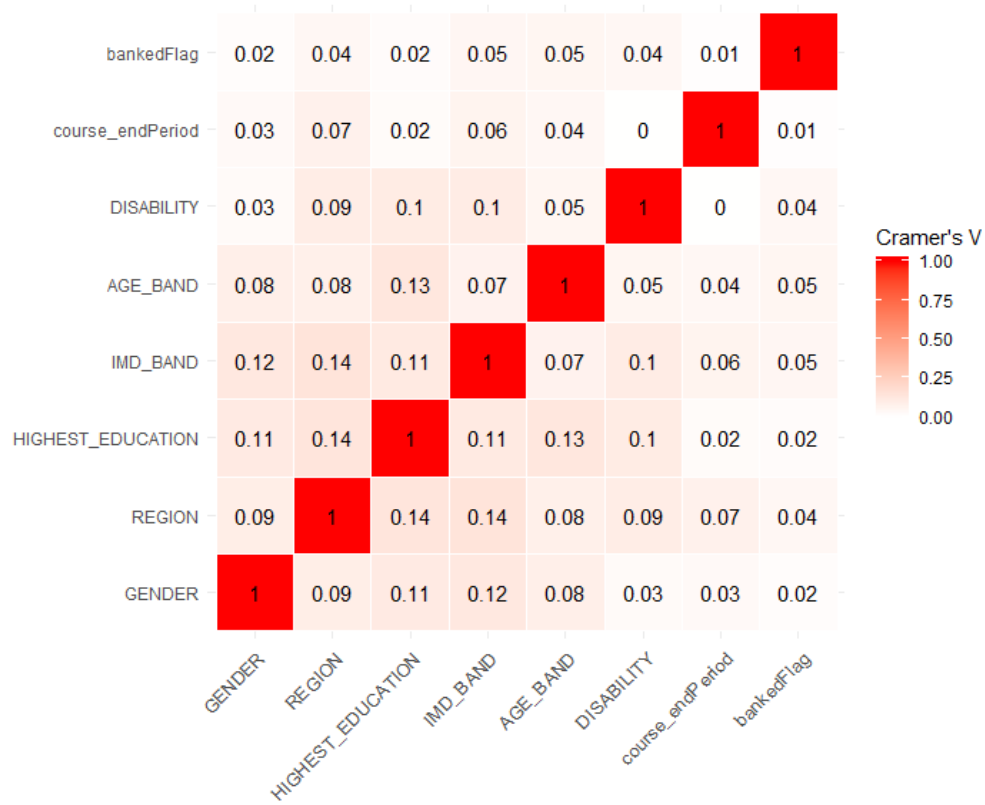


Figure 84. Cramer's V matrix of variable set's (categorical)

No relationship between categorical data can be elicited from the results presented in above's figure. Thus, no removal of categorical data is inferred from it.

Intra-Class Correlation – Categorical vs. Numerical data:

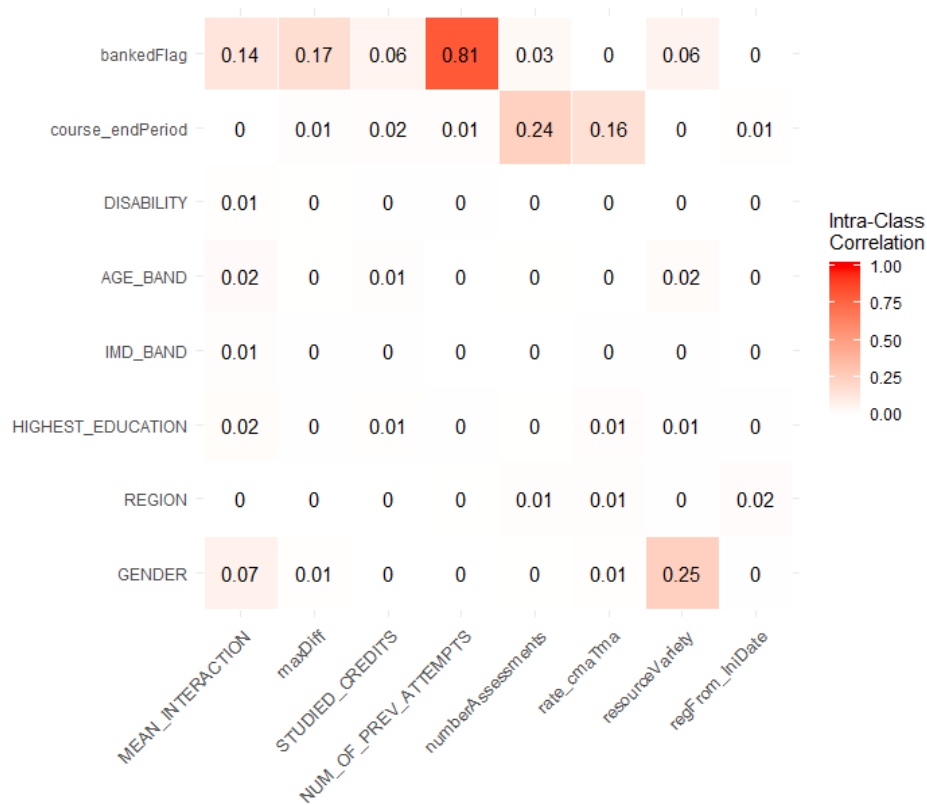


Figure 85. Intra-Class Correlation matrix of variable set's (categorical vs. numerical)

Same observations as made in mean interaction's modelling process are extracted from these results, which showcase the implicit relationship between having transferred scores from a previous course and the number of times a student has attempted a course (which is a necessary condition for having scores transferred).

Once again, two different sets are extracted attending to the different rank these two variables have for information gain ($NUM_OF_PREV_ATTEMPTS > bankedFlag$) and information gain ratio ($bankedFlag > NUM_OF_PREV_ATTEMPTS$) values.

The following variable sets results from the collinearity assessment conducted:

Predictors		
mean interaction	maxDiff	rate_cmaTma
course_endPeriod	gender	studied credits
imd band	disability	bankedFlag
region	regFrom_iniDate	resourceVariety
highest education	age band	

Table 55. Predictors resulting from applying a Collinearity filter based on Information Gain (average score's regression tasks)

Predictors		
avgScore	maxDiff	rate_cmaTma
course_endPeriod	gender	studied_credits
imd_band	disability	num_of_prev_attempts
region	regFrom_iniDate	resourceVariety
highest_education	age_band	

Table 56. Predictors resulting from applying a Collinearity filter based on Information Gain Ratio (average score's regression tasks)

6.2.3 Cases of study

This section will give detail of the set of features included in each of the cases of study identified, along with each variable's source and their IDs.

- Original variable set
 - ID: RAW

Predictors			
Name	Source	Name	Source
mean_interaction	--	rate_cmaTma	--
numberAssessments	--	studied_credits	--
imd_band	--	bankedFlag	--
region	--	disability	--
maxDiff	--	resourceVariety	--
gender	--	highest_education	--
num_of_prev_attempts	--	age_band	--
regFrom_iniDate	--	course_endPeriod	--

Table 57. Predictors present in "RAW" case of study (average score's regression tasks)

- Clustered factors of categorical variables
 - ID: CLUST

Predictors			
Name	Source	Name	Source
mean_interaction	--	rate_cmaTma	--
numberAssessments	--	studied_credits	--
imd_band	Category clustering	bankedFlag	--
region	Category clustering	disability	--
maxDiff	--	resourceVariety	--
gender	--	highest_education	Category clustering
num_of_prev_attempts	--	age_band	--
regFrom_iniDate	--	course_endPeriod	--

Table 58. Predictors present in "CLUST" case of study (average score's regression tasks)

- Transformation of categorical variables to numerical
 - **ID: NUM**

Predictors			
Name	Source	Name	Source
mean interaction	--	rate cmaTma	--
numberAssessments	--	studied credits	--
imd_band	Numerical transformation	bankedFlag	--
region	Numerical transformation	disability	--
maxDiff	--	resourceVariety	--
gender	--	highest_education	Numerical transformation
num of prev attempts	--	age band	--
regFrom iniDate	--	course endPeriod	--

Table 59. Predictors present in “NUM” case of study (average score's regression tasks)

- Information gain filtering
 - **ID: IG**

Predictors			
Name	Source	Name	Source
mean interaction	--	maxDiff	--
numberAssessments	--	rate cmaTma	--
resourceVariety	--		

Table 60. Predictors present in “IG” case of study (average score's regression tasks)

- Information gain ratio filtering
 - **ID: IGR**

Predictors			
Name	Source	Name	Source
mean interaction	--	maxDiff	--
numberAssessments	--	rate cmaTma	--
resourceVariety	--	bankedFlag	--
num of prev attempts	--		

Table 61. Predictors present in “IGR” case of study (average score's regression tasks)

- Collinearity set #1
 - **ID: COLL1**

Predictors			
Name	Source	Name	Source
mean interaction	--	regFrom iniDate	--
numberAssessments	--	studied credits	--
imd_band	--	disability	--
region	--	resourceVariety	--
maxDiff	--	highest education	--
gender	--	age band	--
num of prev attempts	--	course endPeriod	--

Table 62. Predictors present in “COLL1” case of study (average score's regression tasks)

- Collinearity set #2
 - **ID: COLL2**

Predictors			
Name	Source	Name	Source
mean_interaction	--	course_endPeriod	--
numberAssessments	--	studied_credits	--
imd_band	--	bankedFlag	--
region	--	disability	--
maxDiff	--	resourceVariety	--
gender	--	highest_education	--
regFrom_iniDate	--	age_band	--

Table 63. Predictors present in “COLL2” case of study (average score's regression tasks)

- Collinearity set #1 and categorical factor clustering
 - **ID: COLL1+CLUST**

Predictors			
Name	Source	Name	Source
mean_interaction	--	regFrom_iniDate	--
numberAssessments	--	studied_credits	--
imd_band	Category clustering	disability	--
region	Category clustering	resourceVariety	--
maxDiff	--	highest_education	Category clustering
gender	--	age_band	--
num_of_prev_attempts	--	course_endPeriod	--

Table 64. Predictors present in “COLL1+CLUST” case of study (average score's regression tasks)

- Collinearity set #1 and numerical transformation of categorical variables
 - **ID: COLL1+NUM**

Predictors			
Name	Source	Name	Source
mean_interaction	--	regFrom_iniDate	--
numberAssessments	--	studied_credits	--
imd_band	Numerical transformation	disability	--
region	Numerical transformation	resourceVariety	--
maxDiff	--	highest_education	Numerical transformation
gender	--	age_band	--
num_of_prev_attempts	--	course_endPeriod	--

Table 65. Predictors present in “COLL1+NUM” case of study (average score's regression tasks)

- Collinearity set #2 and categorical factor clustering
 - **ID: COLL2+CLUST**

Predictors			
Name	Source	Name	Source
mean_interaction	--	course_endPeriod	--
numberAssessments	--	studied_credits	--
imd_band	Category clustering	bankedFlag	--
region	Category clustering	disability	--
maxDiff	--	resourceVariety	--
gender	--	highest_education	Category clustering
regFrom_iniDate	--	age_band	--

Table 66. Predictors present in “COLL2+CLUST” case of study (average score's regression tasks)

- Collinearity set #2 and numerical transformation of categorical variables
 - **ID: COLL2+NUM**

Predictors			
Name	Source	Name	Source
mean_interaction	--	course_endPeriod	--
numberAssessments	--	studied_credits	--
imd_band	Numerical transformation	bankedFlag	--
region	Numerical transformation	disability	--
maxDiff	--	resourceVariety	--
gender	--	highest_education	Numerical transformation
regFrom_iniDate	--	age_band	--

Table 67. Predictors present in “COLL2+NUM” case of study (average score's regression tasks)

- Information gain filtering and collinearity
 - **ID: IG+COLL**

Predictors			
Name	Source	Name	Source
mean_interaction	--	maxDiff	--
resourceVariety	--	rate_cmaTma	--

Table 68. Predictors present in “IG+COLL” case of study (average score's regression tasks)

- Information gain ratio filtering and collinearity
 - **ID: IGR+COLL**

Predictors			
Name	Source	Name	Source
mean_interaction	--	maxDiff	--
numberAssessments	--	bankedFlag	--
resourceVariety	--		

Table 69. Predictors present in “IGR+COLL” case of study (average score's regression tasks)

Standardization of variables was performed for every case of study here shown.

6.2.4 Selected algorithms

This process was conducted following the same reasoning as for the mean interaction modelling process (they are both regression problems). Guidelines from ([46]) were followed to maintain diversity of algorithm's family and good overall performance.

Consequently, the same algorithms as for mean interaction modelling were selected:

- Penalized regression
 - **ID:** PEN
- Neural network
 - **ID:** NNET
- Random Forest
 - **ID:** RF
- Support Vector Machine
 - **ID:** SVM
- Gradient Boosting Machine
 - **ID:** GBM

6.2.5 Results

RAW	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	23.91083	16.04891	14.73977	19.16286	14.80208
TsRMSE	23.8471	15.09063	6.833463	17.88211	13.14192
TrRRSE	0.847886	0.56847	0.521987	0.678325	0.524102
TsRRSE	0.845562	0.535099	0.242298	0.633943	0.466003
ResVar	568.7633	228.2565	46.70112	322.0113	173.005

Table 70. Results from "RAW" case of study (average score's regression tasks)

CLUST	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	23.56813	16.18888	14.76907	17.35532	15.54528
TsRMSE	23.51779	14.98794	6.835109	15.40797	12.10007
TrRRSE	0.833116	0.57201	0.521938	0.613292	0.5494
TsRRSE	0.831272	0.529762	0.241592	0.544607	0.427664
ResVar	553.4142	224.636	46.72348	235.8992	147.378

Table 71. Results from "CLUST" case of study (average score's regression tasks)

NUM	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	23.65691	16.32132	14.79216	18.98315	15.19694
TsRMSE	23.60006	15.33732	6.831363	18.03886	13.67094
TrRRSE	0.83794	0.576842	0.522788	0.670933	0.53713
TsRRSE	0.835825	0.543172	0.241929	0.638952	0.484096
ResVar	557.1626	236.1617	46.6717	323.9812	186.984

Table 72. Results from "NUM" case of study (average score's regression tasks)

IG	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	23.73831	15.98582	14.6976	17.59911	14.94067
TsRMSE	23.58675	16.15733	6.822615	16.063	13.48332
TrRRSE	0.841128	0.566265	0.520859	0.623638	0.529346
TsRRSE	0.835624	0.572483	0.241705	0.56904	0.477644
ResVar	556.5604	261.7592	46.55281	257.0861	182.7594

Table 73. Results from "IG" case of study (average score's regression tasks)

IGR	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	23.71684	16.04978	14.70272	17.22779	14.76684
TsRMSE	23.6016	15.73912	6.813029	15.39009	12.94326
TrRRSE	0.840543	0.571336	0.523445	0.613317	0.525611
TsRRSE	0.836475	0.557778	0.241474	0.54547	0.458723
ResVar	557.2821	248.1866	46.42093	235.3627	168.4114

Table 74. Results from "IGR" case of study (average score's regression tasks)

COLL1	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	24.0506	16.12997	14.73486	20.76894	14.89863
TsRMSE	24.05533	15.2014	6.822787	19.57671	13.96756
TrRRSE	0.850313	0.569687	0.520426	0.733436	0.526251
TsRRSE	0.850349	0.537355	0.241191	0.6923	0.493777
ResVar	578.745	231.2571	46.55616	383.1529	195.1444

Table 75. Results from "COLL1" case of study (average score's regression tasks)

COLL2	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	24.02551	16.13682	14.81231	18.72853	15.14396
TsRMSE	23.87764	15.22956	6.834561	17.40573	12.61187
TrRRSE	0.8499	0.572389	0.525429	0.664455	0.537179
TsRRSE	0.844664	0.538721	0.241769	0.615814	0.446135
ResVar	570.2841	232.1579	46.71757	303.3197	159.0795

table 76. Results from "COLL2" case of study (average score's regression tasks)

COLL1+CLUST	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	23.94992	15.97412	14.64031	18.16664	15.79761
TsRMSE	23.90614	15.59901	6.792535	16.74532	13.05028
TrRRSE	0.850137	0.567492	0.520145	0.645281	0.561182
TsRRSE	0.848539	0.553632	0.241095	0.594266	0.46315
ResVar	571.5944	243.7222	46.14201	280.9955	171.2699

Table 77. Results from "COLL1+CLUST" case of study (average score's regression tasks)

COLL1+NUM	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	24.02206	16.18268	14.73921	17.77673	15.03583
TsRMSE	23.91993	15.19884	6.829107	16.36313	13.79387
TrRRSE	0.850548	0.574683	0.523562	0.631543	0.53426
TsRRSE	0.846902	0.538101	0.241792	0.579266	0.488287
ResVar	572.2138	231.2536	46.64023	266.6838	190.9696

Table 78. Results from "COLL1+NUM" case of study (average score's regression tasks)

COLL2+CLUST	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	24.03333	16.14409	14.74346	18.28124	15.35645
TsRMSE	23.9364	15.53467	6.809969	16.72457	11.8102
TrRRSE	0.8509	0.572893	0.523313	0.648755	0.545194
TsRRSE	0.847479	0.549985	0.241118	0.591959	0.41795
ResVar	573.0154	241.9965	46.38099	280.0666	141.9508

Table 79. Results from "COLL2+CLUST" case of study (average score's regression tasks)

COLL2+NUM	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	23.9122	16.13212	14.7568	17.35224	14.90043
TsRMSE	23.8328	15.45049	6.824363	15.42626	13.17566
TrRRSE	0.847295	0.569492	0.52102	0.612587	0.526018
TsRRSE	0.844434	0.547462	0.2418	0.54658	0.466797
ResVar	568.0657	239.1192	46.5757	236.2273	174.1337

Table 80. Results from “COLL2+NUM” case of study (average score’s regression tasks)

IG+COLL	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	24.02231	16.22569	14.73866	20.80041	14.877
TsRMSE	24.07283	15.11801	6.794641	19.88388	13.1854
TrRRSE	0.848248	0.572984	0.520511	0.734941	0.525447
TsRRSE	0.849895	0.533741	0.239885	0.70203	0.465508
ResVar	579.5551	228.9376	46.17038	393.7723	174.0651

Table 81. Results from “IG+COLL” case of study (average score’s regression tasks)

IGR+COLL	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	23.92678	16.26512	14.76017	17.32027	15.44857
TsRMSE	23.73234	15.95324	6.843258	15.44345	13.62398
TrRRSE	0.848715	0.577455	0.523972	0.615053	0.548408
TsRRSE	0.841742	0.565716	0.242716	0.547747	0.483253
ResVar	563.2917	255.5635	46.8344	236.9527	185.822

Table 82. Results from “IGR+COLL” case of study (average score’s regression tasks)

6.2.5.1 Summary and discussion

The same algorithm-ranking (based on performance) propose for the mean interaction scenario applies for this case:

1. Random Forest
2. Gradient Boosting Machine
3. Neural Network
4. Support Vector Machine
5. Penalized regression

The reasonings applied to infer the possible causes for the performance difference between algorithms are the same as for the mean interaction scenario, and thus, in account of simplicity and to avoid redundancy, are not exposed again in this section.

The best-results summary is presented below:

	RAW	CLUST	NUM	IGR	COLL1
Random Forest					
TrRMSE	14.73977	14.76907	14.79216	14.70272	14.73486
TsRMSE	6.833463	6.835109	6.831363	6.813029	6.822787
TrRRSE	0.521987	0.521938	0.522788	0.523445	0.520426
TsRRSE	0.242298	0.241592	0.241929	0.241474	0.241191
ResVar	46.70112	46.72348	46.6717	46.42093	46.55616
	COLL2	COLL1+CLUST	COLL1+NUM	COLL2+CLUST	COLL2+NUM
Random Forest					
TrRMSE	14.81231	14.64031	14.73921	14.74346	14.7568
TsRMSE	6.834561	6.792535	6.829107	6.809969	6.824363
TrRRSE	0.525429	0.520145	0.523562	0.523313	0.52102
TsRRSE	0.241769	0.241095	0.241792	0.241118	0.2418
ResVar	46.71757	46.14201	46.64023	46.38099	46.5757
	IG	IG+COLL	IGR+COLL		
Random Forest					
TrRMSE	14.6976	14.73866	14.76017		
TsRMSE	6.822615	6.794641	6.843258		
TrRRSE	0.520859	0.520511	0.523972		
TsRRSE	0.241705	0.239885	0.242716		
ResVar	46.55281	46.17038	46.8344		

Table 83. Summary of best regression models and algorithms (average score)

For this scenario, all the elaborated models performed better on a Random Forest algorithm.

It can be observed that, although two clear performance peaks exist for “C1+CLUST” and “IG+COLL” cases, there is no clear difference among the other cases of study, thus not allowing for an objective inference of causality for improvement or impairment of the measures used.

For deciding between the two most favourable cases previously mentioned, a simple “parsimony criteria” has been applied. This is, the model comprising the less attributes is the one selected (although there may exist an argument for the appropriateness of this criteria when applied a-priori, that fact that performance results of each model are present allow for an objective choice based on it, since only the most valuable attributes are preserved).

Consequently, “IG+COLL” (4 attributes against 14 from “C1+CLUST”) model is selected as the most appropriate for conducting regression tasks (by applying Random Forests) on this scenario.

It can be concluded that the raw model, treated with a Random Forest algorithm, is the most suitable for the conduction of regression tasks on this scenario.

6.2.5.1.1 Contrast of our approach's validity

In order to support the approach given to our analytics tasks (described in its correspondent section), results from a regression model accounting uniquely for mean interaction with the VLE platform are provided:

	PEN	NNET	RFOREST	SVM	GBM
TrRMSE	24.2868	19.8919	22.0068	28.3431	21.545
TsRMSE	24.2839	19.8454	13.4646	27.9415	17.7691
TrRRSE	0.8665	0.7099	0.7852	1.0113	0.7691
TsRRSE	0.8664	0.7080	0.4804	0.9969	0.634
ResVar	589.75	393.87	181.3096	706.8425	315.7658

Table 84. Results from an approach based on mean interaction alone

Given the significantly worse performance measures for this case (compared to the ones previously discussed), it can be safely stated that, in order to develop adequately informed Learning Analytics processes, data from different ambits other than mere interaction should be considered.

7 TIME SERIES MODELS

Following the same top-down methodology given to the regression processes performed, time series forecasting has been conducted by taking a generalist approach to its modelling. Consequently, mean values of the observed variable are first assessed for each course as a whole, so that conclusions elicited during this process aid the adaptation of the model to individual issues (forecasting of unique students' indicators).

Although there exists data enough to develop time series for the evolution of the average score, both for each course (mean value) and unique students, its stepped behaviour, with changes taking place only on assessments, being constant in the intervals between them and thus indifferent to any external force acting over time, led to the conclusion that no significant information would be extracted from its treatment. Consequently, interaction during a course (mean value) and that of unique students have been defined as the forecast's subjects.

A detail of the framework developed for this processes' development is following presented:

- Graphical assessment and definition of time series: information relevant to the further development of time series models is extracted from the visual analysis of its plots.

Additionally, any change to the arrangement of the data with analytics purposes is detailed.

- Characteristics of the experiments to be conducted: details referred to considerations with respect to training and testing and measures of performance taken into account.

- Identification and definition of external regressors

- Outliers' effects

As it will be discussed in the following sections, the appearance of outlier in time series may imply a certain modification in its shape.

To account for these events, it is needed to include their possible effects in the model as external regressors. This will be made by confecting indicator variables (widely known as dummy variables, which are composed of 0s and 1s, with the latter appearing to "signal" the outlier's effect) for each outlier having a significant effect.

- Seasonality assessment: the identification of a time series' main seasonal components leads to different treatment possibilities to address its influence:

- Decomposition and adjustment of time series so that seasonality doesn't appear as the main force altering its behaviour, which may improve the performance of certain algorithms.

- Correspondent Fourier series' terms and harmonic regression for modelling seasonality

While treating with daily date (as in our case) the possibilities of encountering multiple seasonality along the time series is high.

Although most algorithms are ready to deal with simple seasonality, some of them cannot consider anything more complex than that, as it is

the case of ARIMA (which apart of being an algorithm itself, some others draw from it for errors' computation).

As a solution, there's the possibility of incorporating Fourier terms to the modelling of time series:

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos \frac{2n\pi}{T}t + b_n \sin \frac{2n\pi}{T}t \right]$$

Figure 86. Fourier series formula

$$x_{1,t} = \sin\left(\frac{2\pi t}{m}\right), x_{2,t} = \cos\left(\frac{2\pi t}{m}\right), x_{3,t} = \sin\left(\frac{4\pi t}{m}\right),$$

$$x_{4,t} = \cos\left(\frac{4\pi t}{m}\right), x_{5,t} = \sin\left(\frac{6\pi t}{m}\right), x_{6,t} = \cos\left(\frac{6\pi t}{m}\right),$$

Figure 87. First terms of a Fourier series

Its periodic behaviour (as a cosine/sine function) makes it ideal to incorporate cyclic seasonality as an external regressor of a time series' model.

It is important to point out that, although being regressors added to the model, the fact that they are necessary for certain models' appropriate performance will make its unique inclusion to not count as a case study referred to external regressors. This is, the base case for algorithms which need this solution to consider multiple seasonality imply this regressors.

A list of cases of study is proposed as a summary of how each of the previously detailed processes will be studied:

Cases of study	
Original data	Original data with regressors
Seasonally adjusted data	Seasonally adjusted data with regressors

Table 85. Forecasting tasks' cases of study

It is important to remark the fact that not every algorithm on which forecasting has been conducted considers the possibility of adding external regressors to their models. Coherently with this, the list of study cases presented is may be reduced to the first column's cases for the cases in which no external regressors can be passed to the model.

7.1 Graphical assessment and definition of time series

Following, time series plots for each of the 22 courses present in our database are collected and discussed. Visual assessment of these graphs will act as a source of indications for the development of further analytics tasks (blue dashed lines represent assessment dates).

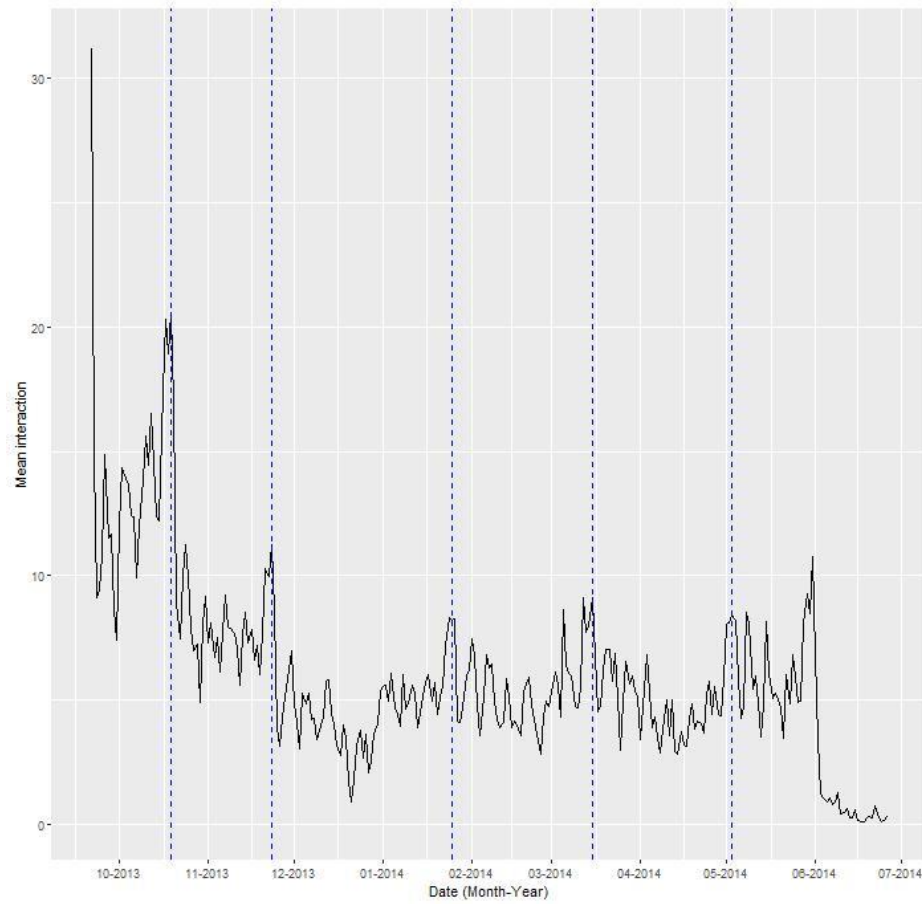


Figure 88. Time series plot of course “AAA-2013J”

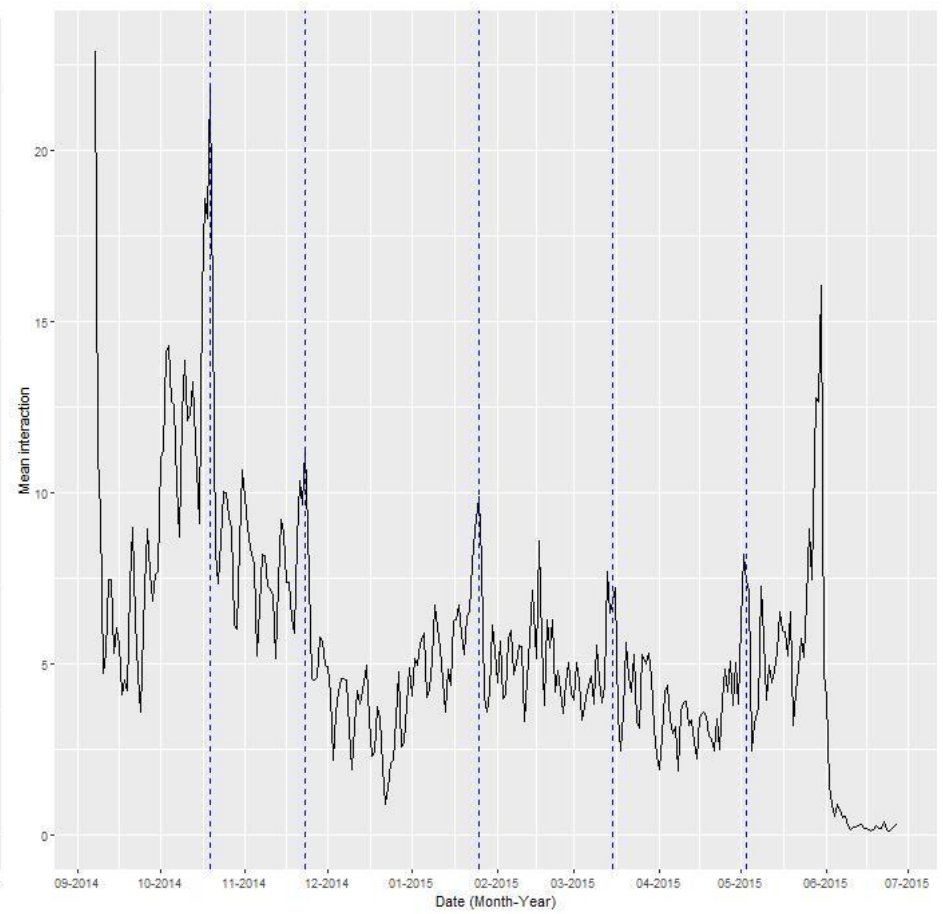


Figure 89. Time series plot of course “AAA-2014J”

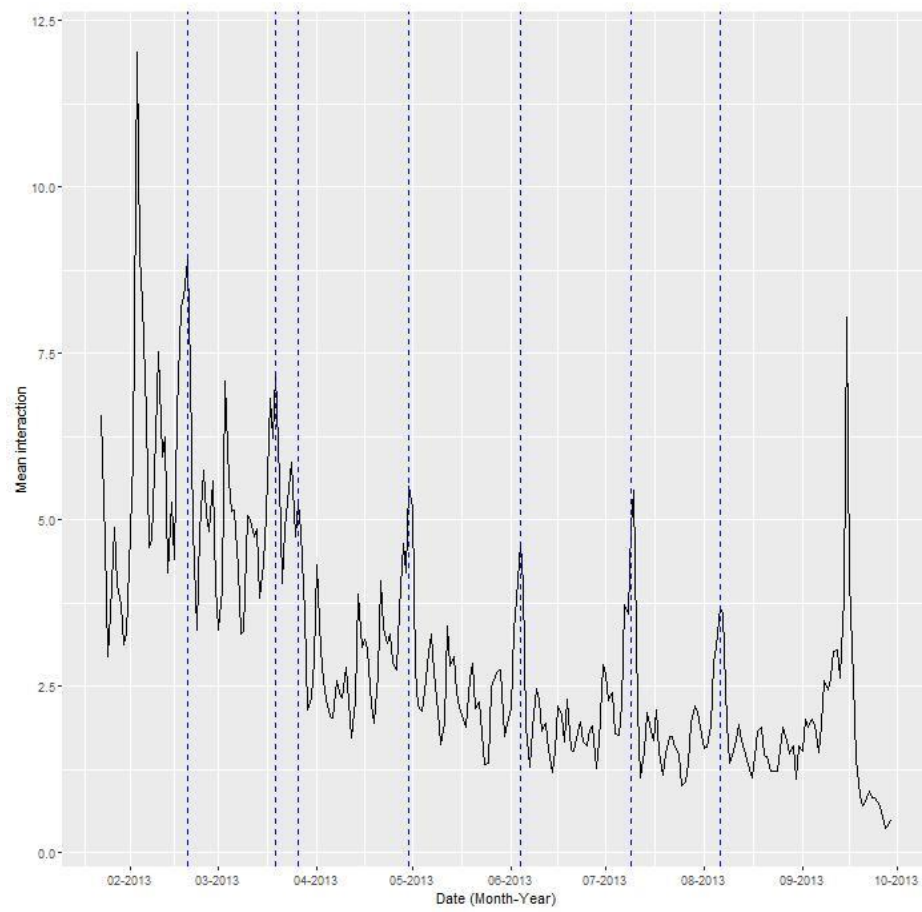


Figure 90. Time series plot of course “BBB-2013B”

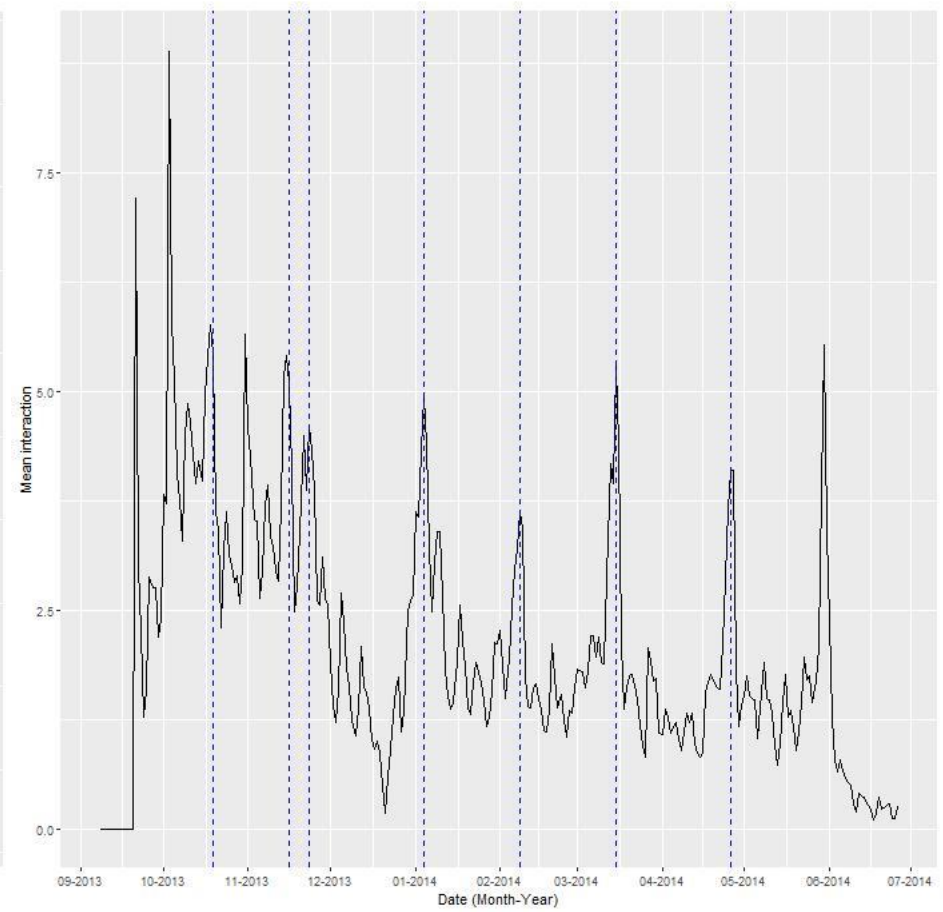


Figure 91. Time series plot of course “BBB-2013J”

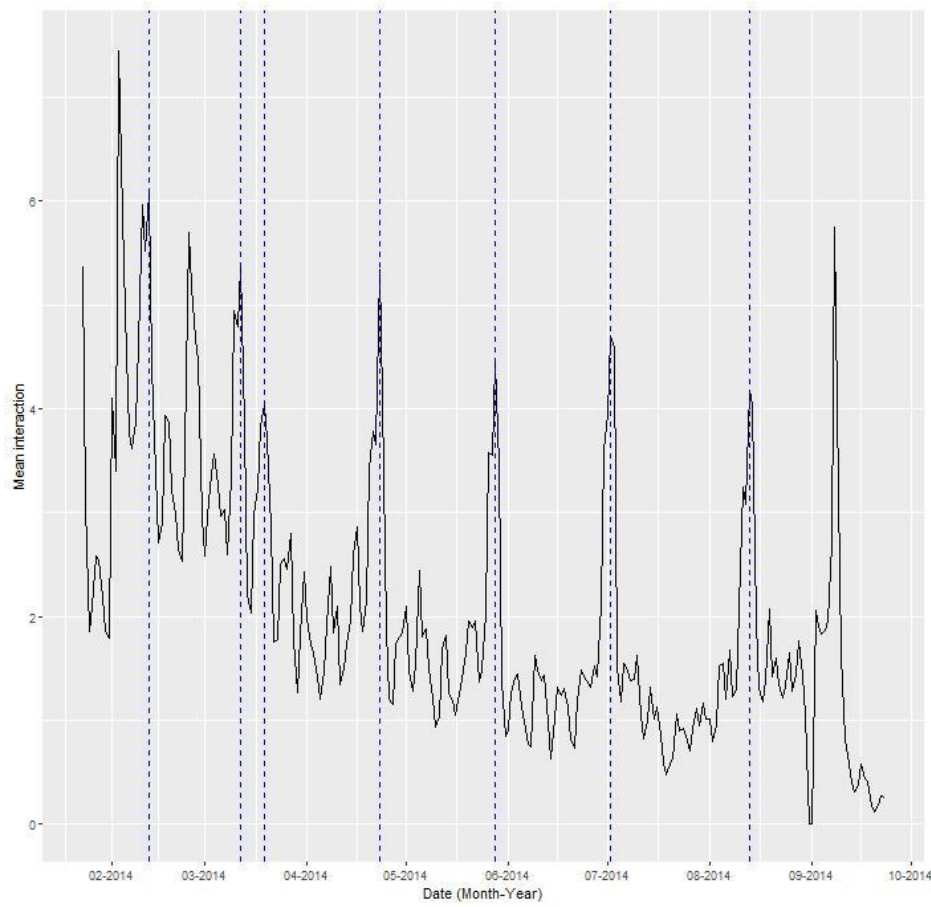


Figure 92. Time series plot of course “BBB-2014B”

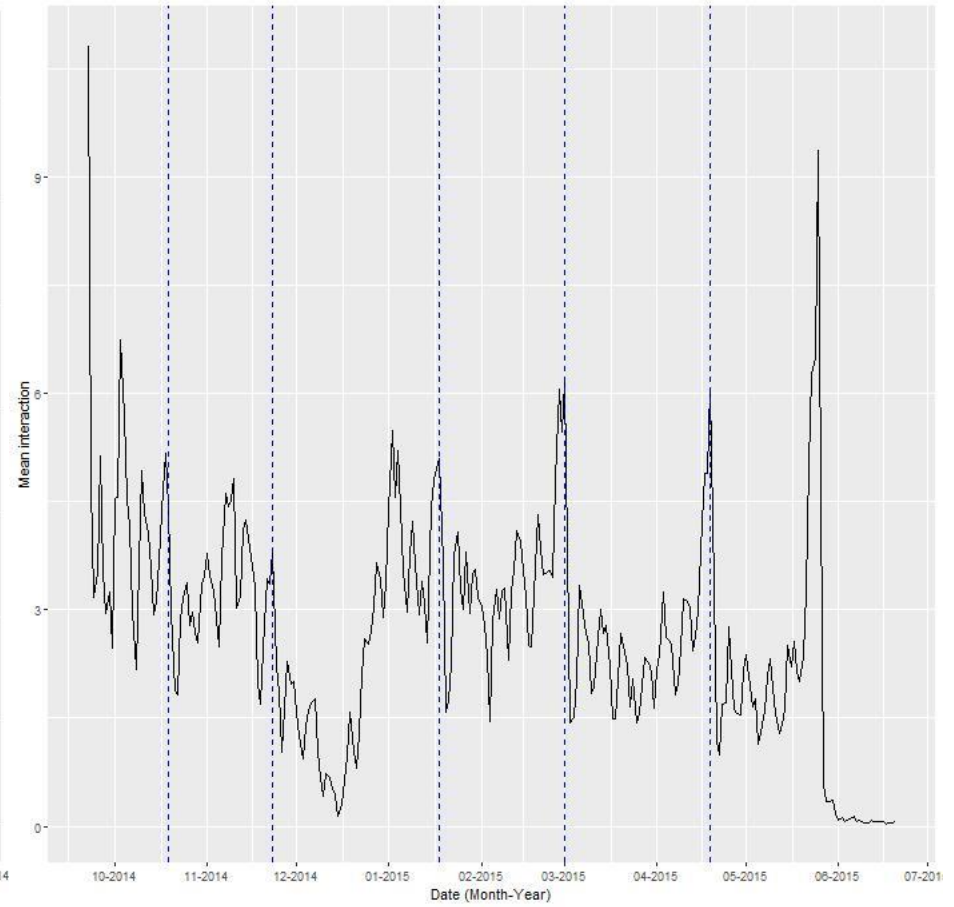


Figure 93. Time series plot of course “BBB-2014J”

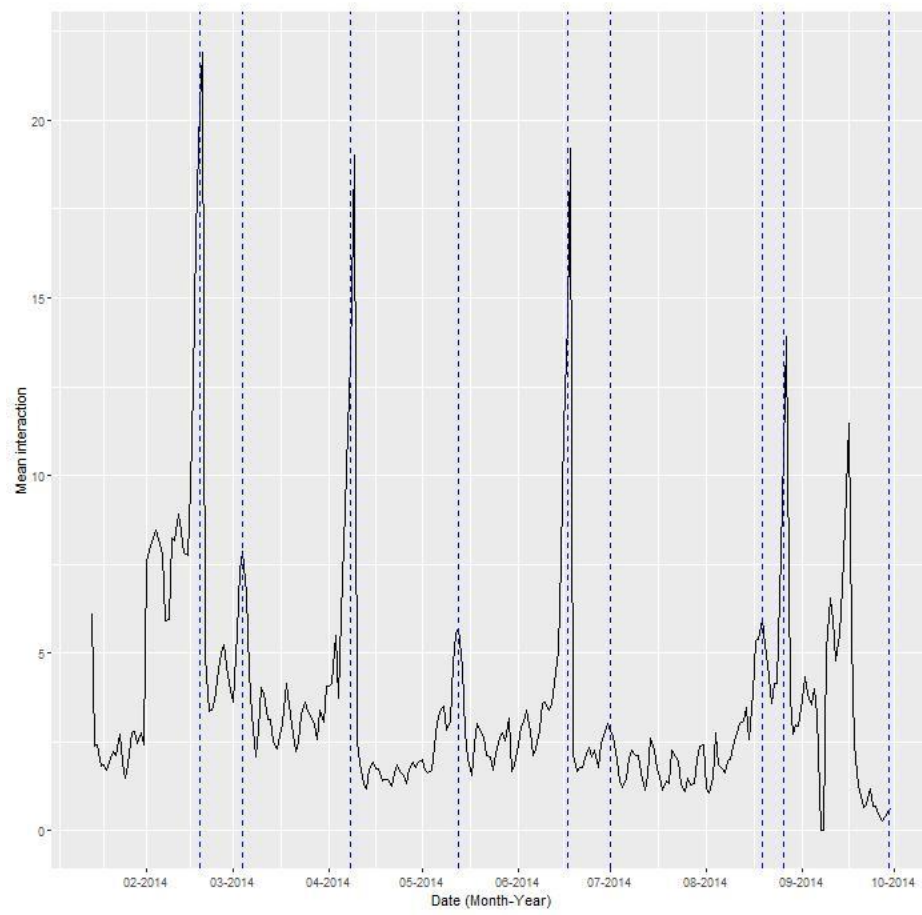


Figure 94. Time series plot of course “CCC-2014B”

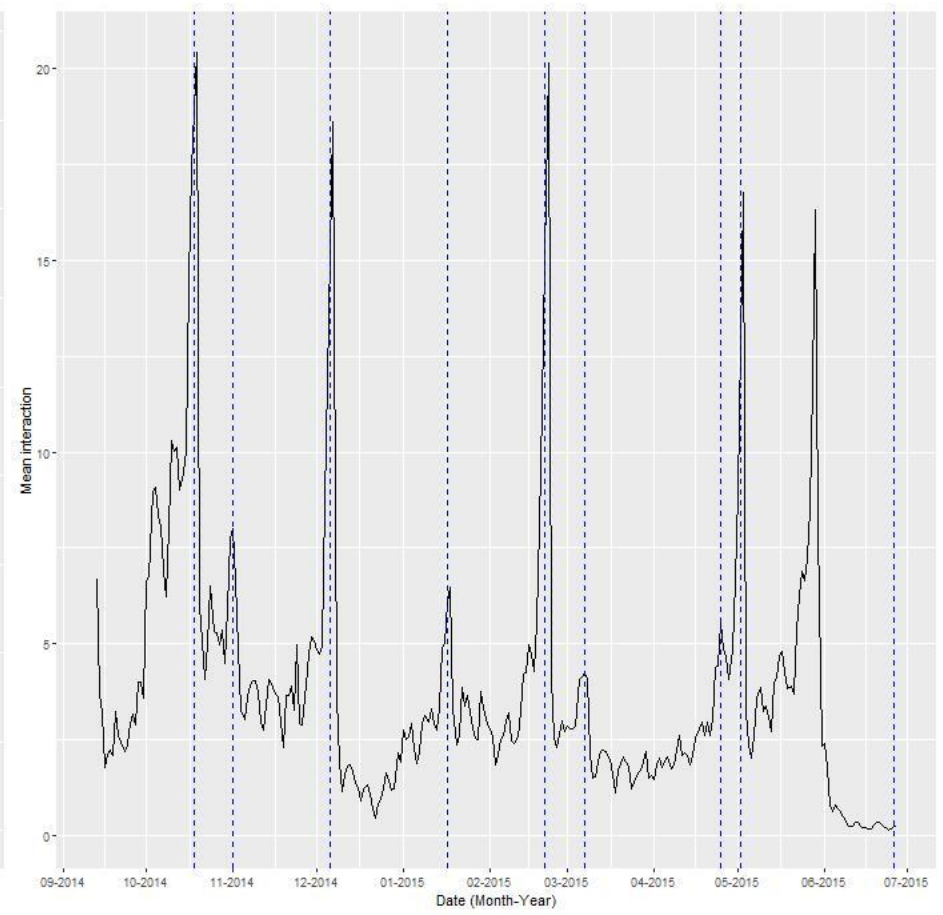


Figure 95. Time series plot of course “CCC-2014J”

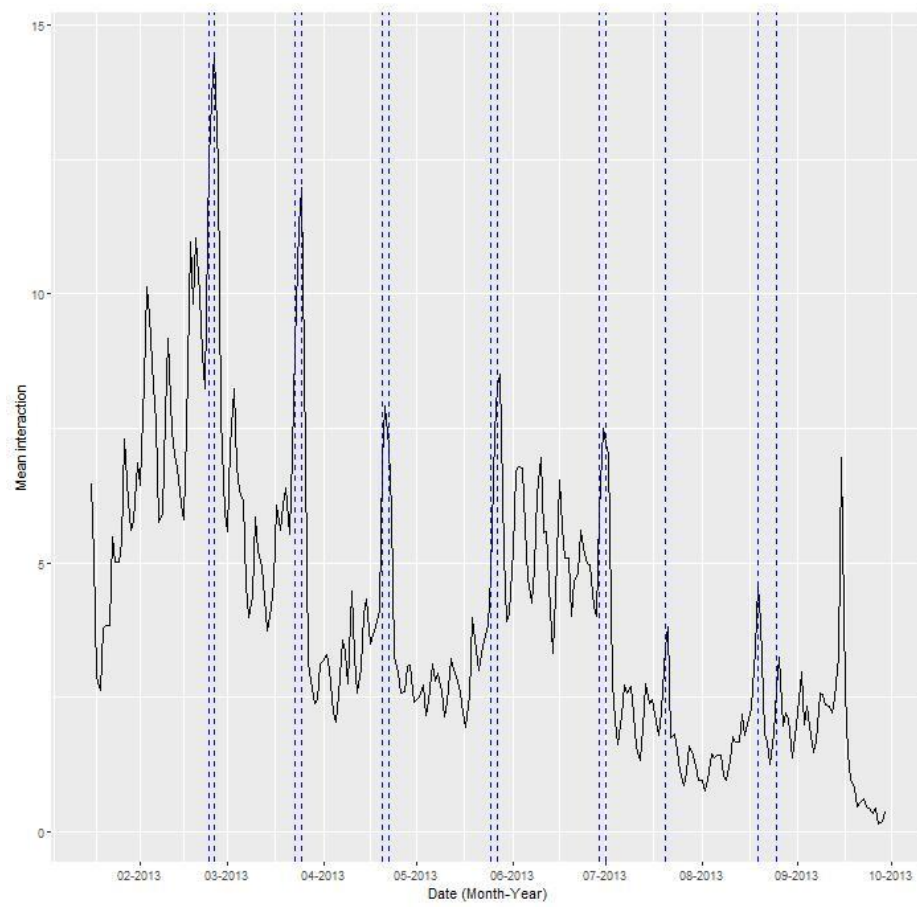


Figure 96. Time series plot of course “DDD-2013B”

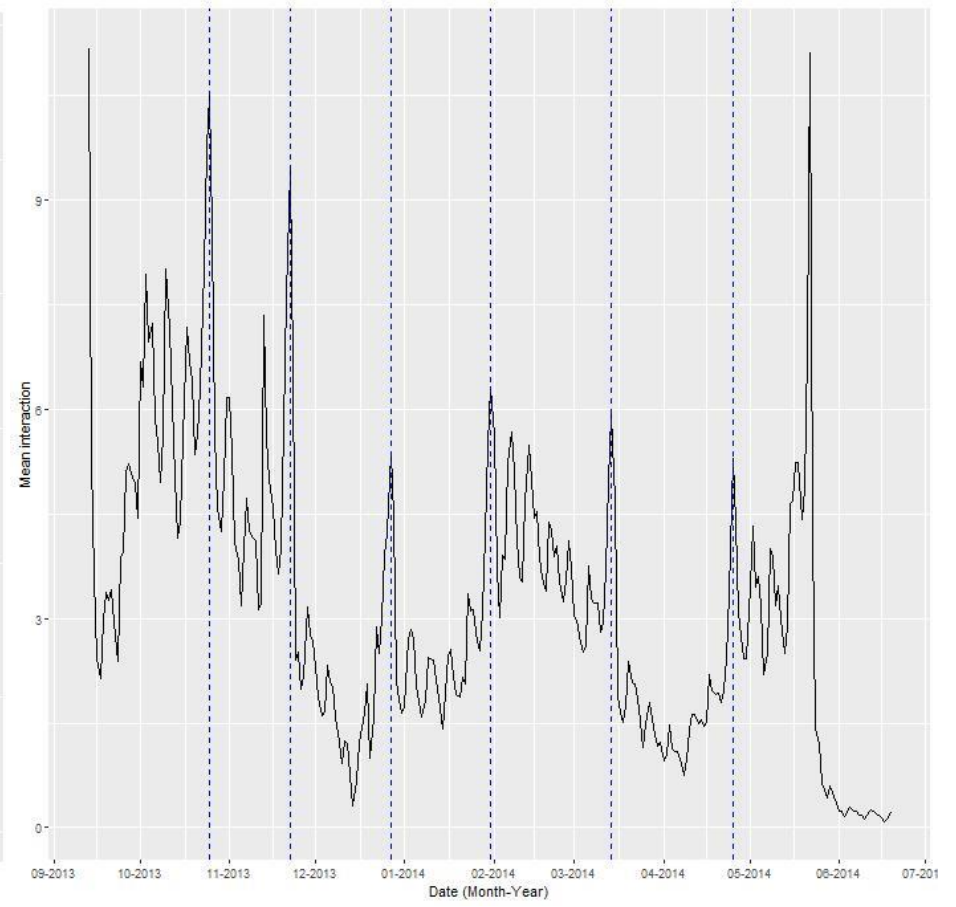


Figure 97. Time series plot of course “DDD-2013J”

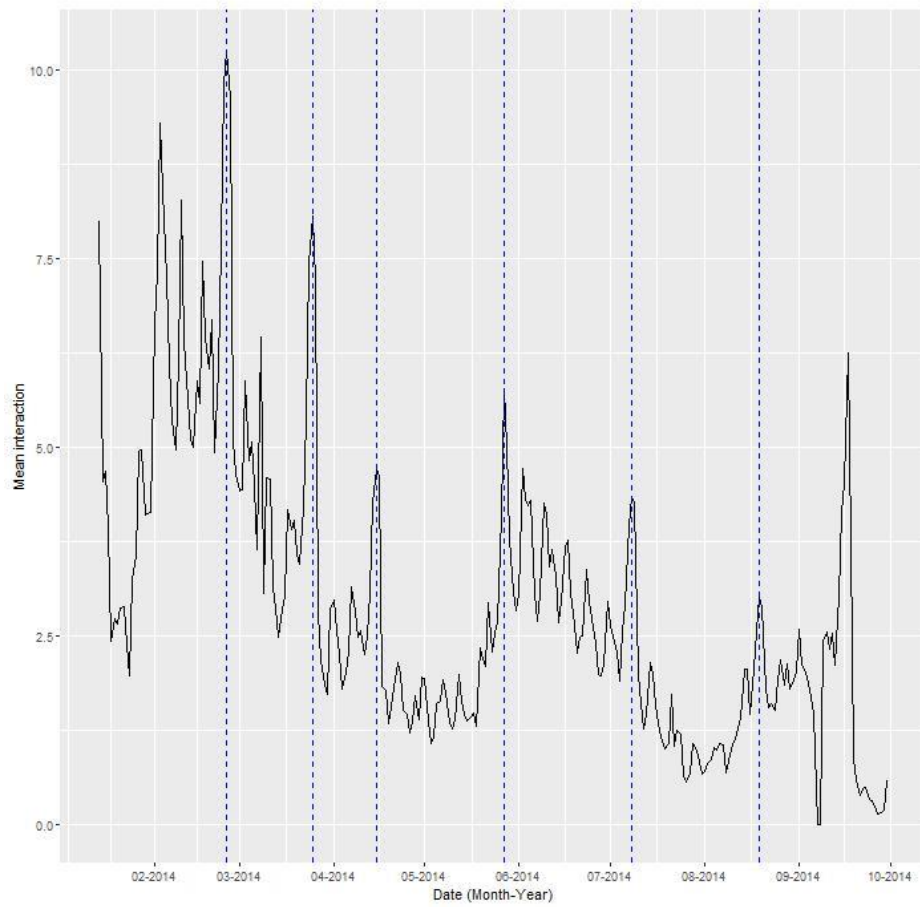


Figure 98. Time series plot of course "DDD-2014B"

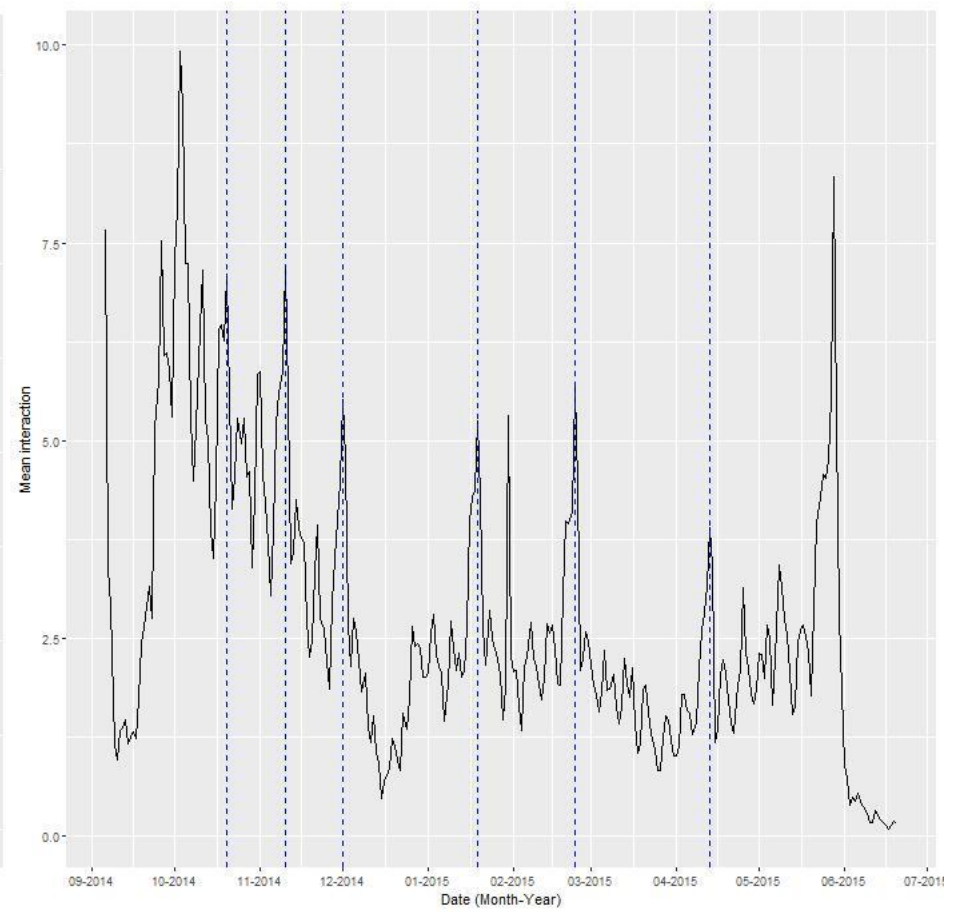


Figure 99. Time series plot of course "DDD-2014J"

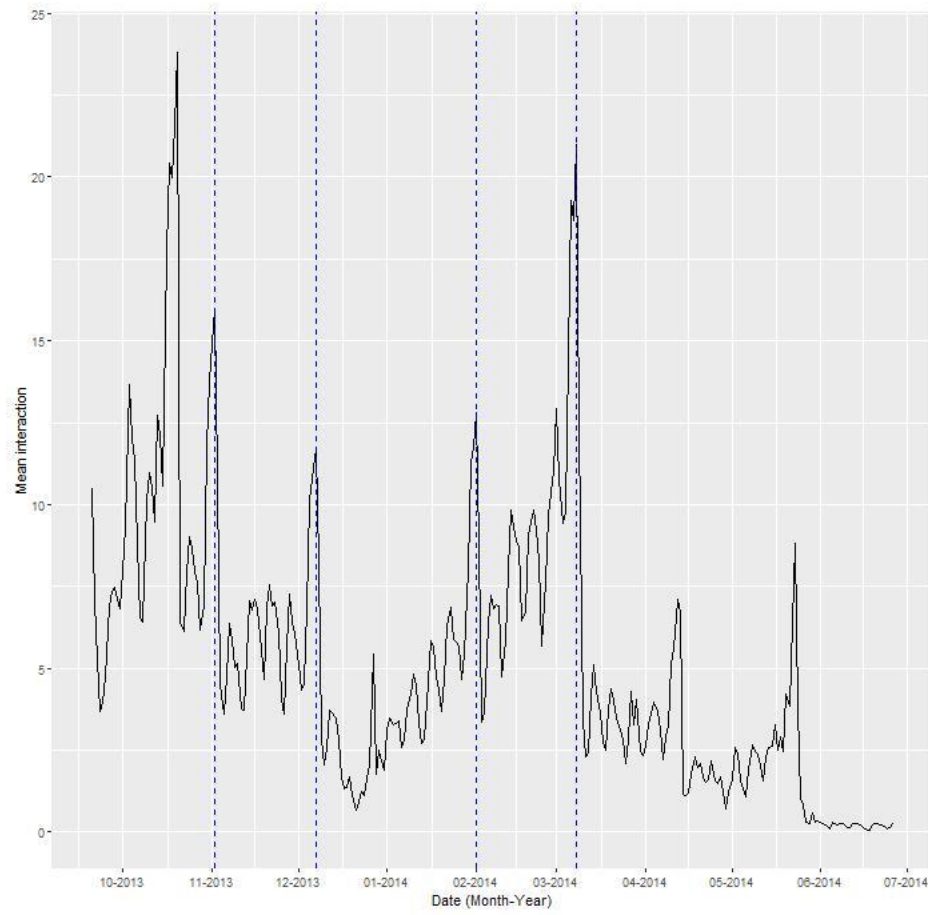


Figure 100. Time series plot of course "EEE-2013J"

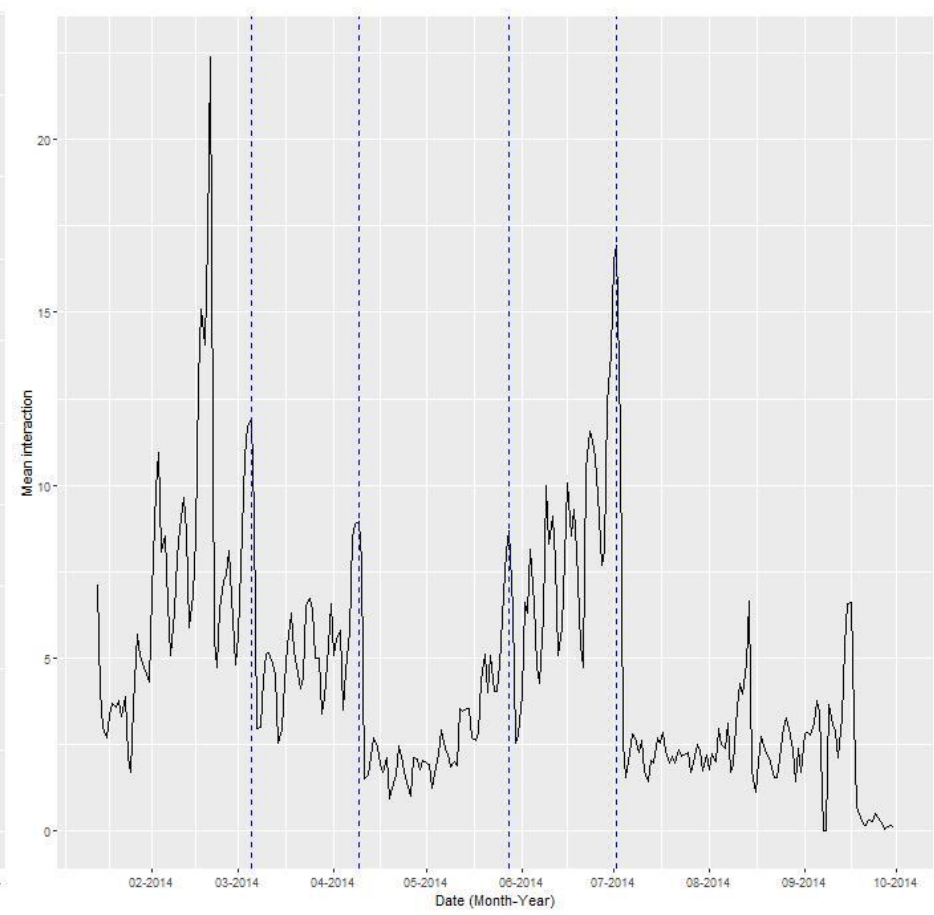


Figure 101. Time series plot of course "EEE-2014B"

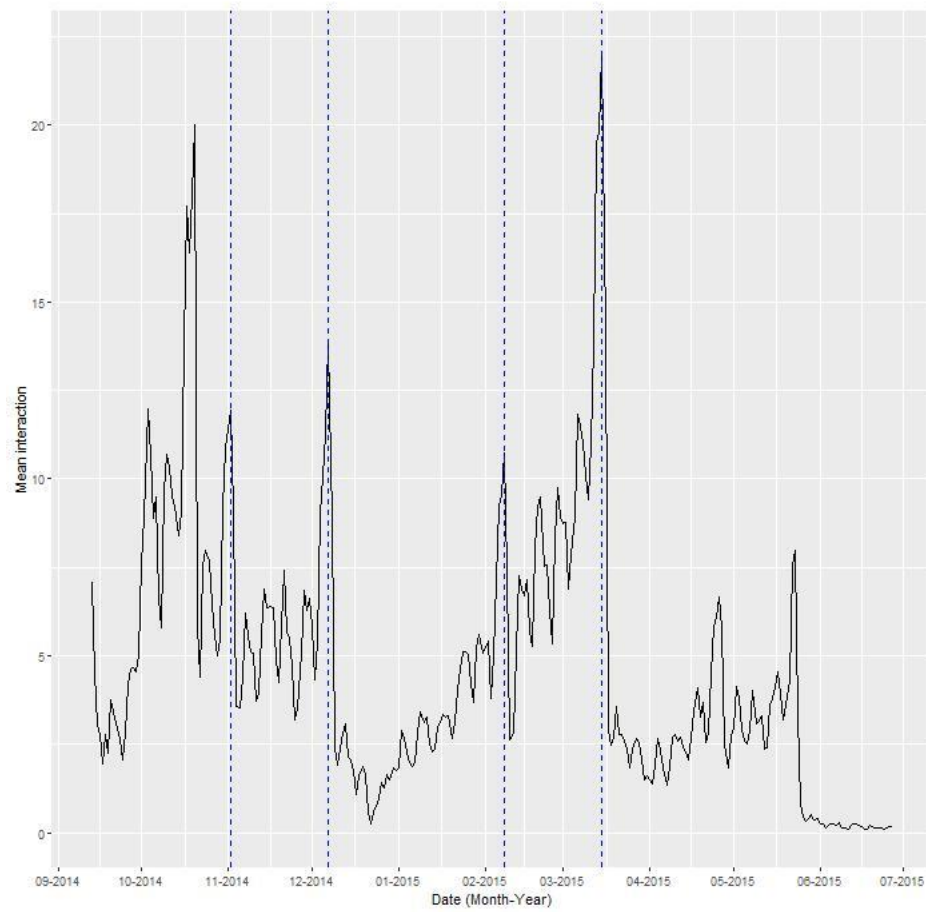


Figure 102. Time series plot of course "EEE-2014J"

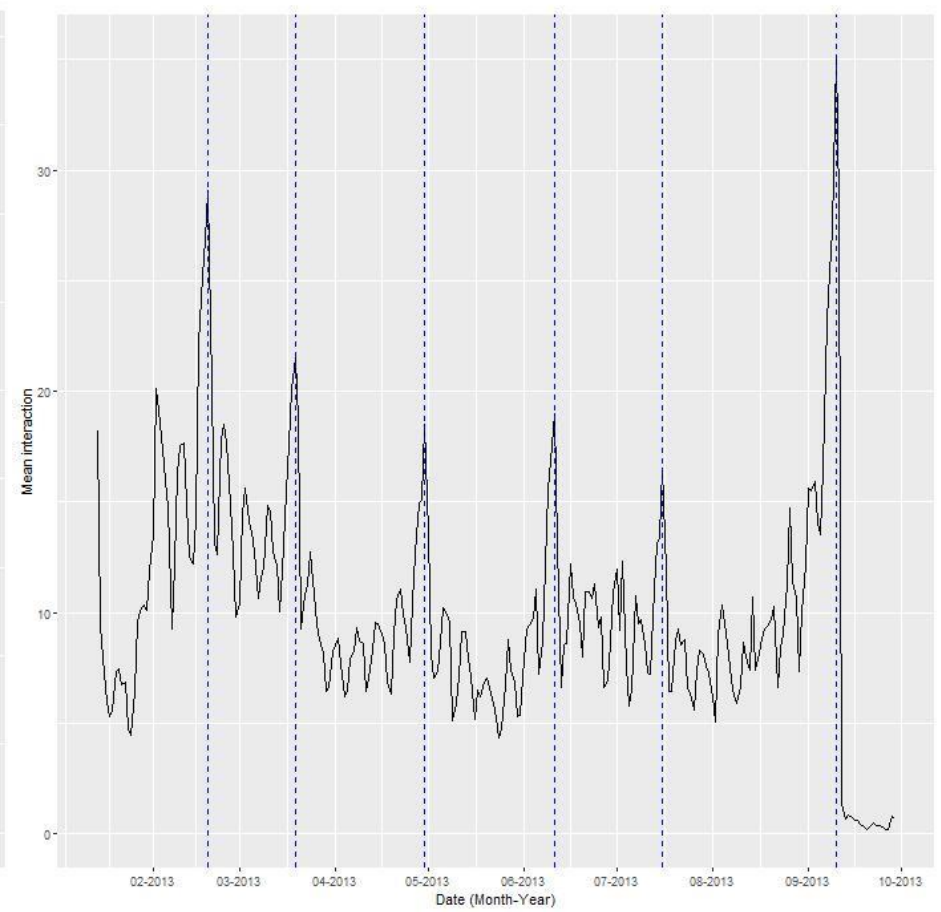


Figure 103. Time series plot of course "FFF-2013B"

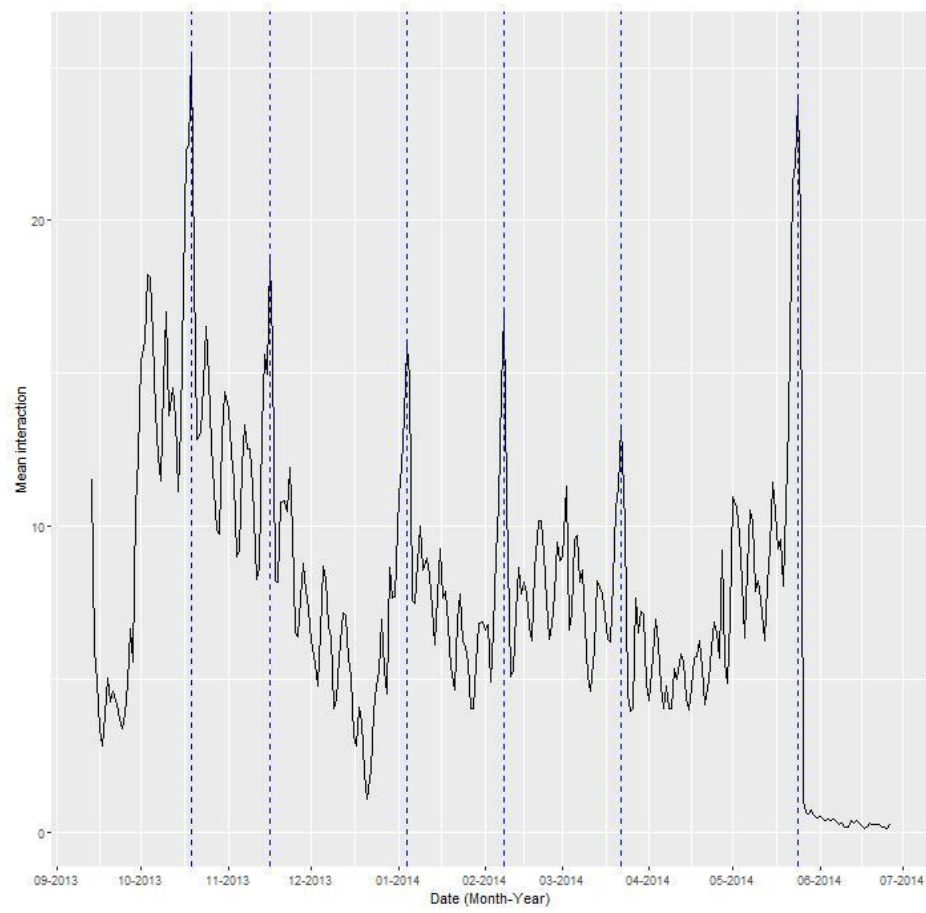


Figure 104. Time series plot of course "FFF-2013J"

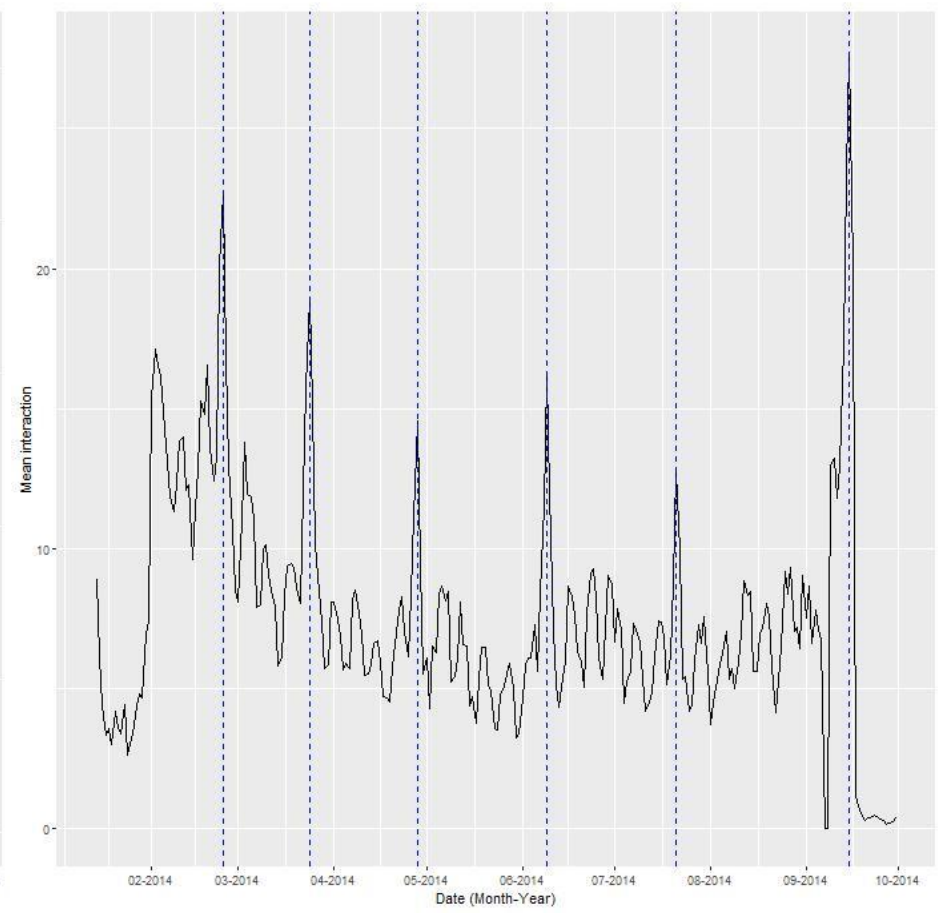


Figure 105. Time series plot of course "FFF-2014B"

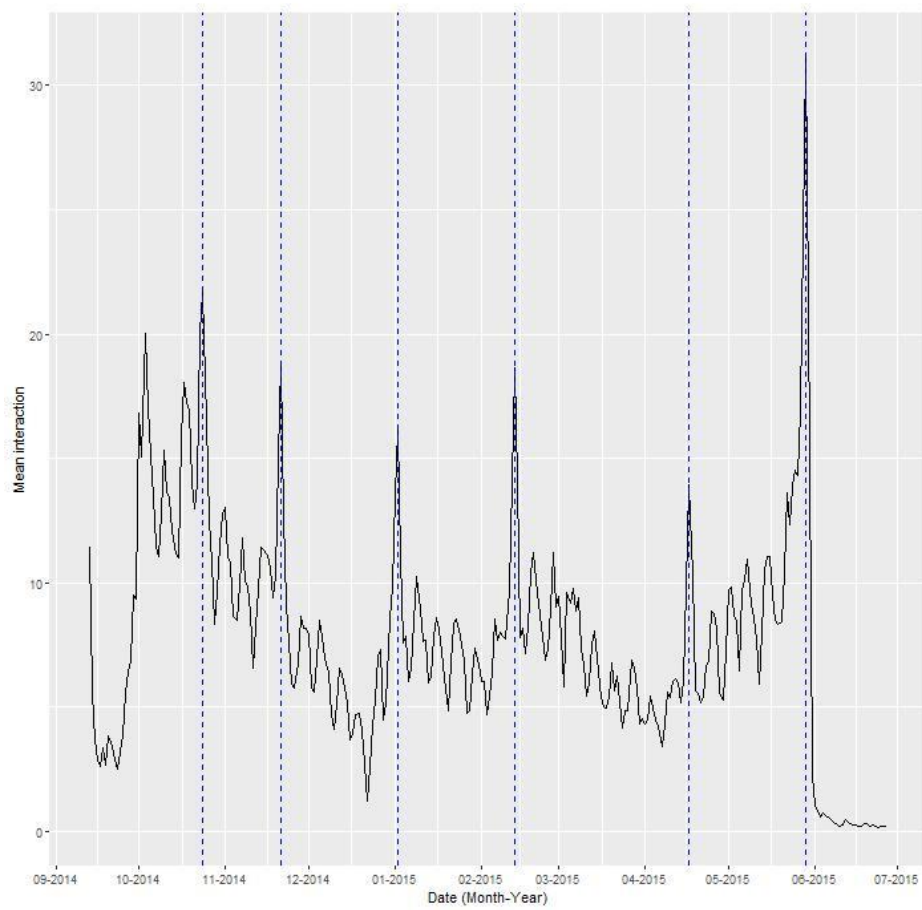


Figure 106. Time series plot of course "FFF-2014J"

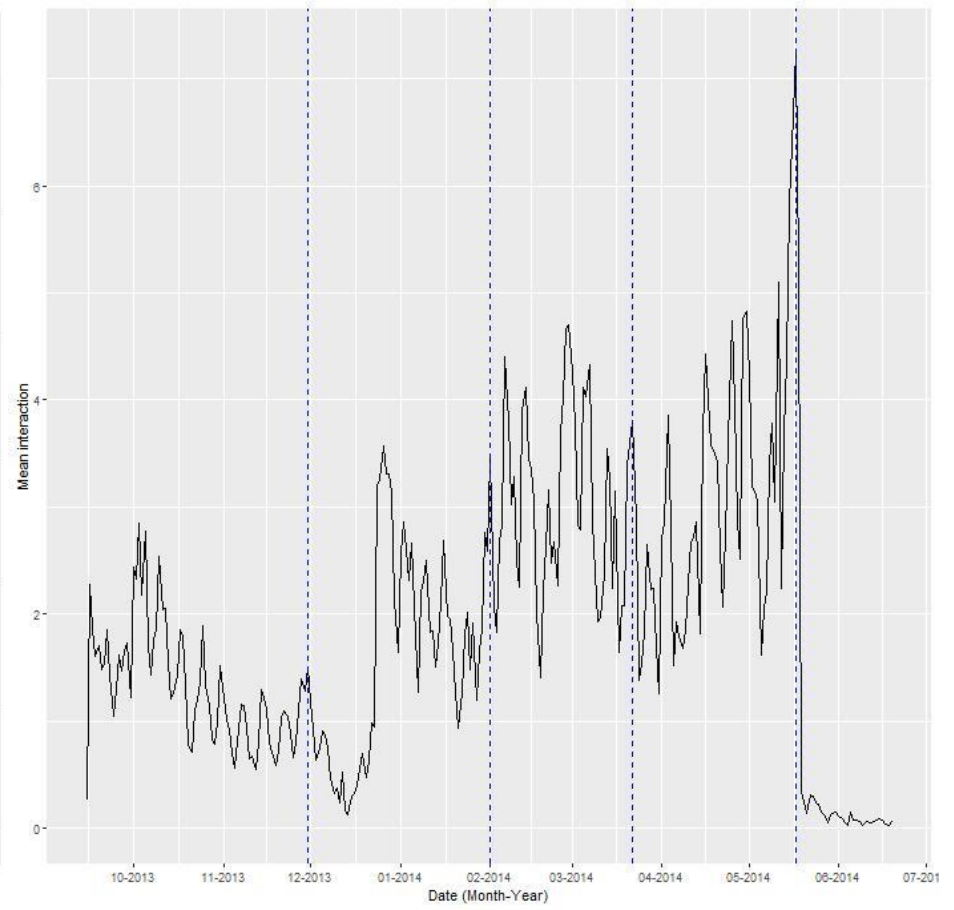


Figure 107. Time series plot of course "GGG-2013J"

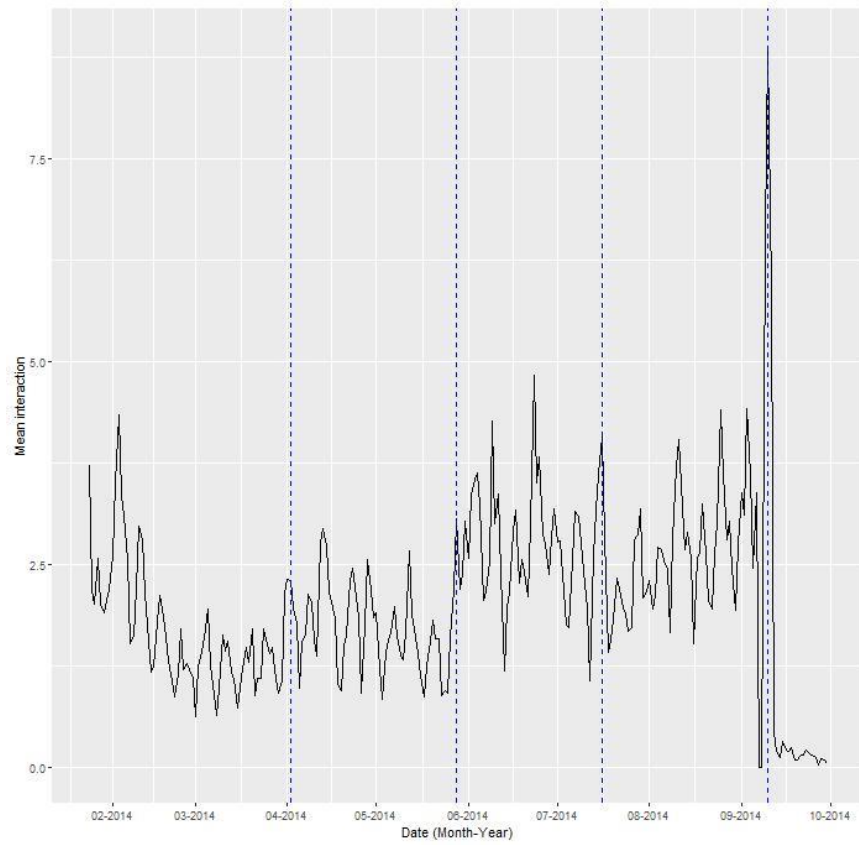


Figure 108. Time series plot of course "GGG-2014B"

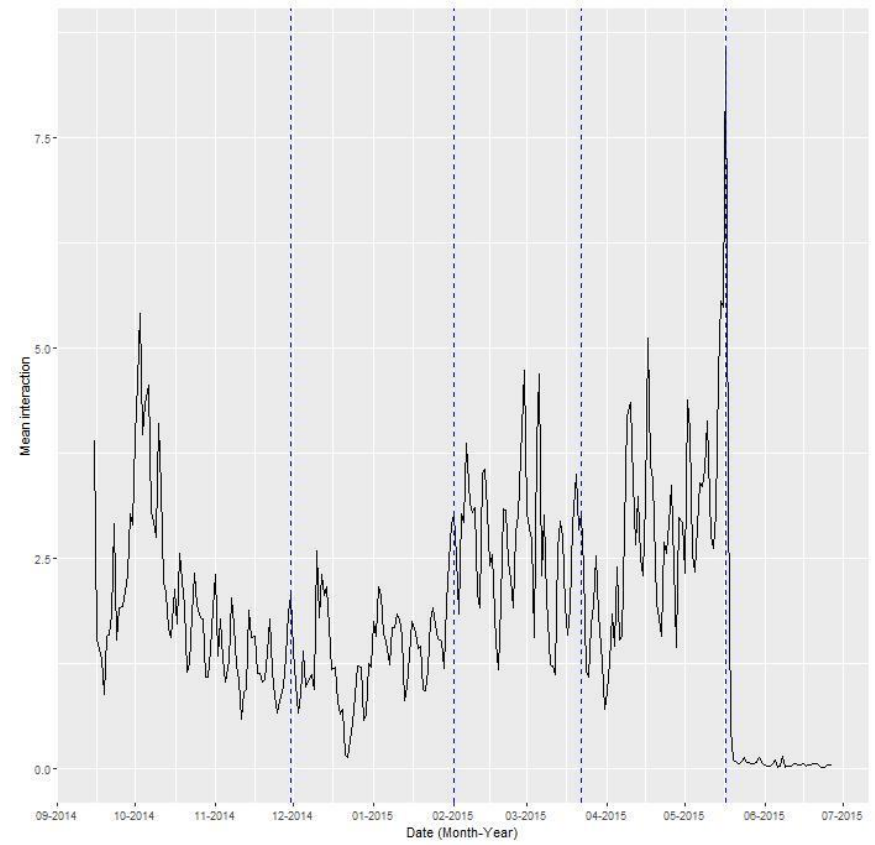


Figure 109. Time series plot of course "GGG-2014J"

At first glance, the occurrence of a repetitive pattern within every graph is noticeable, consisting in the presence of peaks with a rate of appearance lower than a month, thus implying the possibility of existence of a weekly seasonality.

No other seasonality other than weekly can be inferred from the observation of this set of plots. In fact, any other remarkable fluctuation in the data might be better explained by the effect of certain outliers, mostly correspondent to assessment dates.

In this respect, it is worth mentioning the fact that the number and distribution of assessments over time among courses of the same module (first three letters of its identification) is the exact same and, in addition, evolution of the observed variable (mean interaction) follows a significantly similar pattern among these “module-groups”.

The assessment of the influence of assessment dates in the observed variable has been aimed towards an Intervention Analysis-based approach (i.e. the analysis of the fluctuation of a variable's behaviour over time caused by external forces). Accordingly, a process of identification of outliers and evaluation of their effect on the interaction time series previously presented has been conducted (based on Chen and Liu's time series outlier detection, [47], and its R-language implementation in the “tsoutliers” package, [48]).

Following the procedures shown in ([47]), an ARIMA model with automatized parameter setting (“auto.arima” function from “forecast” package) was set for this process' conduction. Parameter optimization was enforced to search through differenced models $(p,1,q)$, in accordance to the results obtained from operating a KPSS test on each courses' time series (presence of unit roots wasn't discarded and, thus, neither non-stationarity was).

Before presenting the results, it is important to identify the different types of outliers considered in the reviewed literature ([49]):

- Innovational outliers: consist of an initial impact with persistent or increasing effect over time.
- Additive outliers: a considerably small or large value occurring for a single observation, with subsequent observations unaffected.
- Level shift outliers: characterized by the translation of its subsequent observations to a different mean.
- Transient change outliers: remarkably large or small values which effect diminishes over time.
- Seasonal level shifts: equivalent to regular level shift outliers appearing repeatedly at regular intervals.

Taking into account the previously mentioned correspondence between most of the observation's outliers with assessment dates and the nature of these events which, as can be observed from the graphs, generally imply a steady increase in interaction towards the reaching of a peak at the assessment date (the potential outlier), innovational, transient change and seasonal level outliers were discarded from the described evaluation process.

Most assessment dates reflect an interaction maximum that either comes back to normal levels (additive outlier) or diminishes to a new one which persists over time (level

shift outlier) almost immediately (with no progressive decrease proper from a transient change), reinforcing the previous statement about the discarded outlier types.

Following, plots depicting the effect of these outliers are presented.

It is important to mention the illustrative purposes of the figures. Volatility of the data used might not be suitable for an automatized process of this kind and, consequently, adjustments have been needed on some of the resulting models (overestimation of regular observations as outliers occurred). No attention should be paid to y-axis' scale since it will not reflect interpretable values (as a result of the mentioned adjustments).

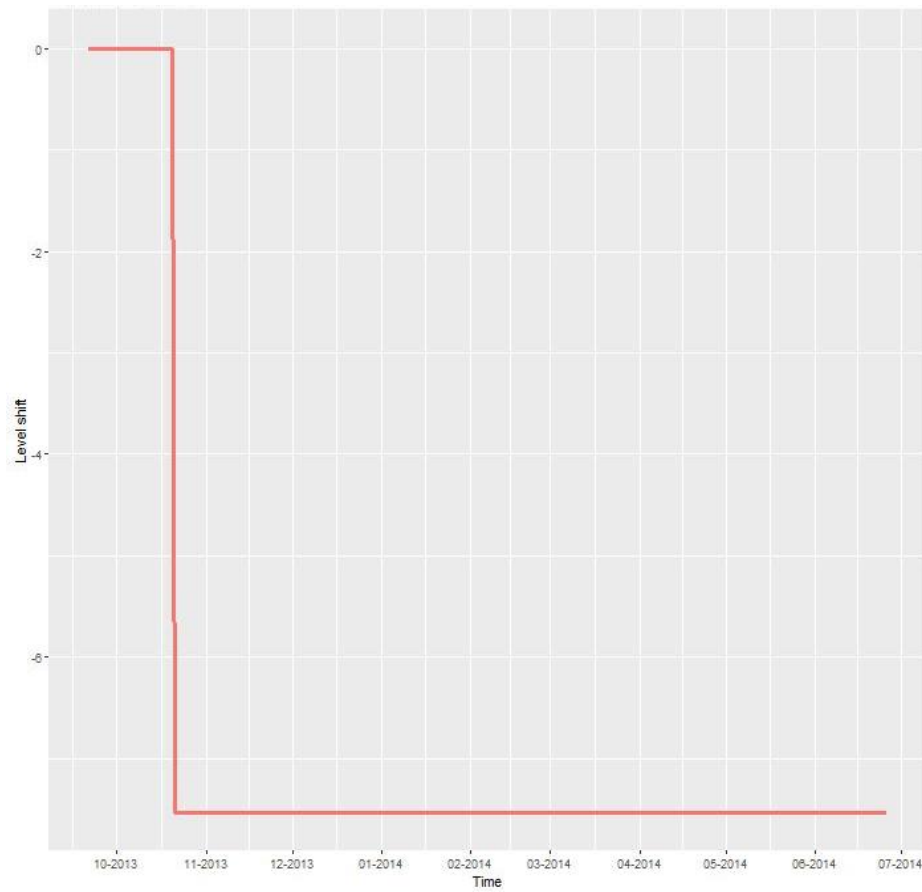


Figure 110. Time series plot of course "AAA-2013J"

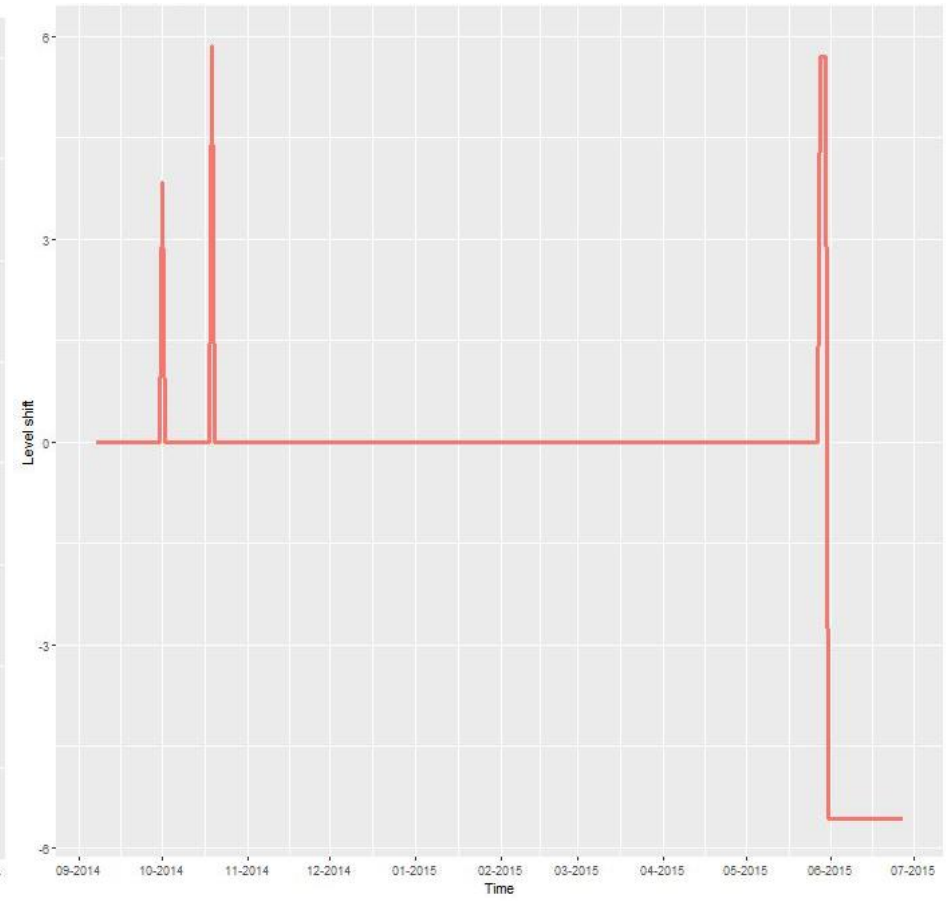


Figure 111. Time series plot of course "AAA-2014J"

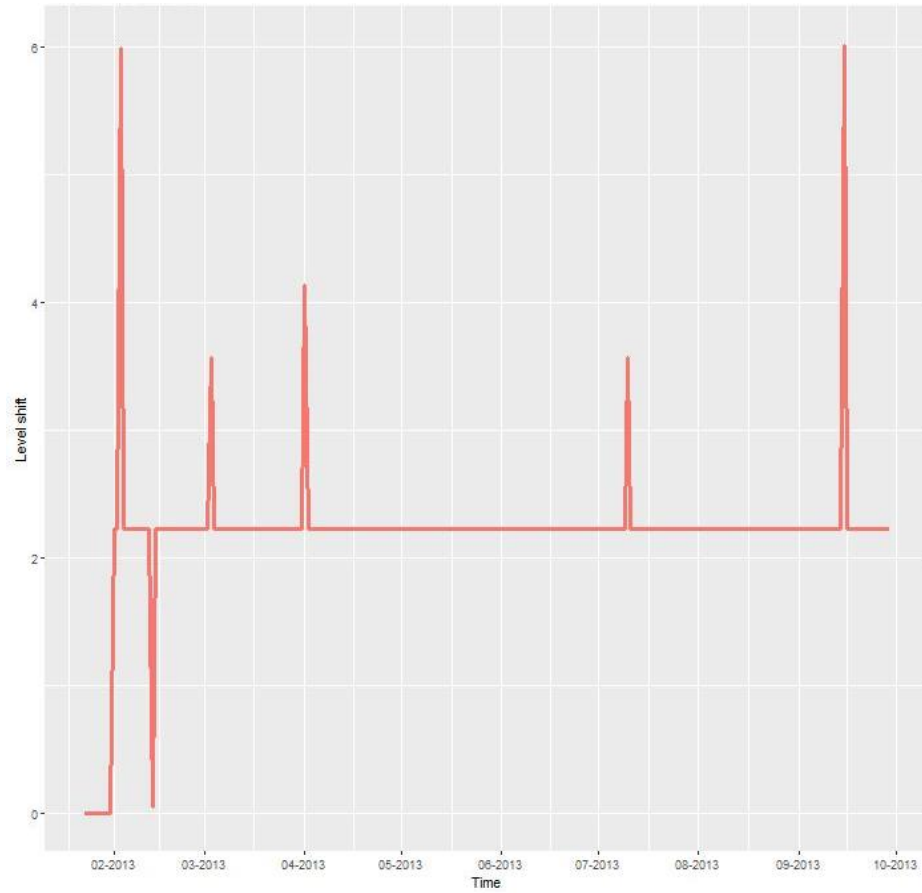


Figure 112. Time series plot of course "BBB-2013B"

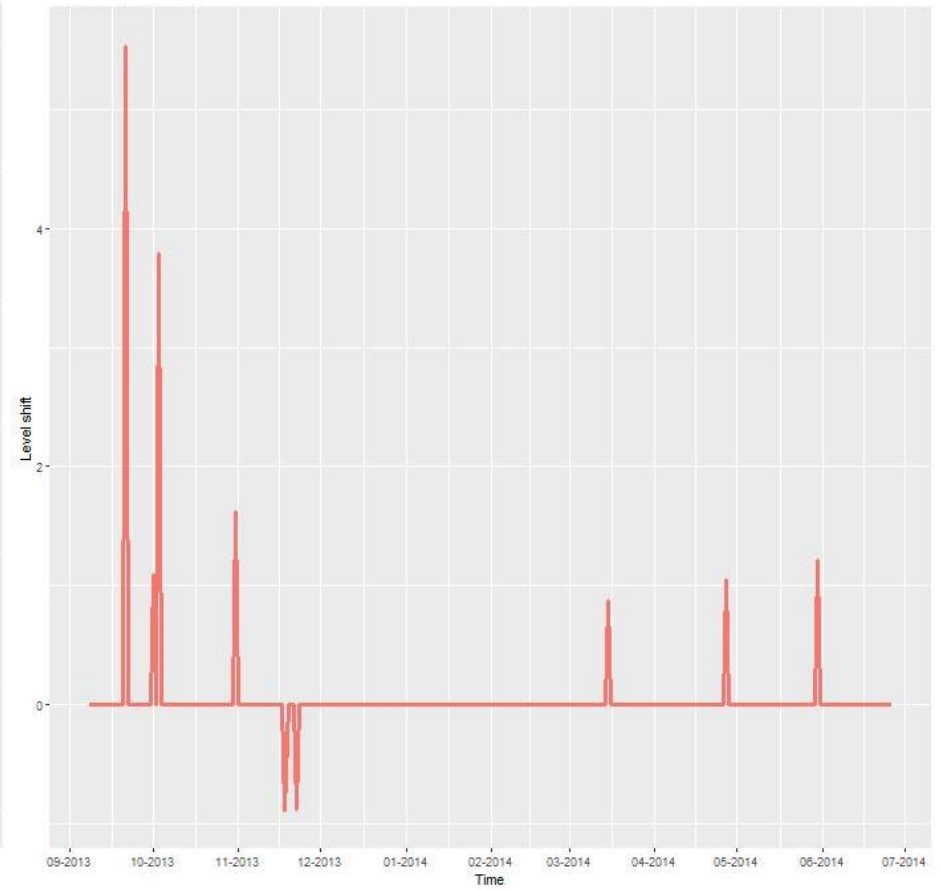


Figure 113. Time series plot of course "BBB-2013J"

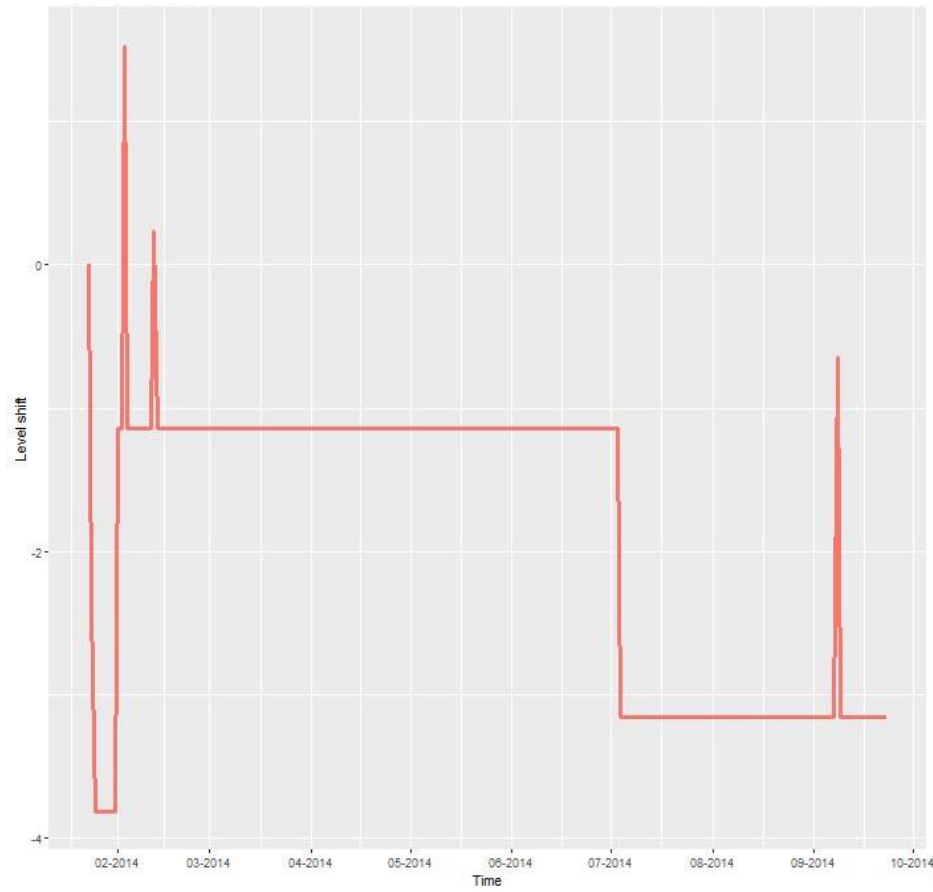


Figure 114. Time series plot of course "BBB-2014B"

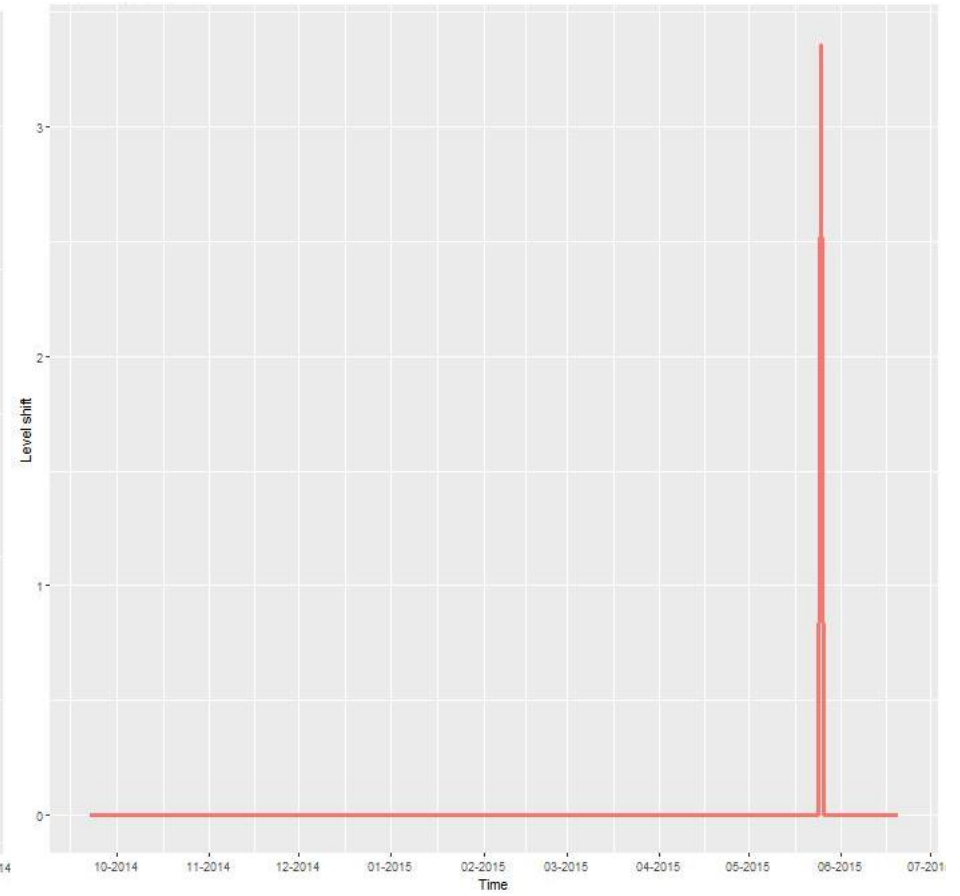


Figure 115. Time series plot of course "BBB-2014J"

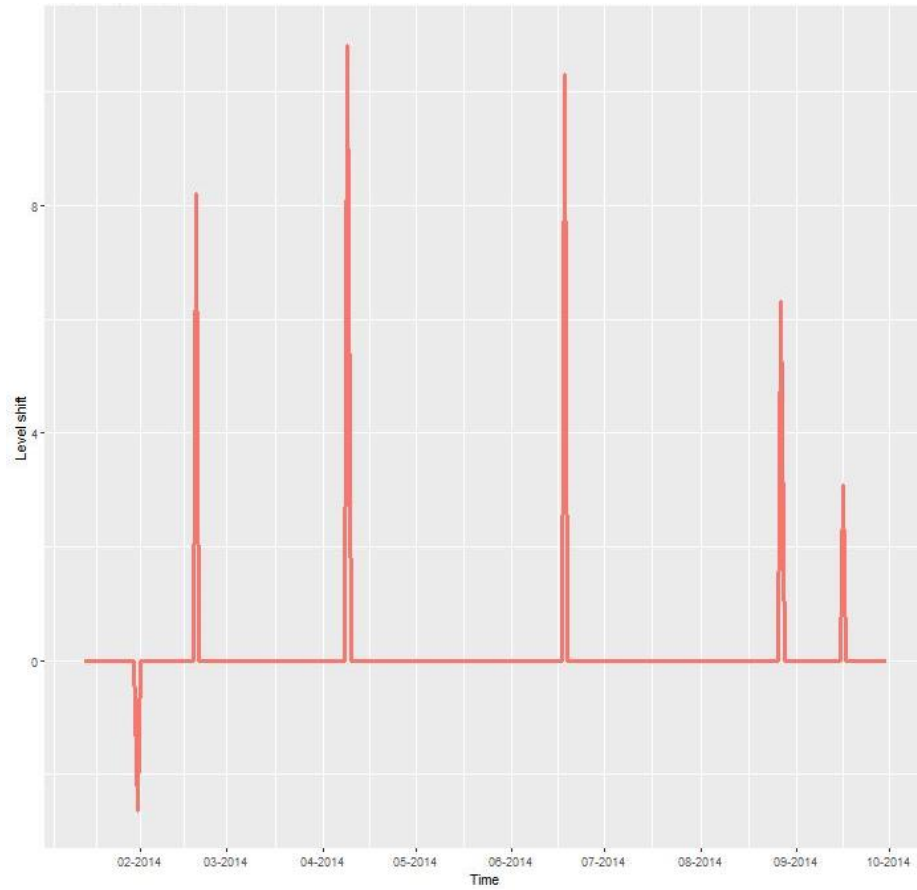


Figure 116. Time series plot of course "CCC-2014B"

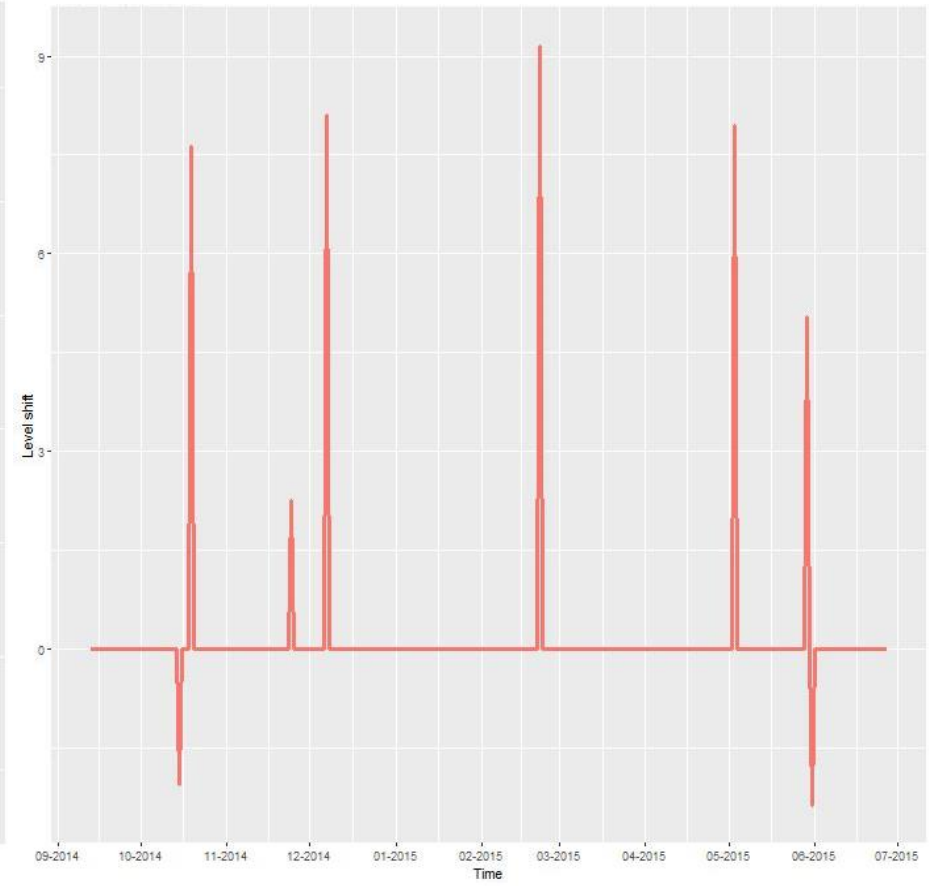


Figure 117. Time series plot of course "CCC-2014J"

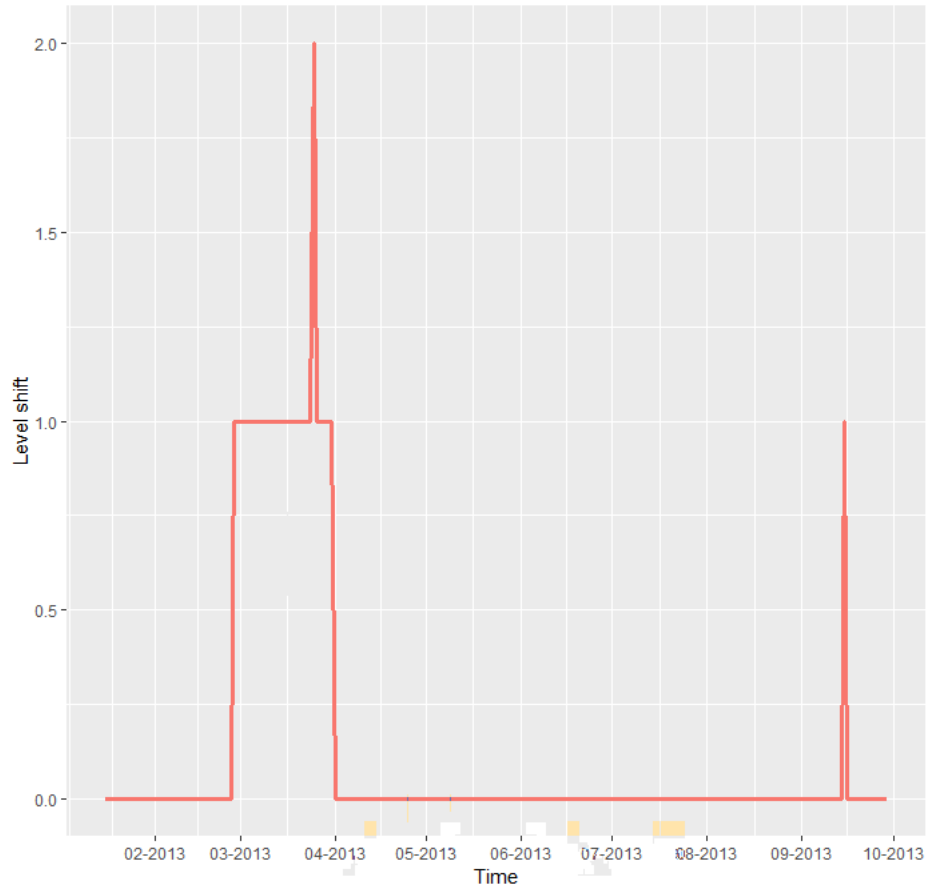


Figure 118. Time series plot of course "DDD-2013B"

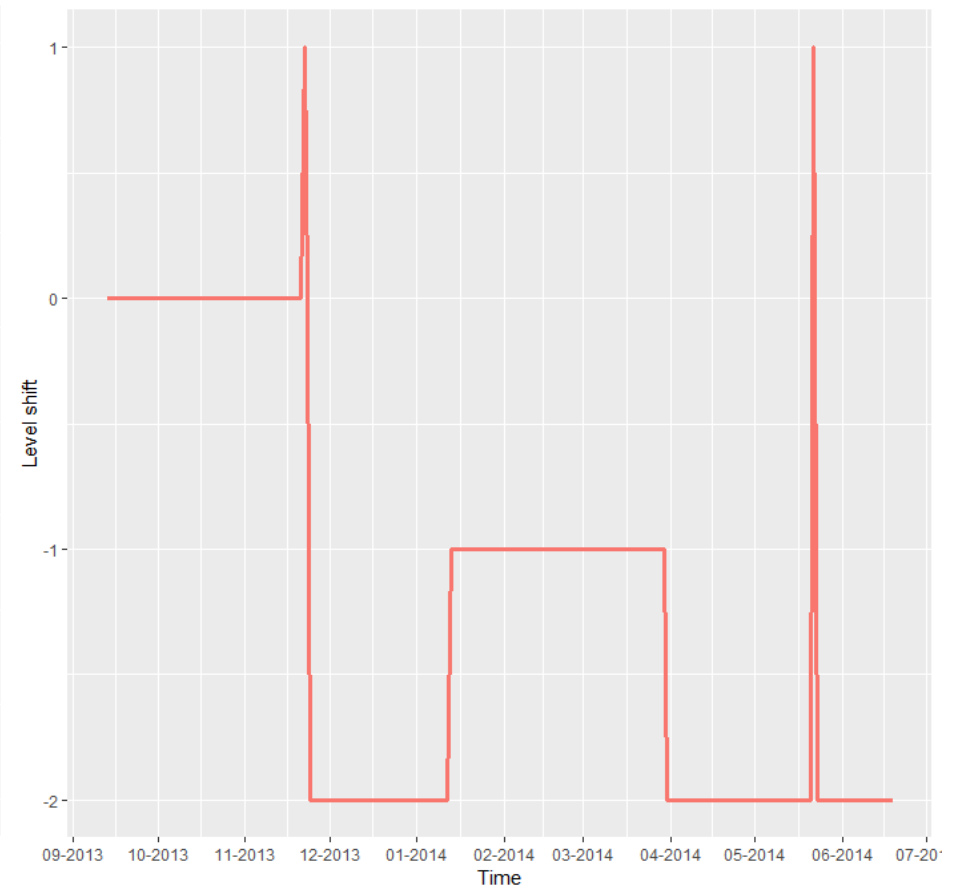


Figure 119. Time series plot of course "DDD-2013J"

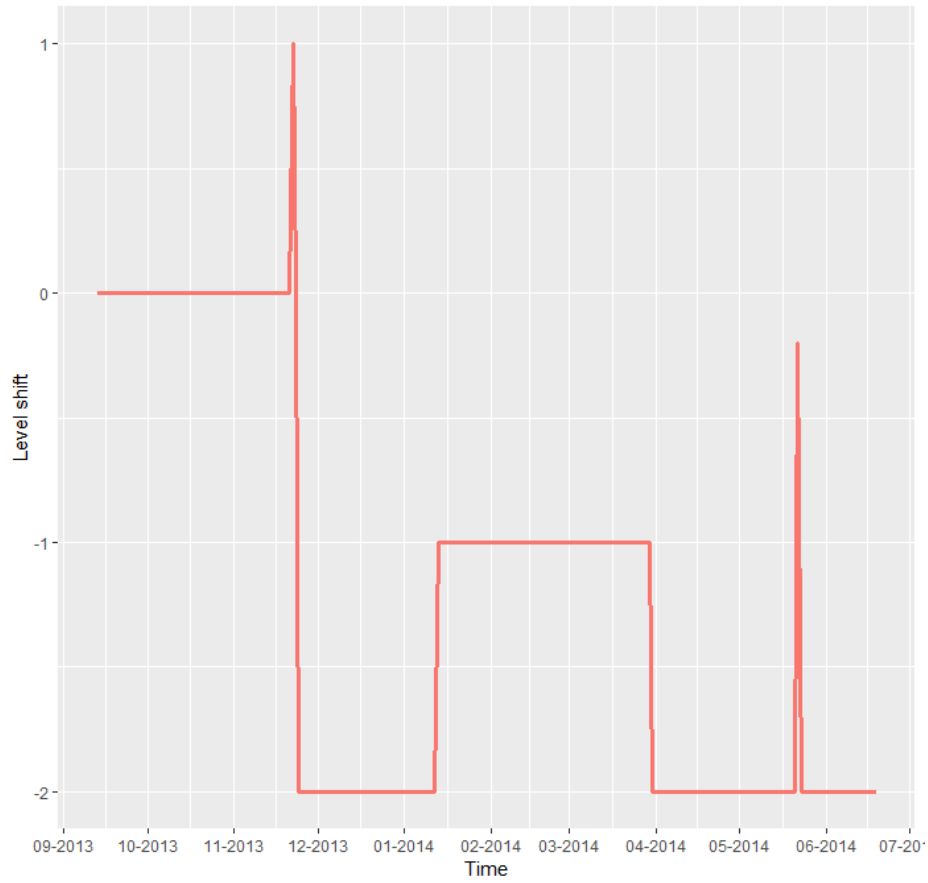


Figure 120. Time series plot of course "DDD-2014B"

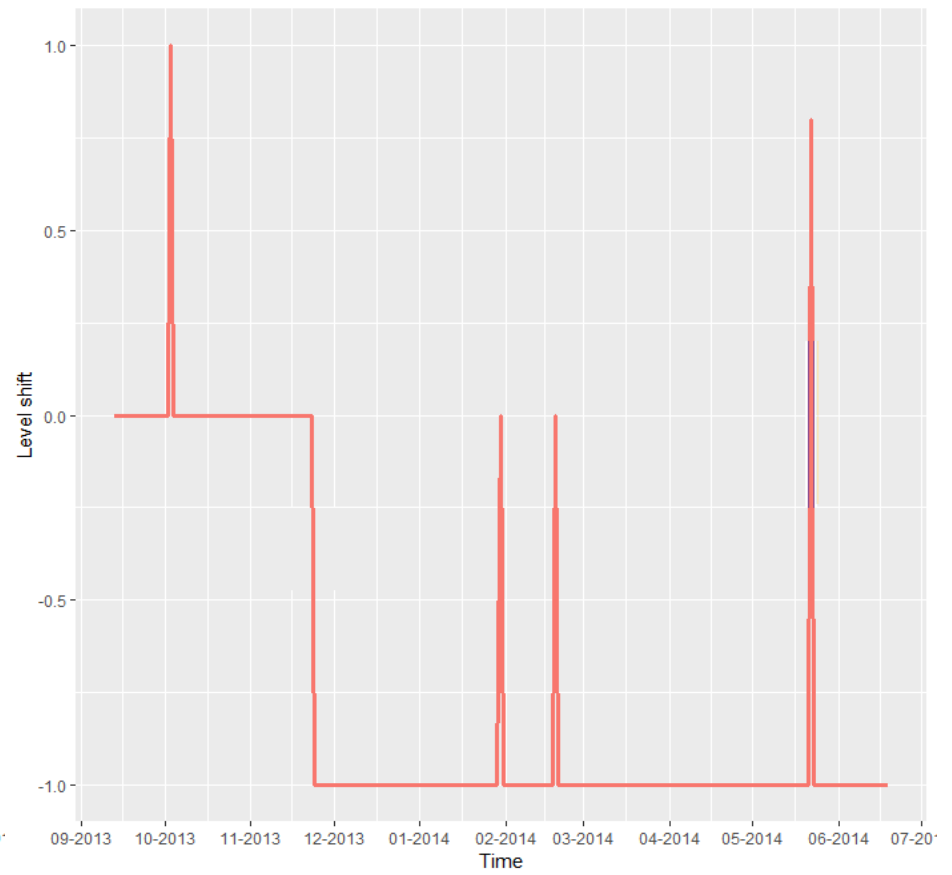


Figure 121. Time series plot of course "DDD-2014J"

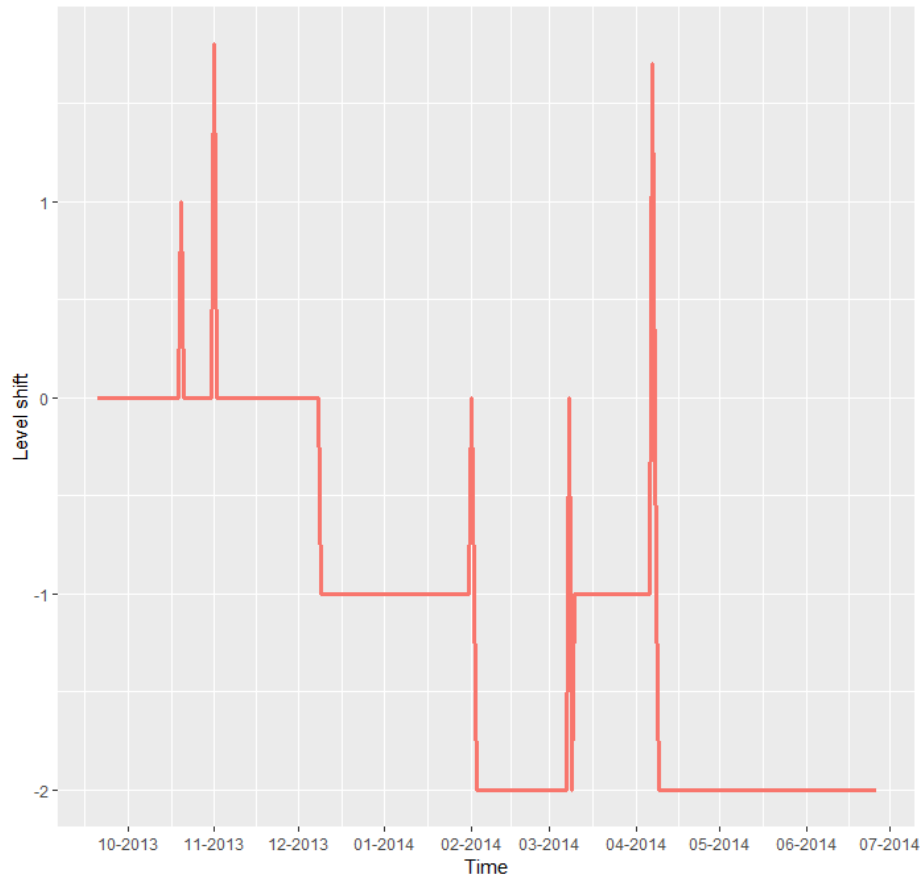


Figure 122. Time series plot of course "EEE-2013J"

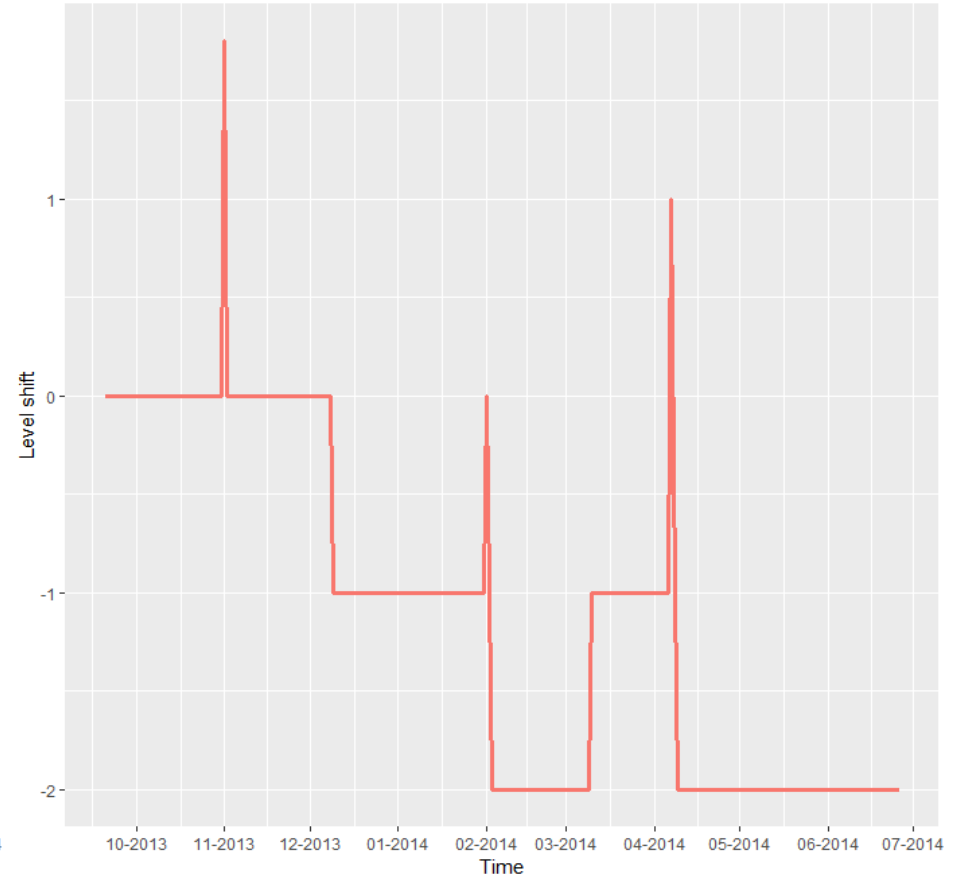


Figure 123. Time series plot of course "EEE-2014B"

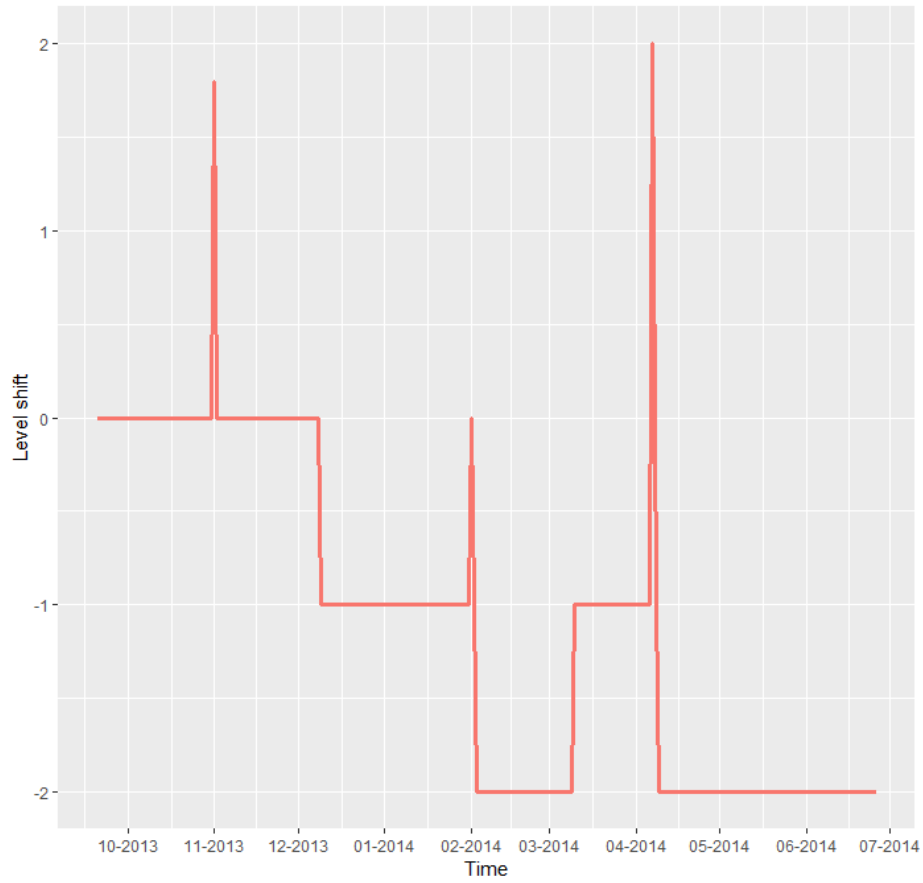


Figure 124. Time series plot of course "EEE-2014J"

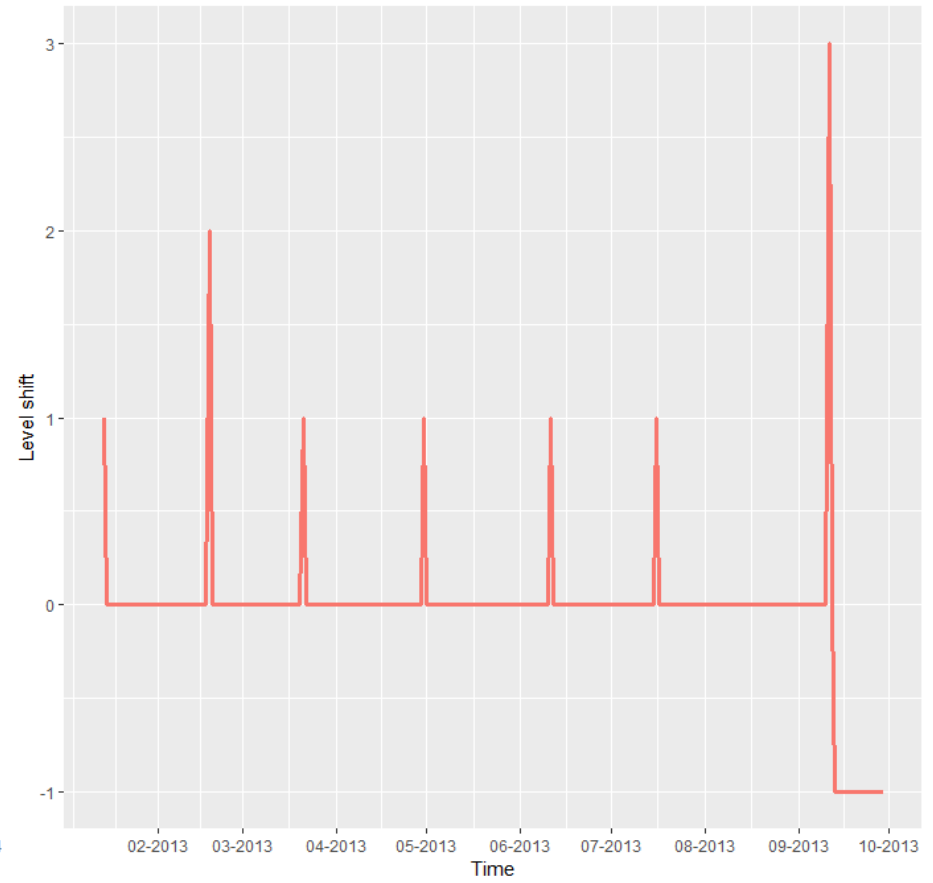


Figure 125. Time series plot of course "FFF-2013B"

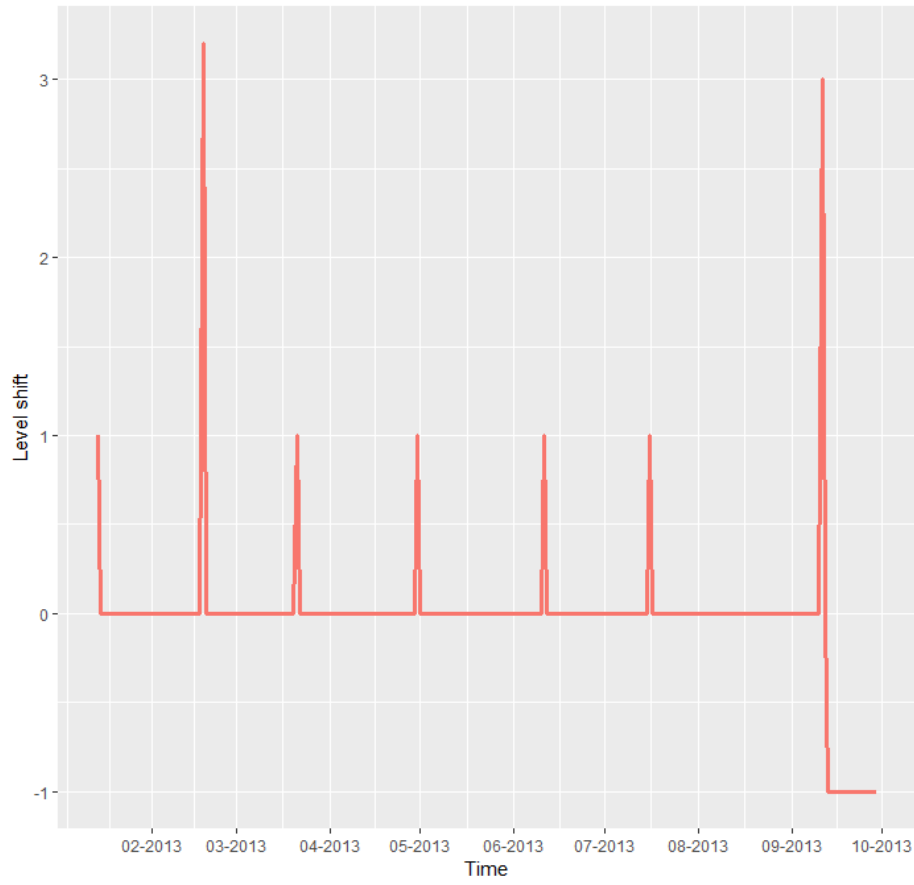


Figure 126. Time series plot of course "FFF-2013J"

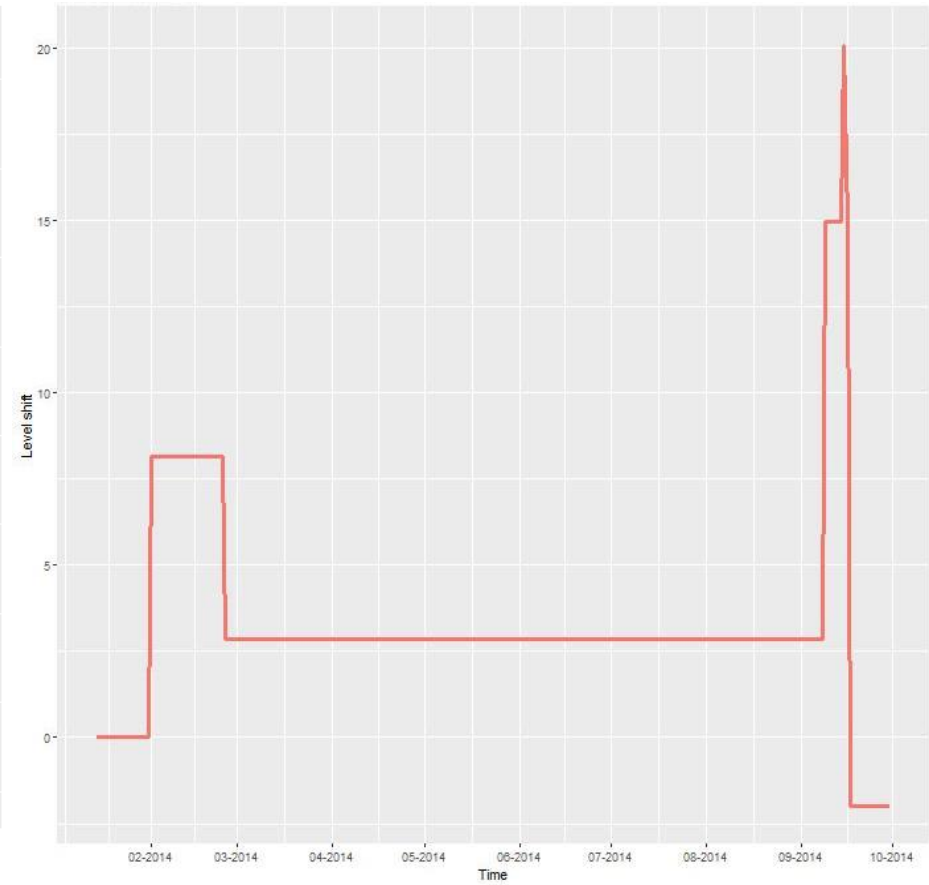


Figure 127. Time series plot of course "FFF-2014B"

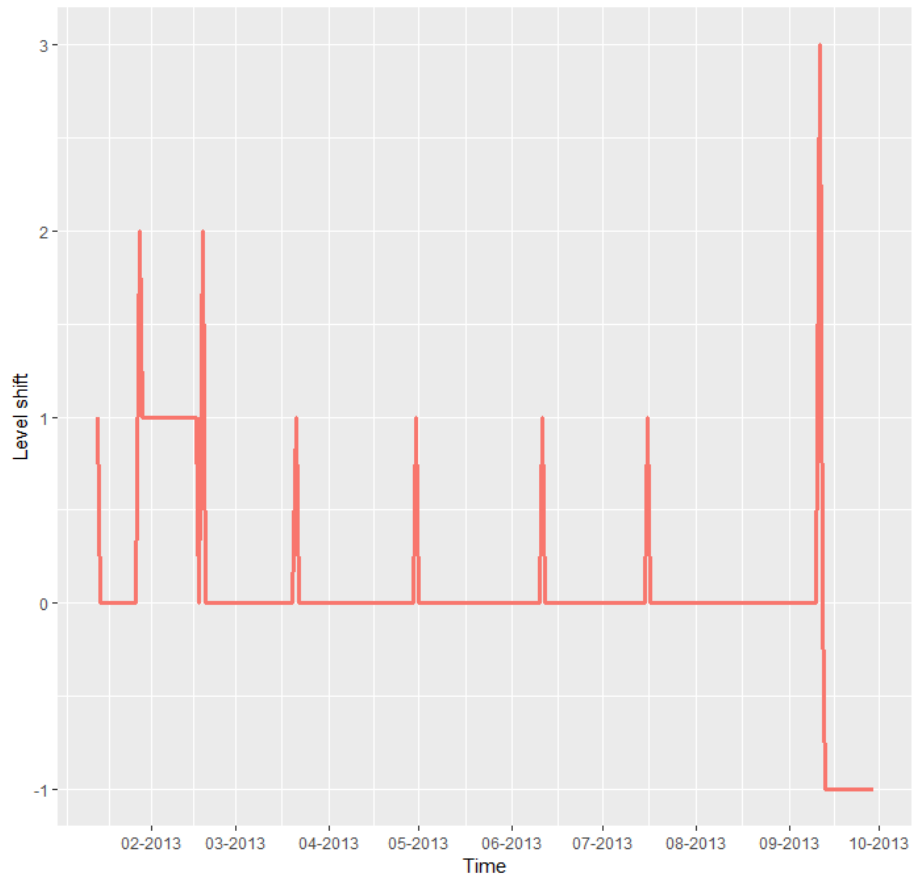


Figure 128. Time series plot of course "FFF-2014J"

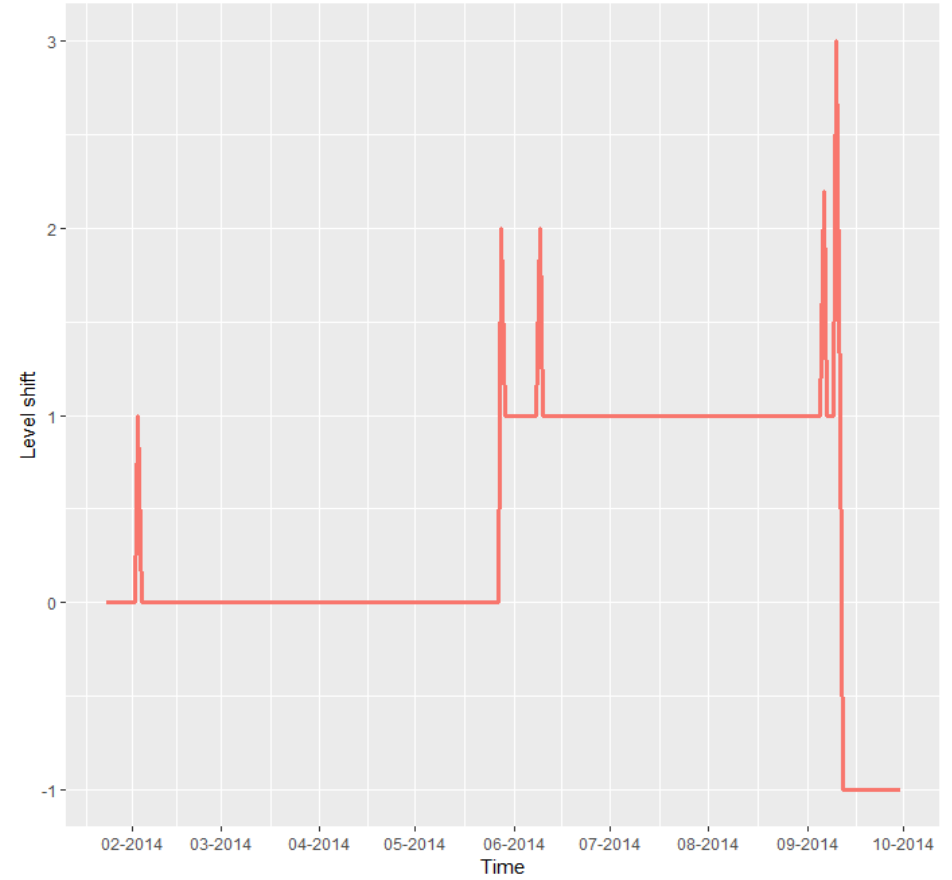


Figure 129. Time series plot of course "GGG-2013J"

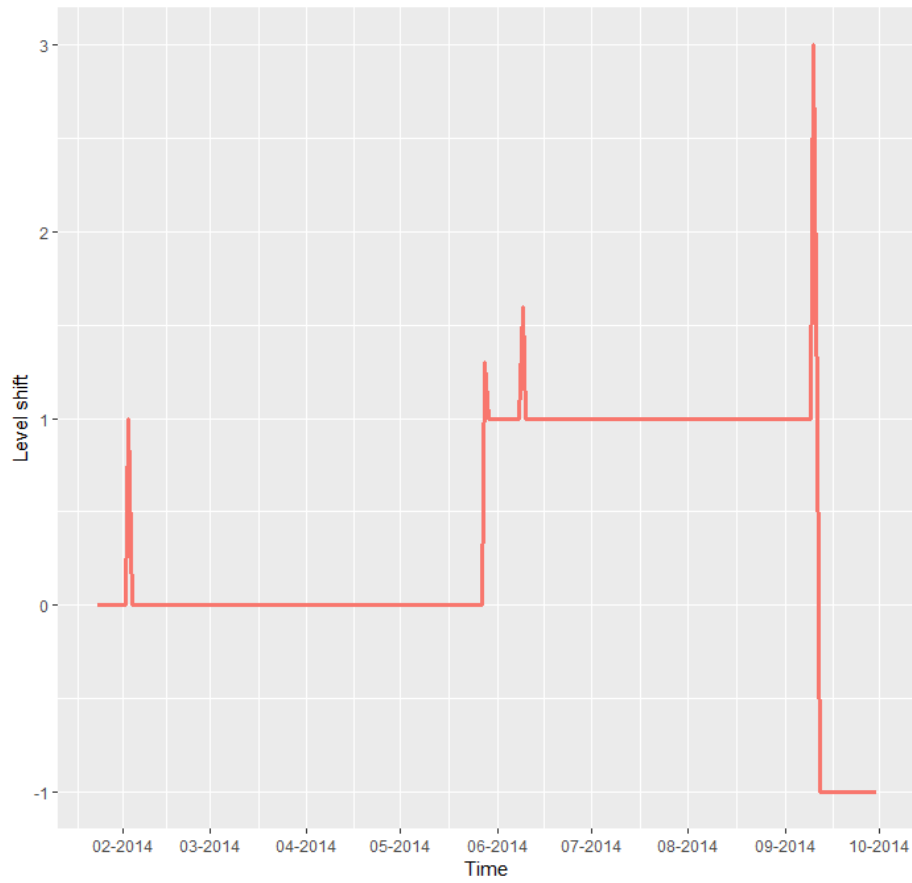


Figure 130. Time series plot of course "GGG-2014B"

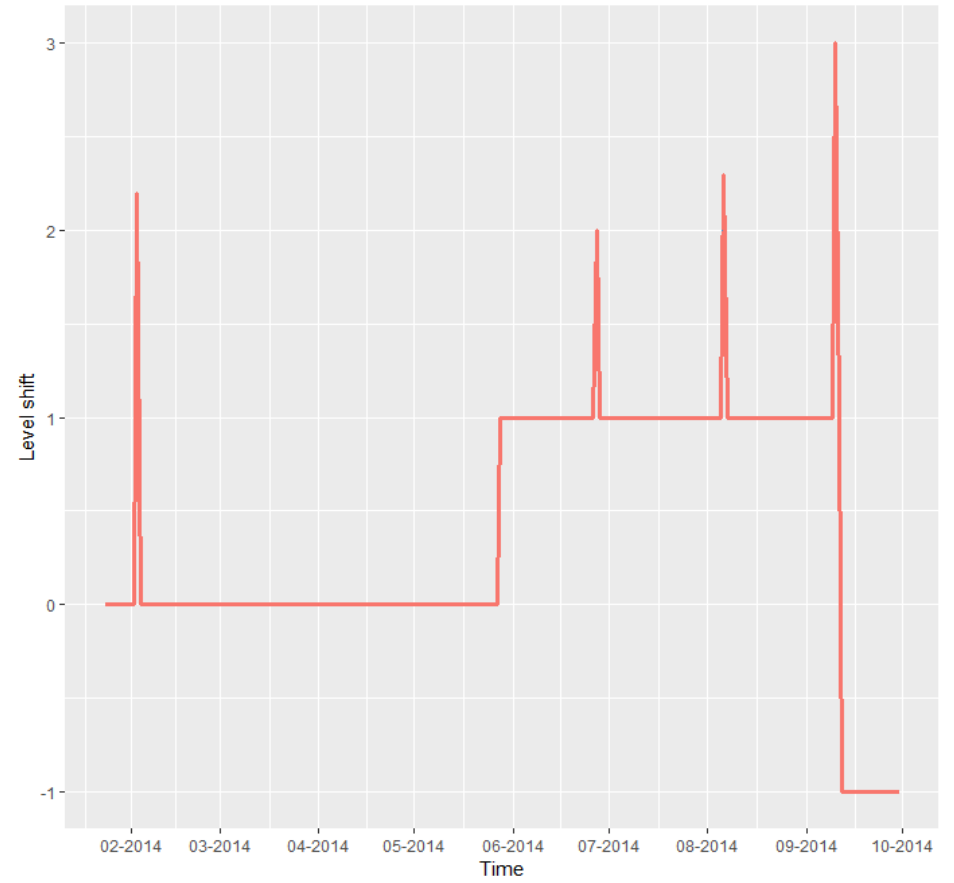


Figure 131. Time series plot of course "GGG-2014J"

As previously acknowledged, outlier's effects described in the previous illustrations are not exhaustively accurate. However, they properly depict the effect of the most prevalent outliers in each of the courses.

Additionally, the heterogeneity present in the way outliers affect each series illustrate the need for an in-depth assessment of each case in order to develop a suitable model.

7.1.1 Time series definition

Finally, it is important to recall the previously detailed similarity of shapes among time series from courses belonging to the same module. Assessment of the outlier's effect results reveal that this resemblance is also present in this scenario. This fact has important implications with respect to the development of a course's forecast model, since, within the correspondent module, the known past effects of an assessment date (an outlier) on the evolution of mean interaction could be used to more accurately predict its effect on an ongoing course. As a result, and attending to the dataset's documentation ([26]) indications to separately study courses correspondent to different presentation groups (presentation codes comprise a number referred to the year and a letter, B or J, correspondent to the beginning of the course in February or September, respectively identifying each presentation group), time series have been rearranged for their analysis as follows:

Course's Module	Presentation group	Years comprised
AAA	J	2013-2014
BBB	B	2013-2014
BBB	J	2013-2014
CCC	B	2014
CCC	J	2014
DDD	B	2013-2014
DDD	J	2013-2014
EEE	B	2014
EEE	J	2013-2014
FFF	B	2013-2014
FFF	J	2013-2014
GGG	B	2014
GGG	J	2013-2014

Table 86. Time series' re-arrangement (forecasting tasks)

One reason behind the separation of these groups according to their presentation group despite their similarity is the overlapping some of their dates present (e.g. a "J" course's end occurs long after the beginning of the next year's "B" course from the same module). However, this does not negate the possibility of reciprocal information between courses from different presentation groups (and same module).

Following, the newly defined time series over which forecasting is to be conducted are presented. Signalled with a red circle in time series referred to "J" courses, there are what are assumed to be interaction decays caused by winter holiday periods. These events will be discussed in the next section.

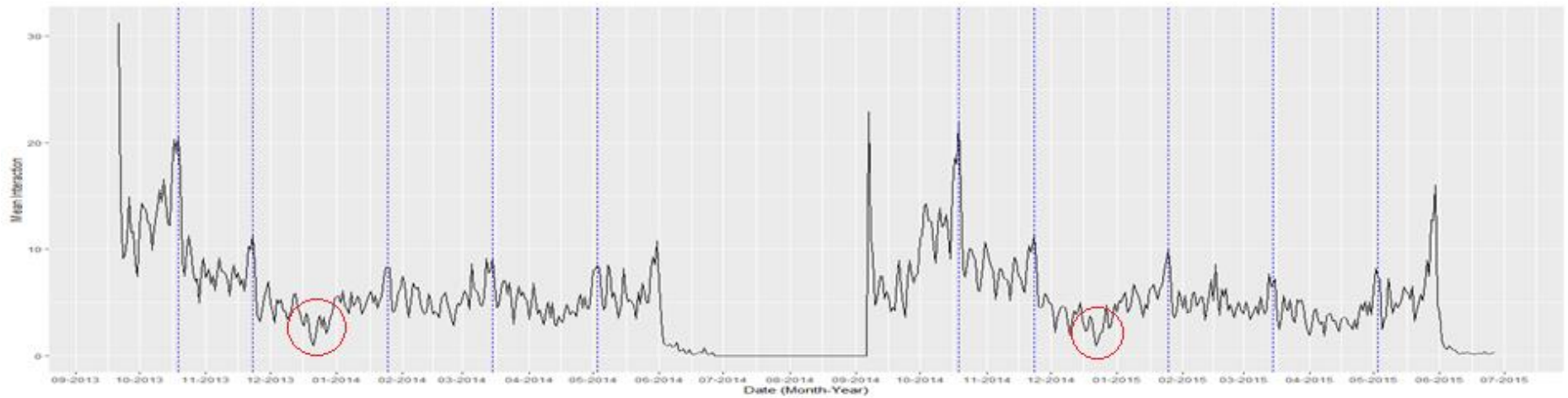


Figure 132. Joint time series plot of courses “AAA-2013J” and “AAA-2014J”

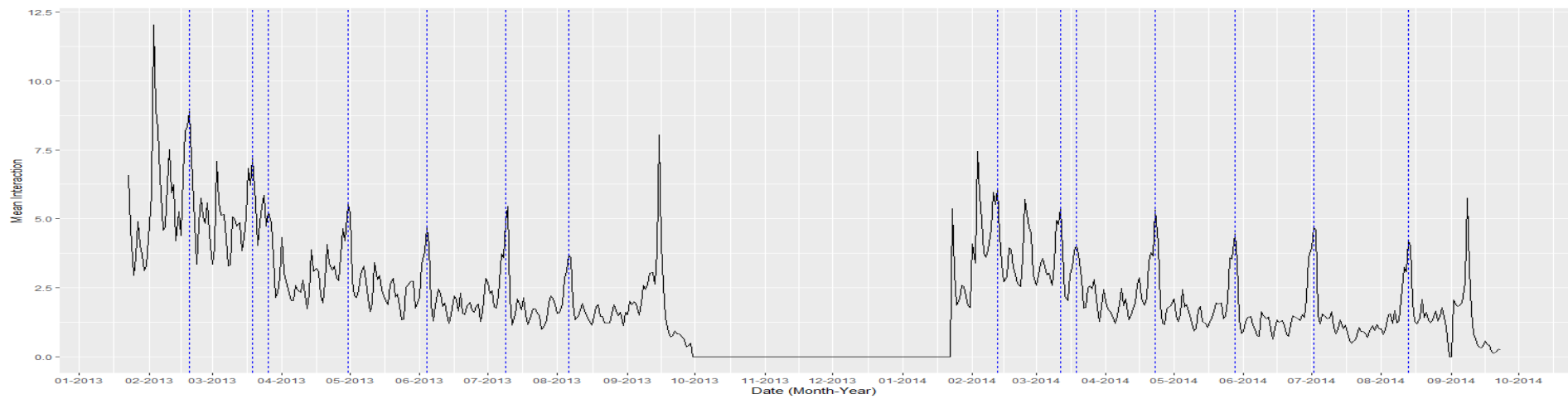


Figure 133. Joint time series plot of courses “BBB-2013B” and “BBB-2014B”

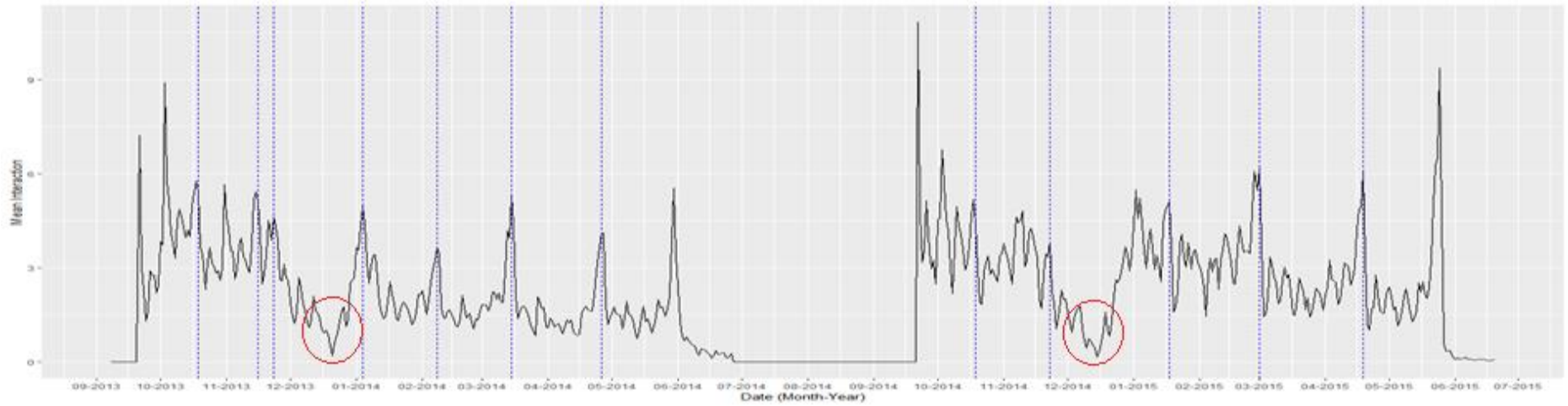


Figure 134. Joint time series plot of courses “BBB-2013J” and “BBB-2014J”

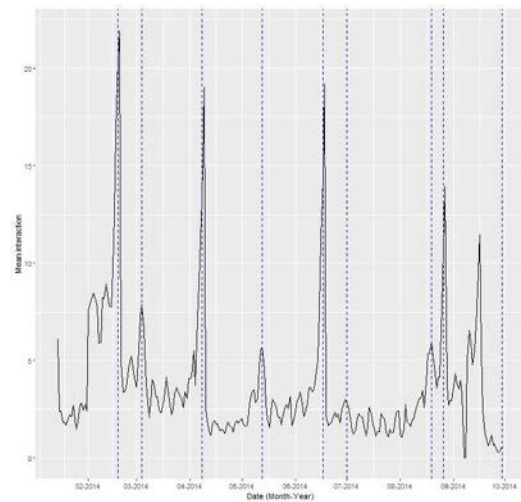


Figure 135. Time series plot of course “CCC-2014B”

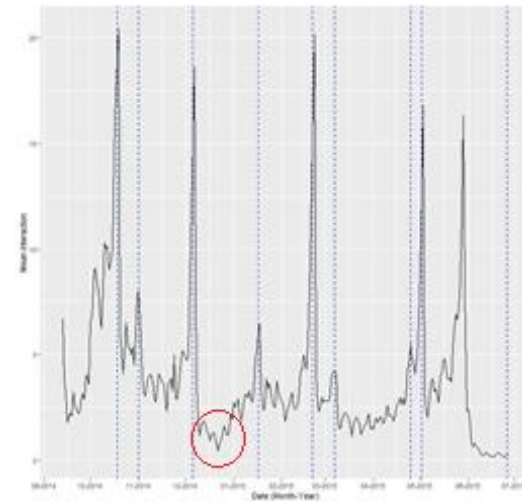


Figure 136. Time series plot of course “CCC-2014J”

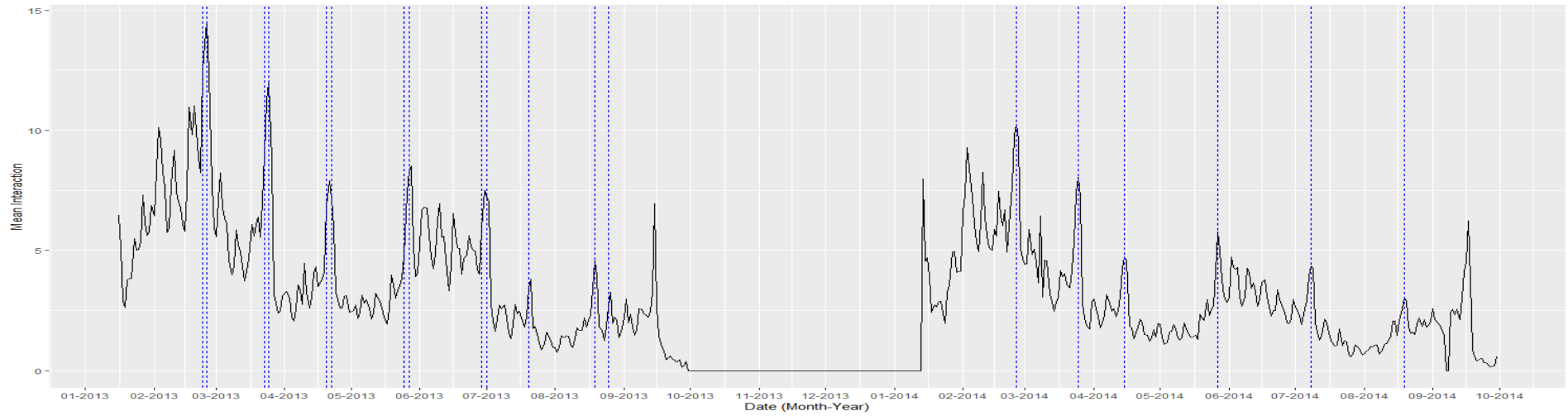


Figure 137. Joint time series plot of courses “DDD-2013B” and “DDD-2014B”

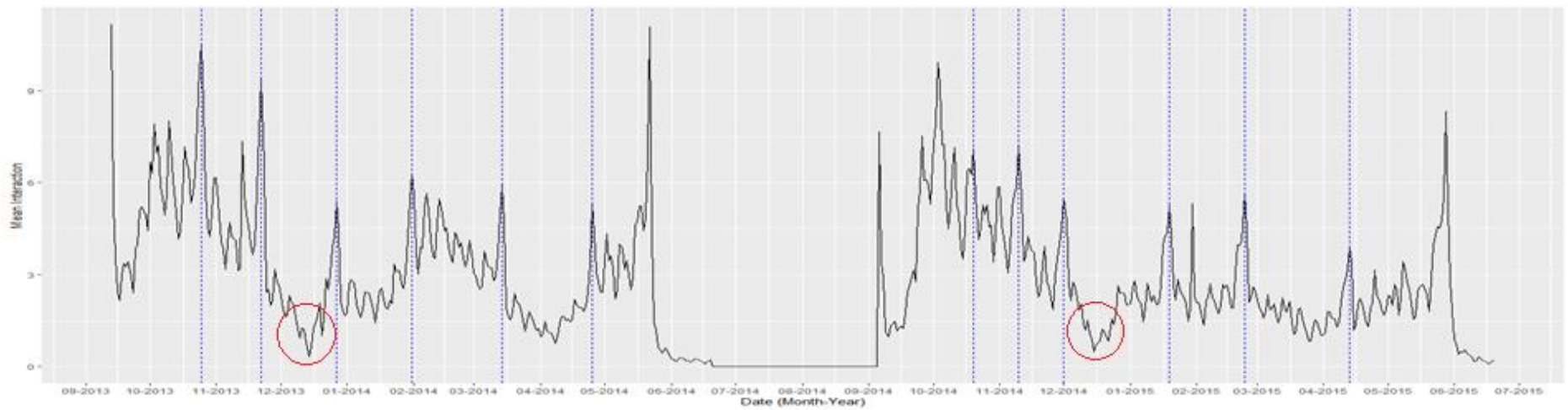


Figure 138. Joint time series plot of courses “DDD-2013J” and “DDD-2014J”

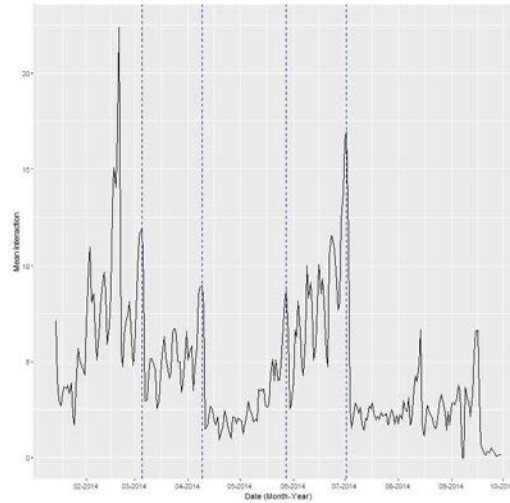


Figure 139. Time series plot of course “EEE-2014B”

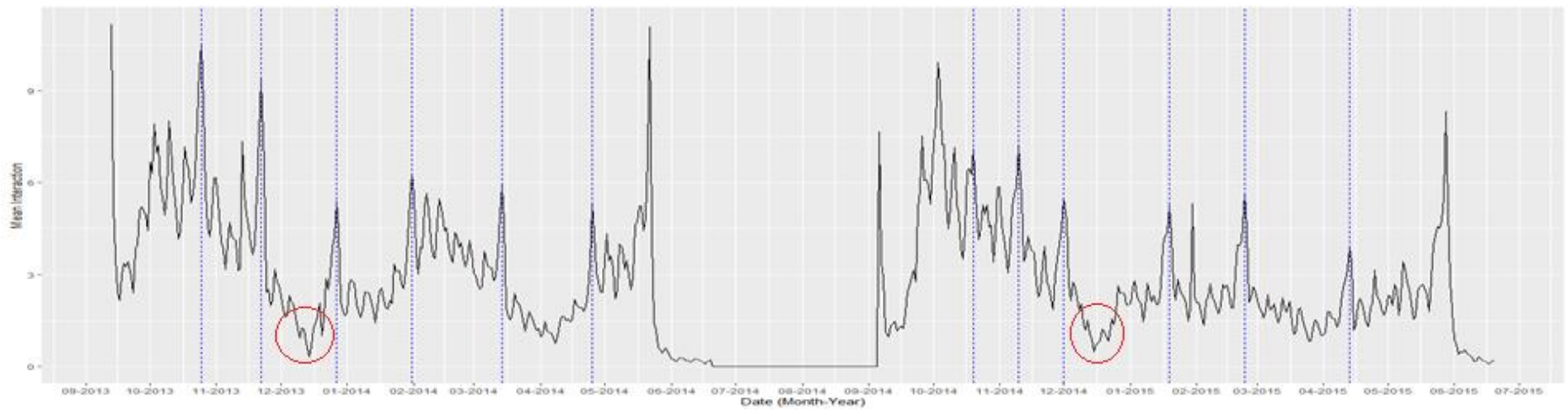


Figure 140. Joint time series plot of courses “EEE-2013J” and “EEE-2014J”

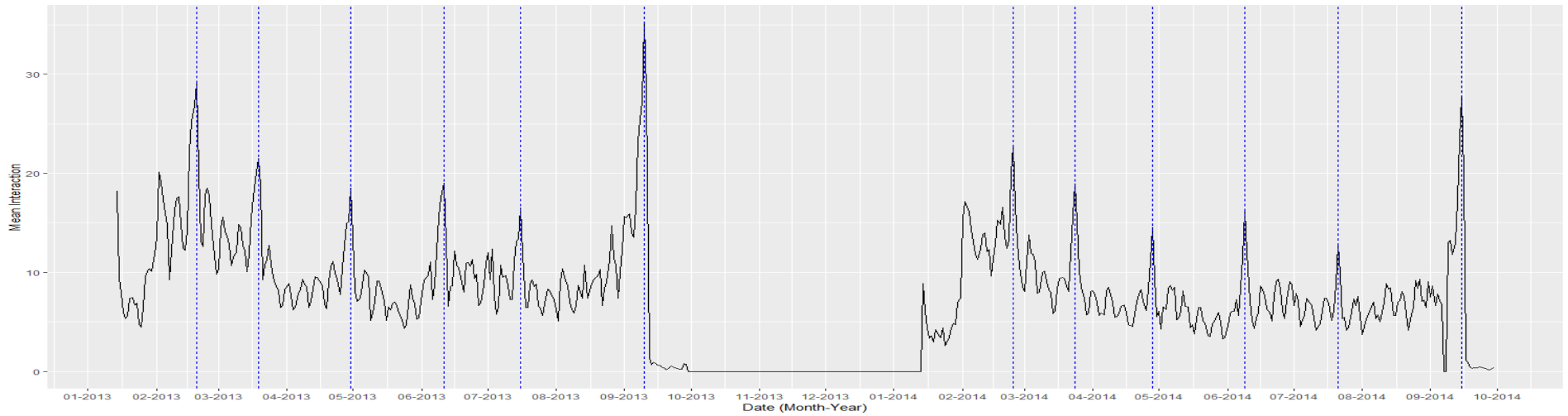


Figure 141. Joint time series plot of courses “FFF-2013B” and “FFF-2014B”

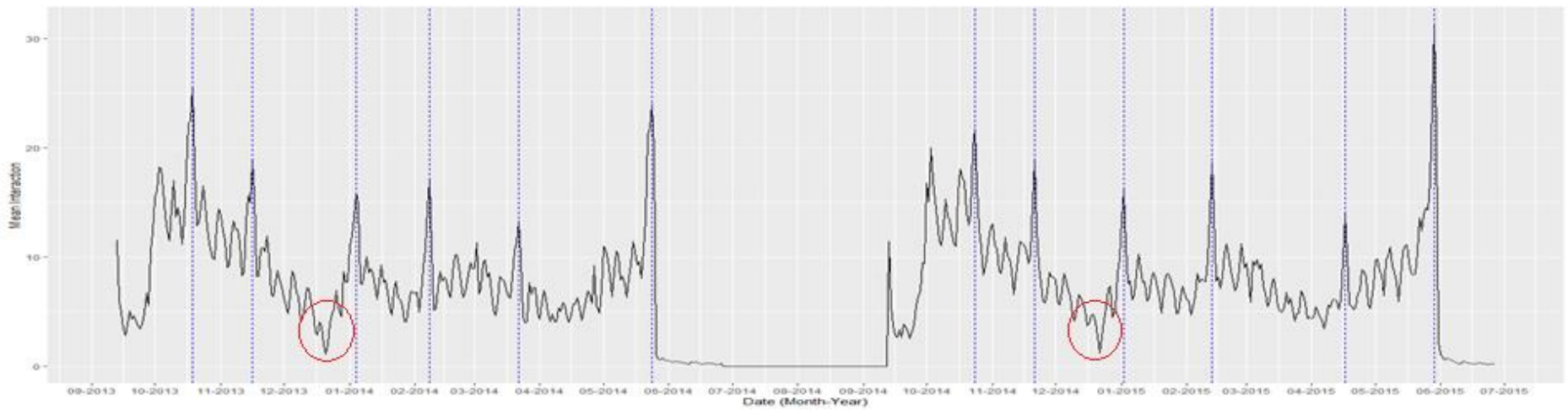


Figure 142. Joint time series plot of courses “FFF-2013J” and “FFF-2014J”

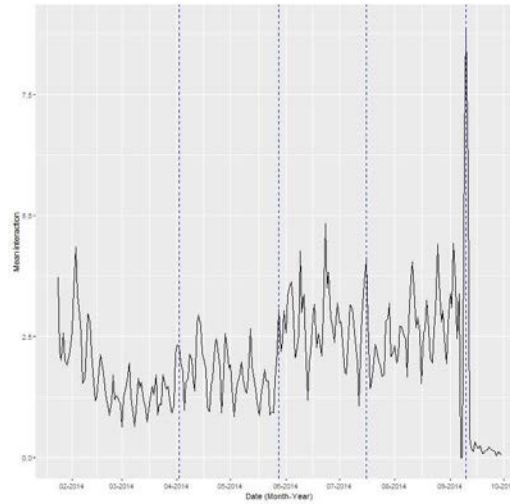


Figure 143. Time series plot of course “GGG-2014B”

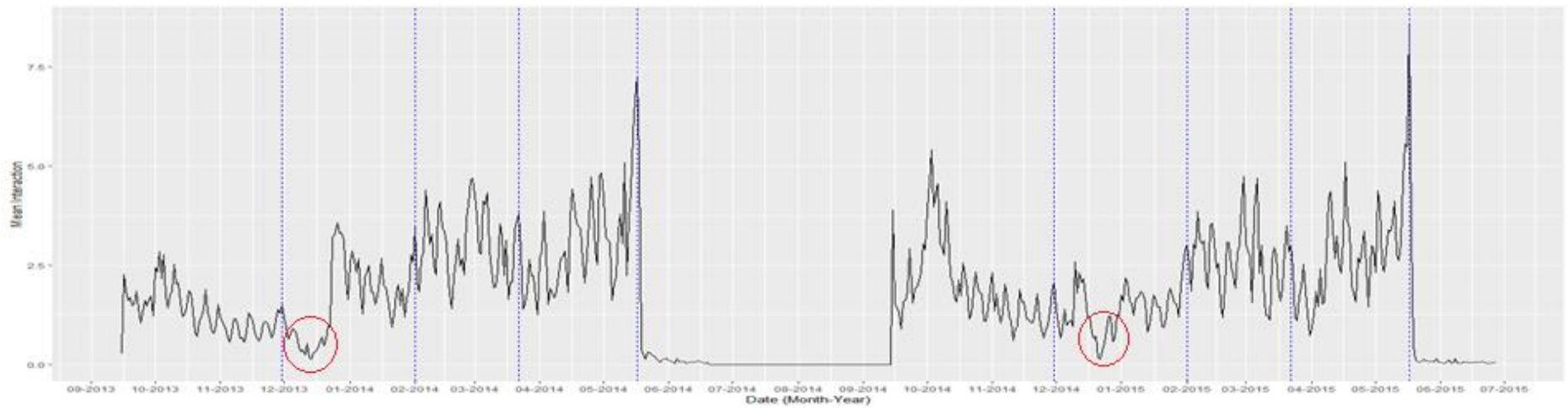


Figure 144. Joint time series plot of courses “GGG-2013J” and “GGG-2014J”

7.2 Identification and definition of external regressors

7.2.1 Outliers' effects

As previously stated, and at sight of what is considered as a general pattern for all courses, the approach given to the treatment of outliers' effects is subject to the assumption that assessment dates imply both an additive outlier (a peak in the graph) and a negative level shift after it.

Additionally, the observation of a local minimum peak (minimum value for each course before its decay to 0 after its ending) located between December and January for courses belonging to the "J" presentation group (signalled with a red circle in the previous graphs) resulted in the consideration of a negative additive outlier located at that point (since no explicit information nor date of such event are available, its localization consisted in the detection of the minimum interaction value for December).

As already mentioned, the encoding of these effects as regressors has been done by confecting indicator variables referred to each of them, in order for the correspondent modelling algorithm to assign a coefficient accounting for its severity.

- Additive outliers consist in an array of 0s, with a unique 1 located in the index corresponding to the addressed outlier.
- Level shifts are made up of 0s until the outlier is reached, from which index 1s populate the array.

For cases in which there is information from a previous course and the next one's period to forecast involves any of these outliers, their indicator variables are summed to their homologous from that previous course, since their dates of occurrence are known, resulting in the possibility to encode regressors addressing their future occurrence.

If no information from previous courses exists, the future indicator variables will be summed to that of the most recent event, so that the coefficient assigned to it is an estimate of the effect of that known outlier.

7.3 Characteristics of the experiments conducted

Before proceeding to detail any of the procedures related to modelling and forecasting of time series, it is necessary to set a framework for the proper understanding of these experiments.

7.3.1 Time constraints

One of the main advantages of the large volume of data with which we count is the possibility of conducting performance tests of the models conducted on enough data to consider them exhaustive. This sets a question about how data should be separated into training and testing sets.

Taking into account both literature treating the amount of “previous data” needed to infer precise forecasts ([50]) and our intendment to represent a real-world scenario, the following specifications have been defined as a guideline to conduct the correspondent processes:

For each time series, analytics are conducted assuming the “present time” as belonging to the most recent course. This defines two different cases referred to the time span comprised by each of the 13-time series previously identified:

- For those time series comprising information referred to two consecutive courses, forecasts will take place within a range comprising one month after the beginning of the second (2014) course and one month before its ending.

It is important to remark the possibility for the correspondent forecasts to be informed with events from the previous course (2013), such as outliers and assessment dates’ effects.

- Time series referred to only one course imply the same time constraints for its forecasts (from one month after its beginning to one month before its ending). However, their referred forecasts count with no information from previous courses and, thus, outliers and assessment dates’ effects will be computed as estimates of those who already took place in it.

It is important to acknowledge the fact that, as mentioned in the dataset’s documentation ([26]), there exists the possibility of informing courses from the same module with data from courses from a different presentation group (e.g. informing “GGG-2014B” with data from “GGG-2013J”). It was decided to not consider this possibility to make our experiments as varied as possible. Since there are already cases in which information from previous courses is drawn, taking

into account those for which there is no such source enriches this aspect, and addresses the possibility of forecasting information about newly created courses (for which there would not be previous information).

7.3.2 Measure of error

As did with regression tasks, the measure of performance for these procedures will be based on the assessment of the random mean squared error (RMSE) on both train and test cases. This will allow for both the evaluation of overfitting and the comparison of different algorithms' efficiency based on the same units and scale as those of the observations (mean interaction).

7.3.3 Measure of residuals' correlation

An important event for which to account in conjunction with the measure of error for a proper evaluation of the models elaborated is the checking of whether or not its residuals are correlated, since a positive case would elicit the fact that there is still relevant information left to be captured.

Although most implementations of forecasting processes address this situation by applying the Ljung-Box test (included R), previous experiments revealed incoherent results for this test, which matches literature's reviews ([51]) of this statistic's low appropriateness for both high volumes of data and auto-regressive models (conditions which our experiments both match). Consequently, and as proposed by the mentioned literature, a Breusch-Godfrey test for serial correlation of a model's residuals has been used instead (both share the same null hypothesis of no correlation).

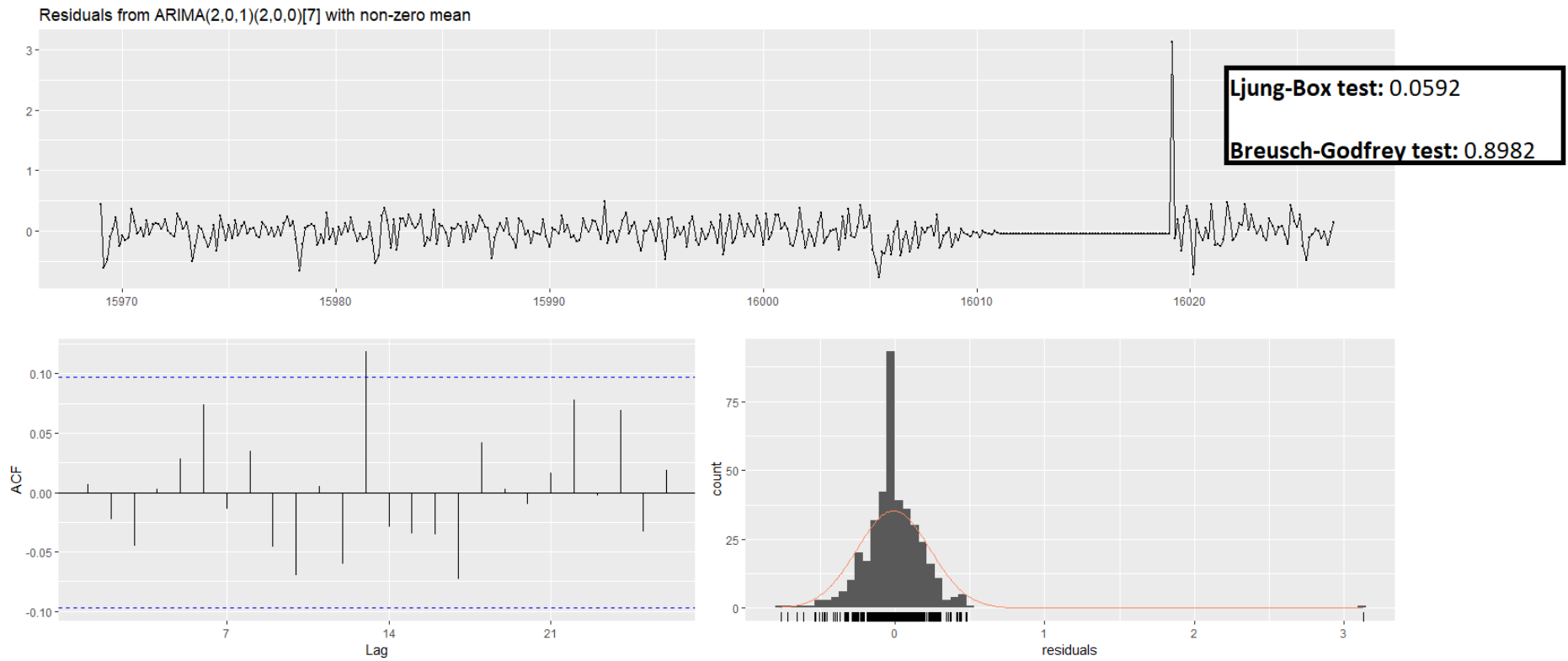


Figure 146. Figure showing discrepancies between Ljung-Box and Breusch-Godfrey test for the same residual evaluation (output from a simple ARIMA model on “AAA-2013J” time series)

This image obtained from the application of a simple ARIMA model (no regressors nor seasonal adjustment applied) goes to show the poor descriptiveness of Ljung-Box test under the previously mentioned circumstances (high volume of data and an auto-regressive model). For the same residuals, which plot observation advances that there is no clear correlation in them, Breusch-Godfrey test offers a more appropriate result in terms of descriptiveness of reality.

7.4 Seasonality assessment

Once the approach given to the conduction of time series modelling has been clearly defined, a specific case proper of this process is used to illustrate the seasonality assessment procedure to be performed on a general basis.

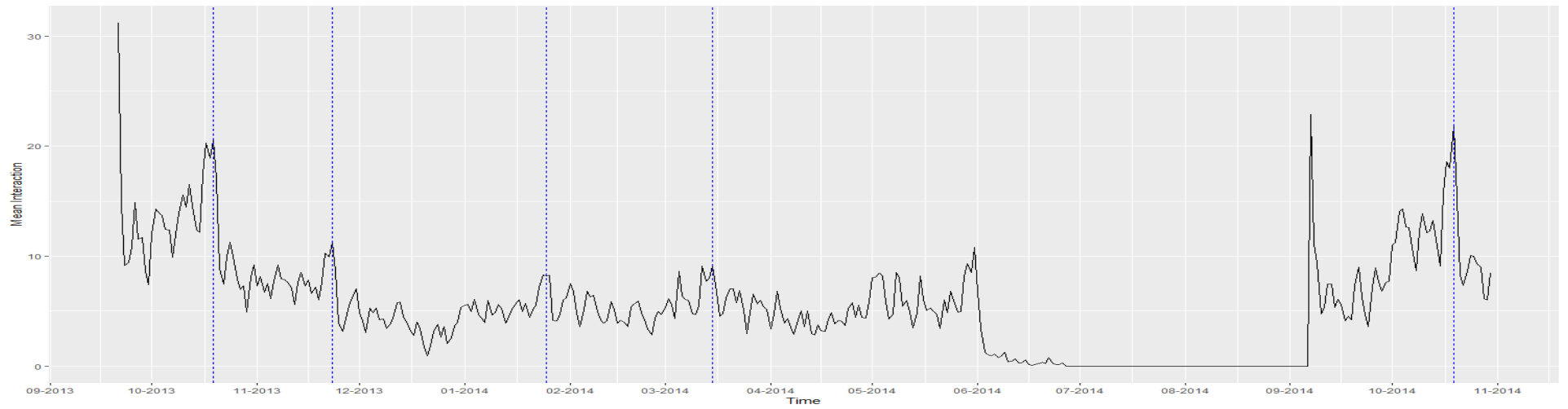
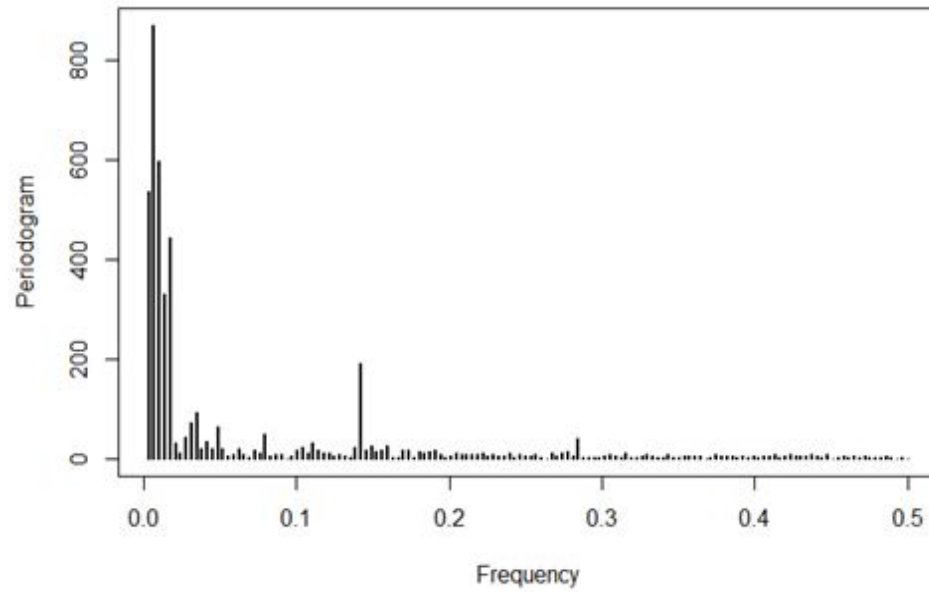


Figure 147. Forecast modelling exemplification: “AAA-2013J” and “AAA-2014J” joint time series

As a first step, a periodogram has been constructed as a description of the observed variable's (mean interaction) distribution over different frequency values (referred to daily data) ([52]-[53]). Only the first course was involved in this process to avoid any distortion caused by the "transition" time span between courses (later on the possibility of addressing it as part of a seasonal component comprising the beginning of a course and that of the next one is discussed).



frequency	spectrum	time
0.006944444	867.99889	144.00000
0.010416667	597.12513	96.00000
0.003472222	534.48250	288.00000
0.017361111	443.84722	57.60000
0.013888889	330.16667	72.00000
0.142361111	191.67118	7.02439
0.034722222	91.14386	28.80000
0.031250000	72.85482	32.00000
0.048611111	64.13393	20.57143
0.079861111	47.61963	12.52174

Figure 148. Periodogram for “AAA-2013J” course time series and its most relevant spectrum spikes (with its correspondent time in days)

Observing the most relevant spikes in the periodogram, it is significant to remark the hypothesis formulated during the graphical assessment of time series with respect to the presence of both weekly and bimonthly seasonalities, both present as potential seasonal frequencies in this analysis (time of 7 and 57 days, respectively).

In which respects to long-term seasonality, it is important to point out that, although it may add valuable information to our model, especially if referred to the period of time between the beginning of a course and that of the next one, the fact that its length (occurrence) is lower than 2 for each of the detailed time series, makes it impossible to account for it. Consequently, 288-days seasonality is discarded.

Following, an agglomerative approach addressing the suitability of different sets of seasonality for an ARIMA model (with automatized parameter setting and no external regressors) applied on the adjusted time series for that set has been conducted to decide for which to account. Seasonal periods are iteratively added to the set only if they significantly improve the root mean squared error of the previous case.

Logarithmic transformation (of time series + 1, to account for the presence of 0 values) took place before the conduction of this process in order to satisfy the homokedasticity (i.e. constant variance) assumption proper of ARIMA models (a non-constant variance test revealed the need for this transformation). It is important to remark that this transformation does not alter the periodogram's assessment results.

Seasonal periods	AICc	Training RMSE	Test RMSE
None	54.96	5.514	3.312
7	-82.18	5.263	3.06
7, 57	-182.48	5.414	2.68
7, 57, 72	-182.48	4.939	2.841
7, 57, 96	-291.22	5.34	2.59
7, 57, 144	-496.19	4.713	3.749

Table 87. Seasonal periods' selection process (forecasting tasks)

Although the inclusion of the 96-days seasonal period still diminishes RMSE if compared to the (7,57) set, this improvement is not considered as significant enough to account for this seasonality (only 0.09) , and it could be concluded that improvement ceases after the inclusion of the bimonthly seasonality (seasonal period of 57 days), thus leading to the conclusion of 7 and 57-days seasonal periods as descriptors of the main seasonal components of this time series.

7.4.1 Time series decomposition and adjustment

Finally, time series decomposition aimed towards the distinction of 7 and 57-days seasonal period's components is conducted in order to adequately adjust the series. The result of this process is presented for both the regular and logarithmic series.

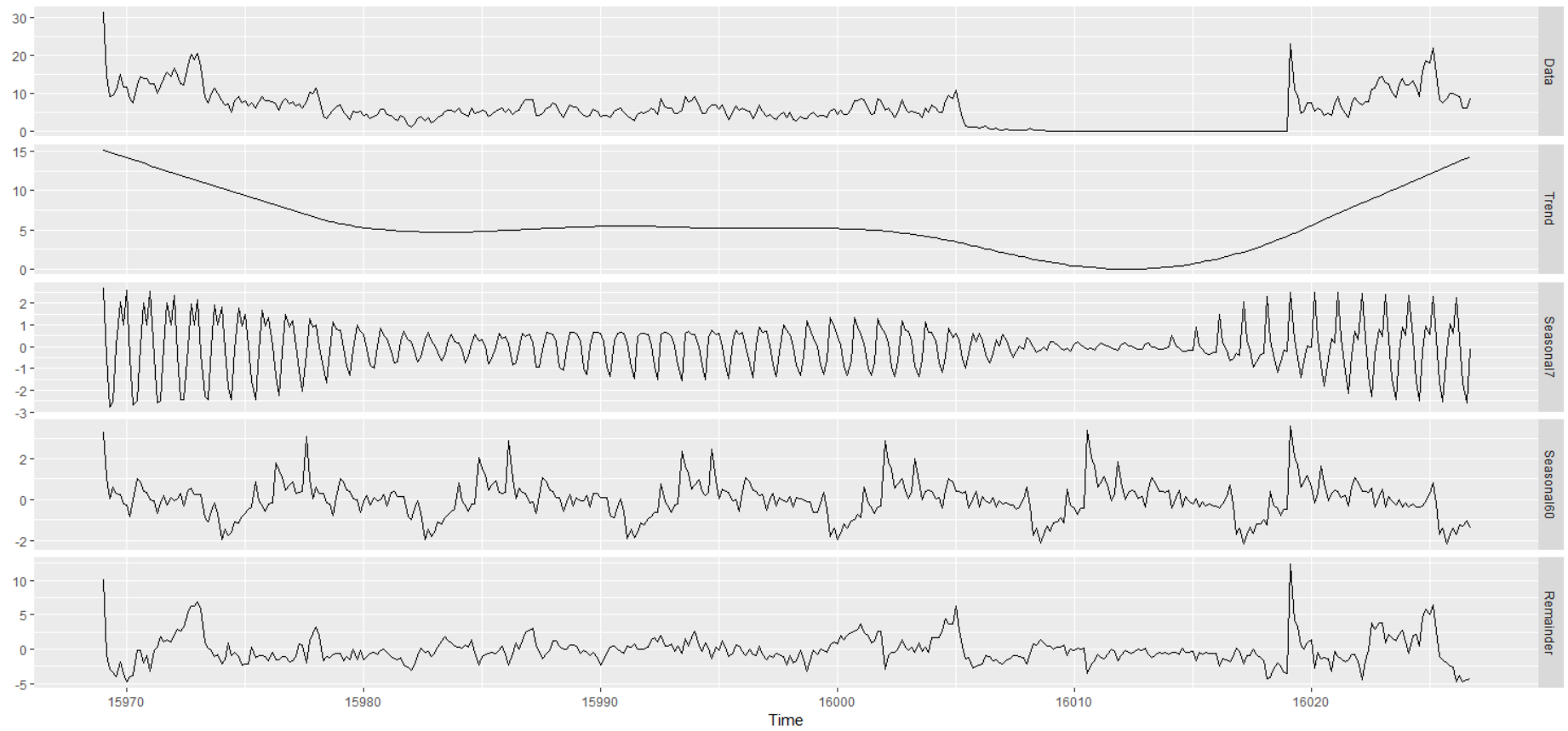


Figure 149. Forecast modelling exemplification: “AAA-2013J” and “AAA-2014J” joint time series decomposition (weekly and bimonthly seasonality)

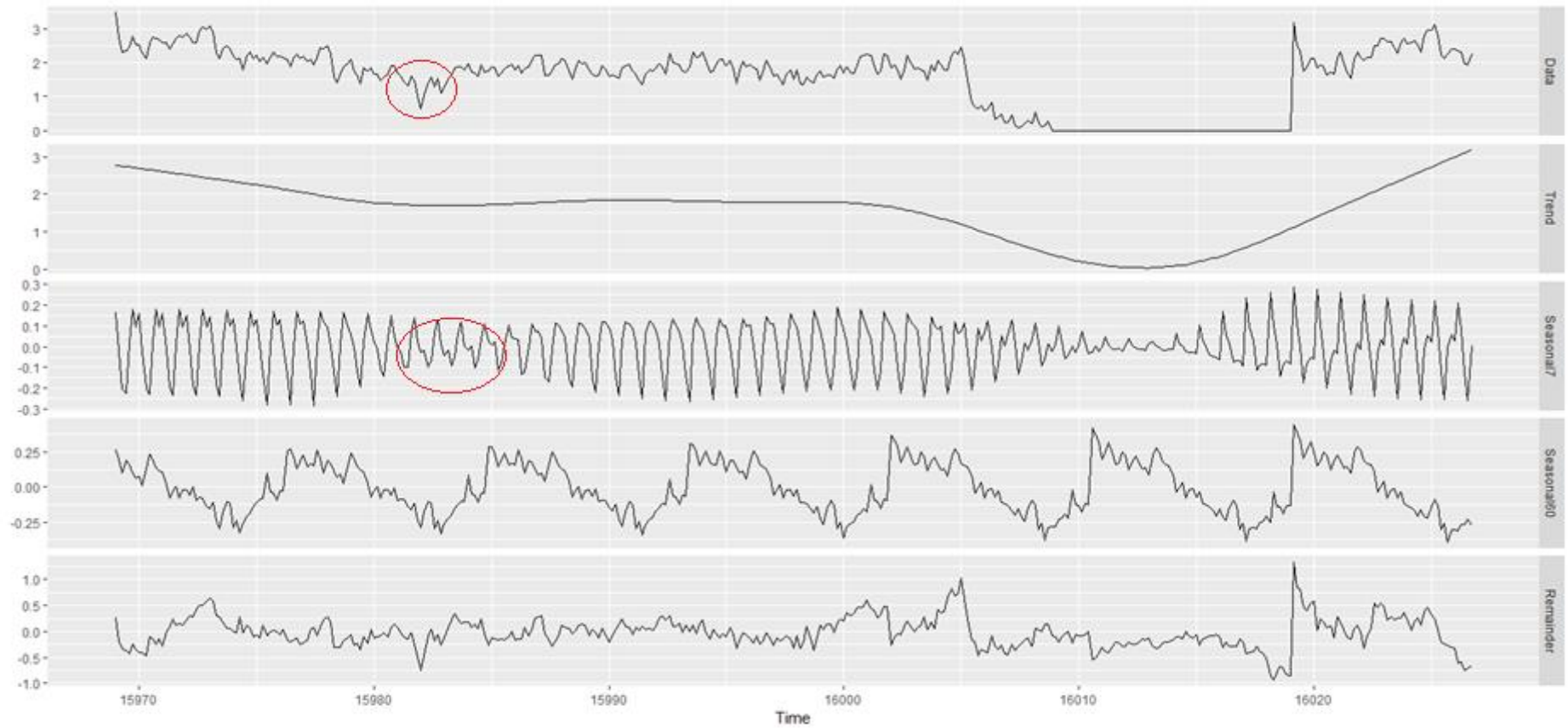


Figure 150. Forecast modelling exemplification: “AAA-2013J” and “AAA-2014J” joint time series logarithmic decomposition (weekly and bimonthly seasonality)

In both cases it is clear the over-time persistence of each of the identified seasonal components. However, it can be seen how variance of the weekly component seems to decrease (and go back to normal) along two distinct periods. As signalled in the second plot with red circles, the first variance-reduction period can be explained by the winter holiday period (between December and January) previously pointed out as occurring in “J” courses, which result in a significant decrease of mean interaction (the event is clearer if observed on logarithmic time series). The second period where variance of the weekly seasonal component decreases corresponds to the time lapse between the two courses, where mean interaction is null.

Concluding, the seasonally adjusted time series is presented (only adjustment for the logarithmic transformation is shown as it is used for most experiments).

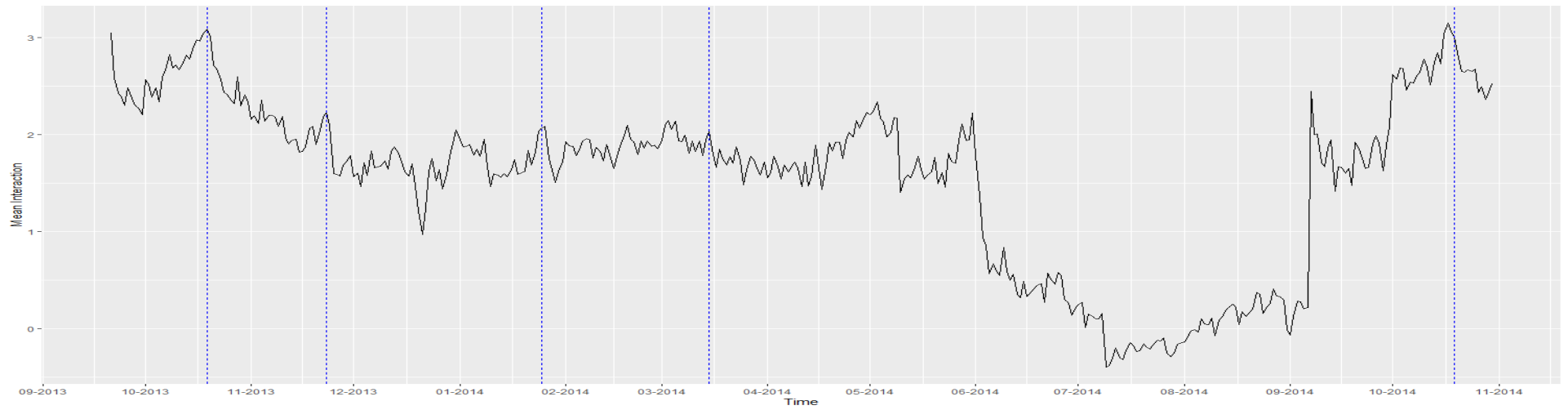


Figure 151. Forecast modelling exemplification: “AAA-2013J” and “AAA-2014J” adjusted (for weekly and bimonthly seasonality) joint time series (logarithm)

7.4.2 Fourier series terms

Once seasonality has been identified it is possible to develop their correspondent Fourier series terms.

$$x_{1,t} = \sin\left(\frac{2\pi t}{m}\right), x_{2,t} = \cos\left(\frac{2\pi t}{m}\right)$$

Figure 152. First terms of a Fourier series

Given the previous formula, and accounting for the identified seasonality, we would count with two different Fourier series describing weekly seasonality ($m=7$) and bimonthly seasonality ($m=57$). The number of terms to consider for each series will be discussed lately, since it substantially affects the dimensionality of the model.

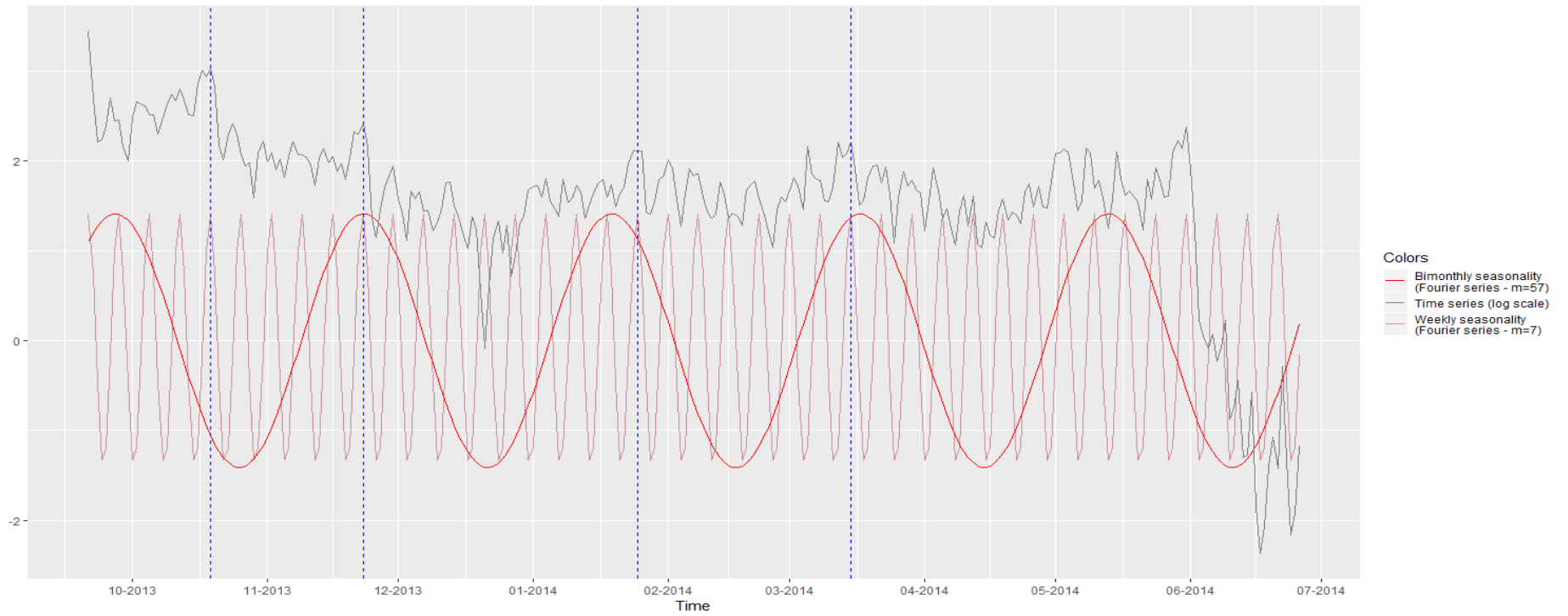


Figure 153. Overlapping of Fourier time series (accounting for weekly and bimonthly seasonality) with “AAA-2013J” time series (logarithm)

The graph corresponds to the overlapping of 1-term weekly ($m=7$) and bimonthly ($m=57$) Fourier series with the first month's time series (which have been logarithmically transformed to fit a more appropriate scale). Although in a real modelling process in which the Fourier terms included as regressors would be adequately scaled by the correspondent coefficients to better fit the time series, this plot clearly illustrates the presence of the mentioned seasonality, with time series' local peaks specific of each day-range considered (7 and 57) matching both those of the weekly and bimonthly Fourier series. Additionally, it is remarkable how assessment dates, which imply interaction local maximums, also match Fourier series' peaks.

7.5 Modelling

Time series modelling detail comprises two sections:

- Evaluation of the functioning and performance of a set of 4 different algorithms on an example case. Each case of study defined in this process' introduction is assessed (if, as previously discussed, the algorithm allows for its conduction).

Cases of study	
Original data	Original data with regressors
Seasonally adjusted data	Seasonally adjusted data with regressors

Table 88. Forecasting tasks' cases of study

The algorithms involved are:

- Auto-regressive Integrated Moving Average (ARIMA): for which stationarity is addressed previous to its application.
- Neural Networks
- Exponential smoothing models:
 - ETS
 - TBATS

Additionally, combination of the best resulting models from the previous set of algorithms will be assessed (consisting on the mean of their predictions).

The study case used for this procedure is the already defined for the detection of seasonality, corresponding to the "AAA-2013J" and "AAA-2014J" joint time series.

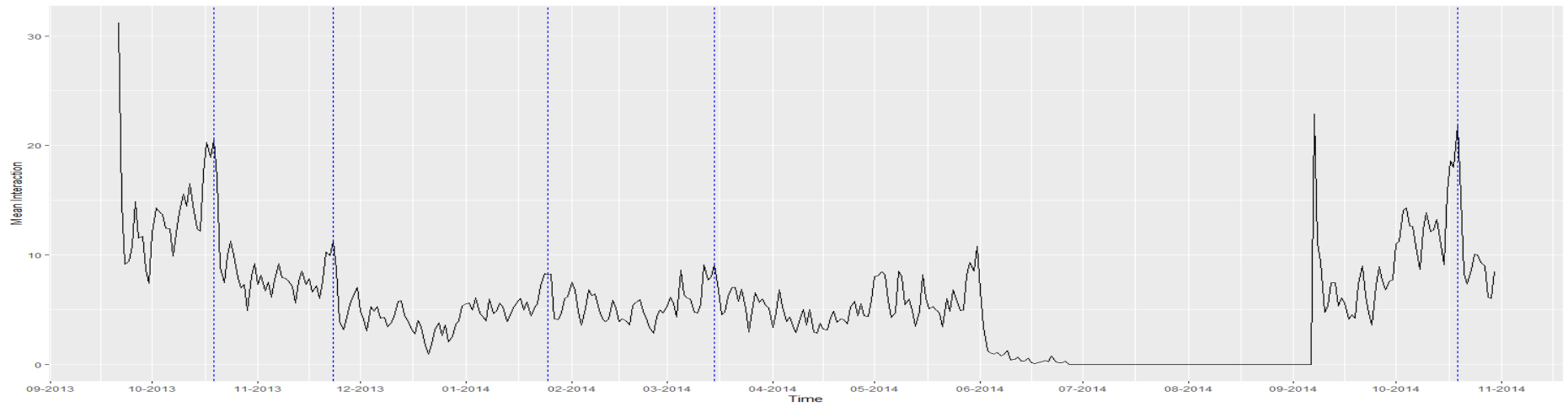


Figure 154. Forecast modelling exemplification: “AAA-2013J” and “AAA-2014J” joint time series

- Summary of results from different forecasting scenarios conducted on each time series identified in accordance to the constraints stated in the experiments' characteristics detail section. Consequently, for each course, different monthly datetimes from which to forecast are set and evaluated along with different monthly forecast horizons (days to forecast).

Only best performing algorithms from the previous process are involved in these experiments.

To avoid redundancy in the exposure of results, it has been considered as appropriate to advance the fact that Breusch-Godfrey test for each model identified as the best-performing in the first section, reflected no correlation existing among their residuals, thus allowing for the assumption of appropriateness of these models. However, residual graphs are presented to allow for the examination of this statement's veracity.

7.5.1 ARIMA

Different parameters sets identified for each study case are the result of an automatized process (provided by "auto.arima()" R function) for which different restrictions were manually set, in order to explore different possible solutions. These restrictions consisted in the inclusion/rejection of both stationarity and seasonality assumptions.

7.5.1.1 Stationarity

The previously advanced addressment of stationarity involves both the evaluation of variance and mean, which must be constant in order for the series to be stationary. Consequently, different tests for the guaranteeing of this characteristics have been conducted (results shown correspond to the application of the test to the example case, although each time series requires this processing previous to its modelling):

- Non-constant variance test (Breusch-Pagan test, [54]): p-value = **0.12**

This result reveals the need for the stabilization of the variance along the time series (although the time span between courses, for which mean interaction is 0, may distort these results and inflate the assumption of non-stability of variance, observation reveals a clear variation of this statistic between different stages of each course), for which a logarithmic transformation is employed.

- Kwiatkowsky-Phillips-Schmidt-Shin (KPSS, [55]) test of logarithmic data: test statistic = **1.521**

This test's output indicates the presence of unit roots in the time series, thus making the application of a differencing process to solve this event and stabilize the mean.

7.5.1.2 ARIMA model (Fourier seasonal terms)

ARIMA parameters	Training RMSE	Test RMSE
(1,1,1)	5.493	5.336
(2,0,1)	5.5	3.073
(2,0,1) (1,0,1)[7]	5.499	2.758

Table 89. Results for regular ARIMA forecasting process

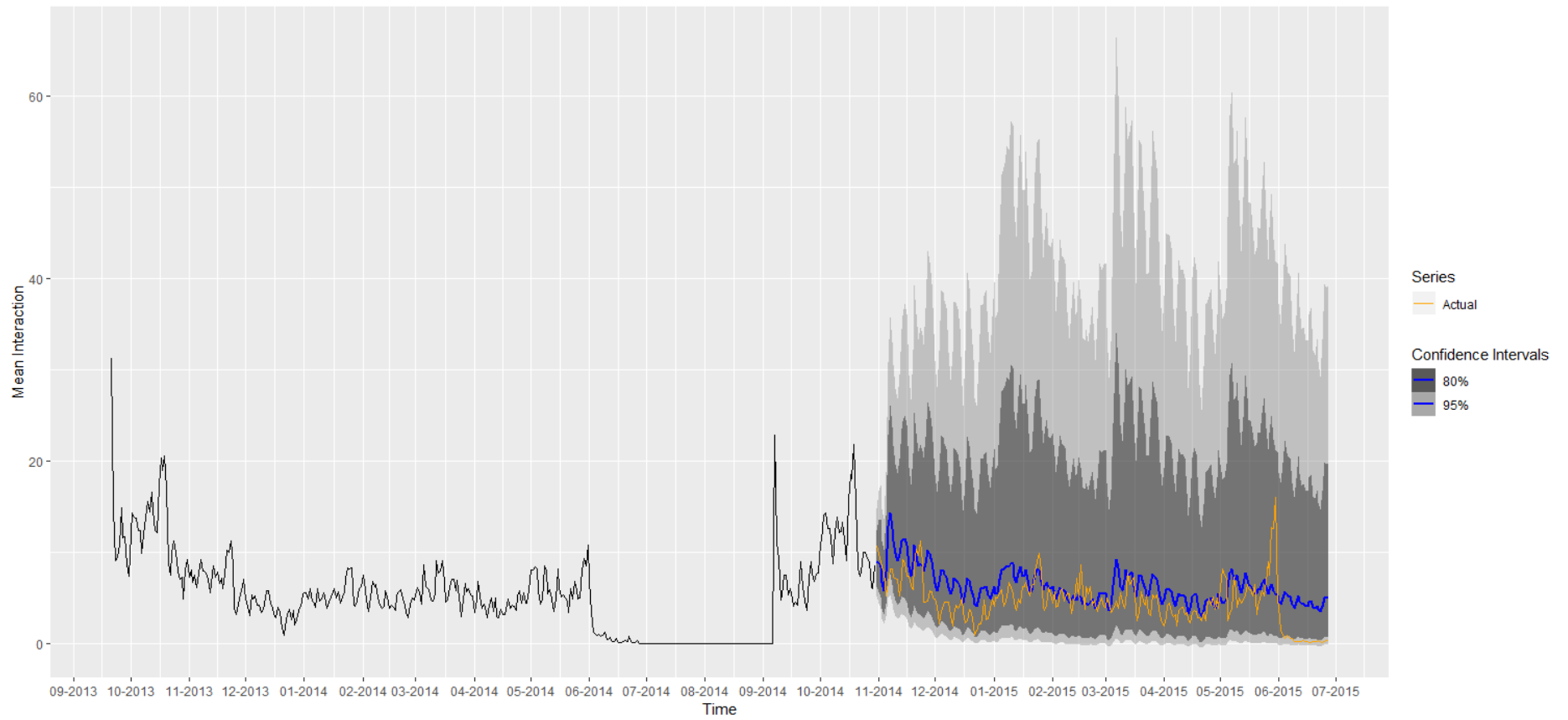


Figure 155. Best forecasting model achieved by using an ARIMA model: “AAA-2013J” and “AAA-2014J” joint time series

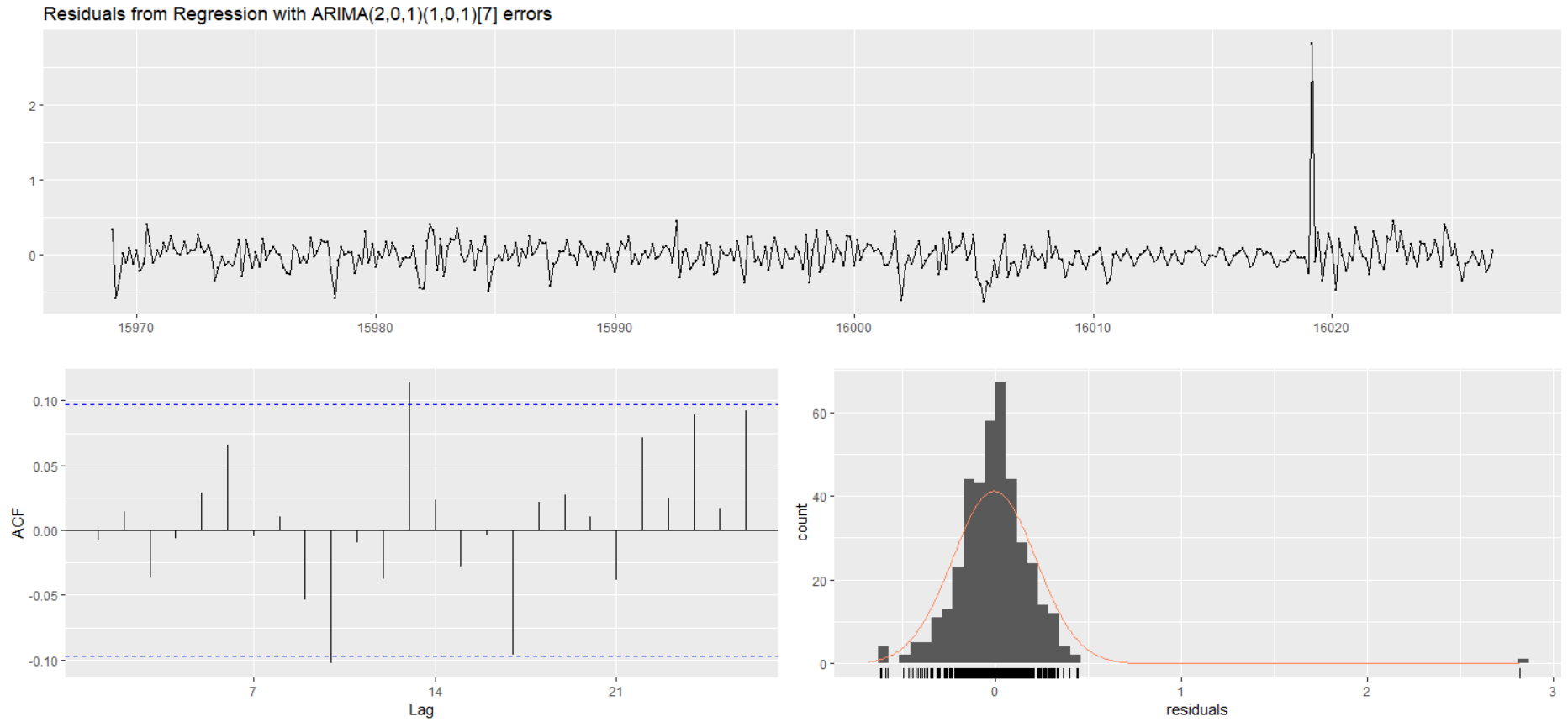


Figure 156. Residuals resulting from the best forecasting model achieved by using an ARIMA model: “AAA-2013J” and “AAA-2014J” joint time series

7.5.1.3 ARIMA model with regressors (Fourier seasonal terms)

ARIMA parameters	Training RMSE	Test RMSE
(2,0,1)	5.504	2.988
(1,1,1)	5.498	3.005

Table 90. Results for ARIMA (with regressors) forecasting process

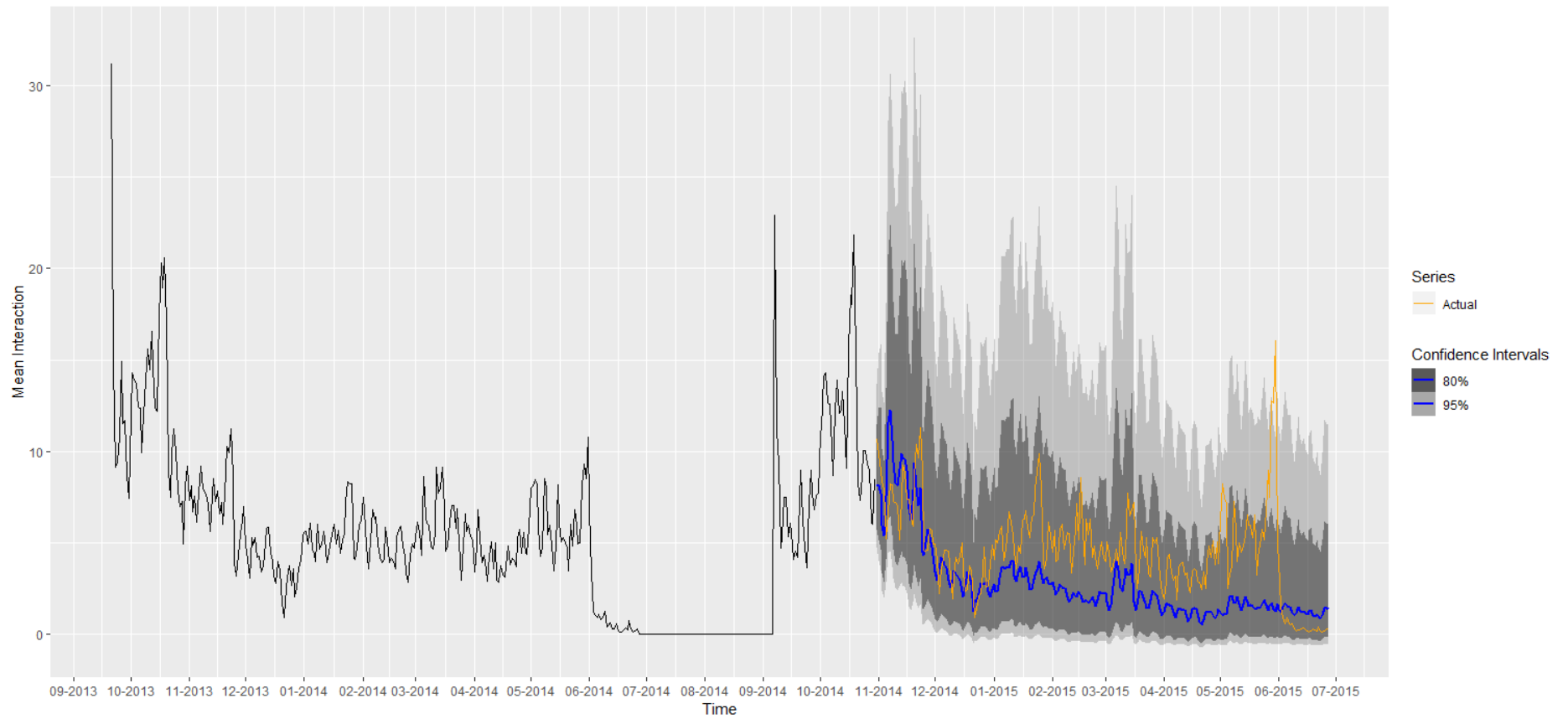


Figure 157. Best forecasting model achieved by using an ARIMA model with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series

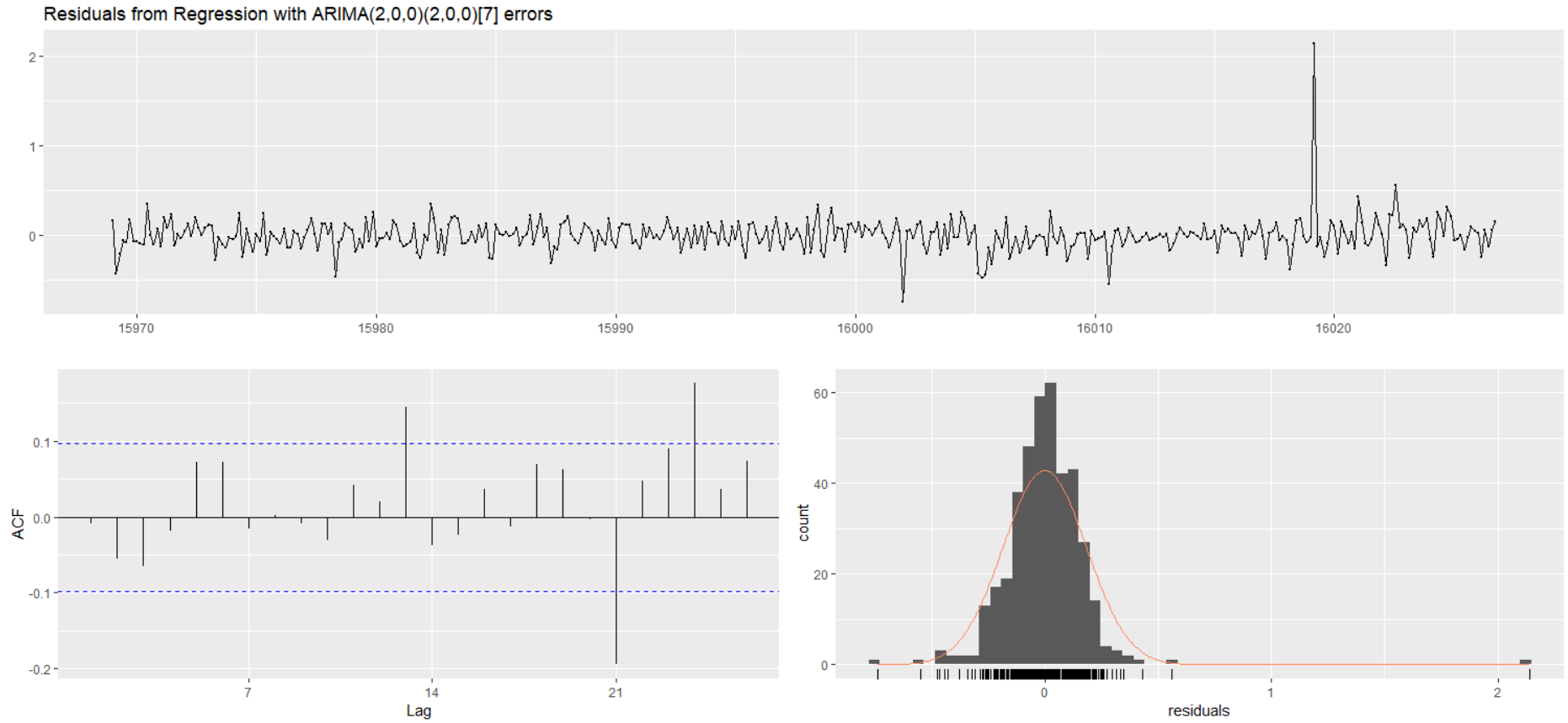


Figure 158. Resulting residuals from the best forecasting model achieved by using an ARIMA model with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series

7.5.1.4 De-seasonalized ARIMA model

ARIMA parameters	Training RMSE	Test RMSE
(1,1,0) (0,0,2)[7]	5.41	6.515
(2,0,1)	5.41	3.466
(2,0,0) (2,0,0)[7]	5.41	3.668
(2,1,1)	5.406	5.807

Table 91. Results for ARIMA (de-seasonalized data) forecasting process

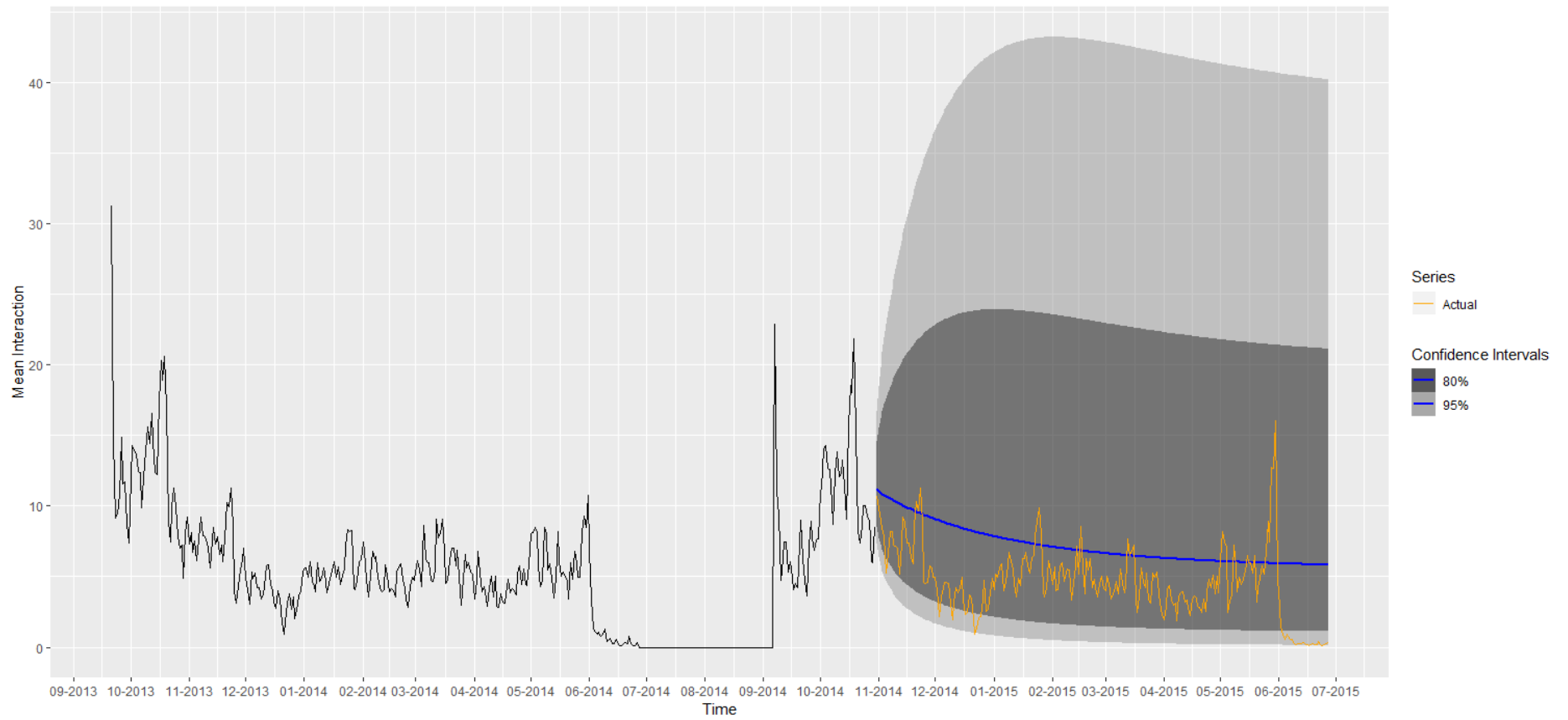


Figure 159. Best forecasting model achieved by using a de-seasonalized ARIMA model: “AAA-2013J” and “AAA-2014J” joint time series

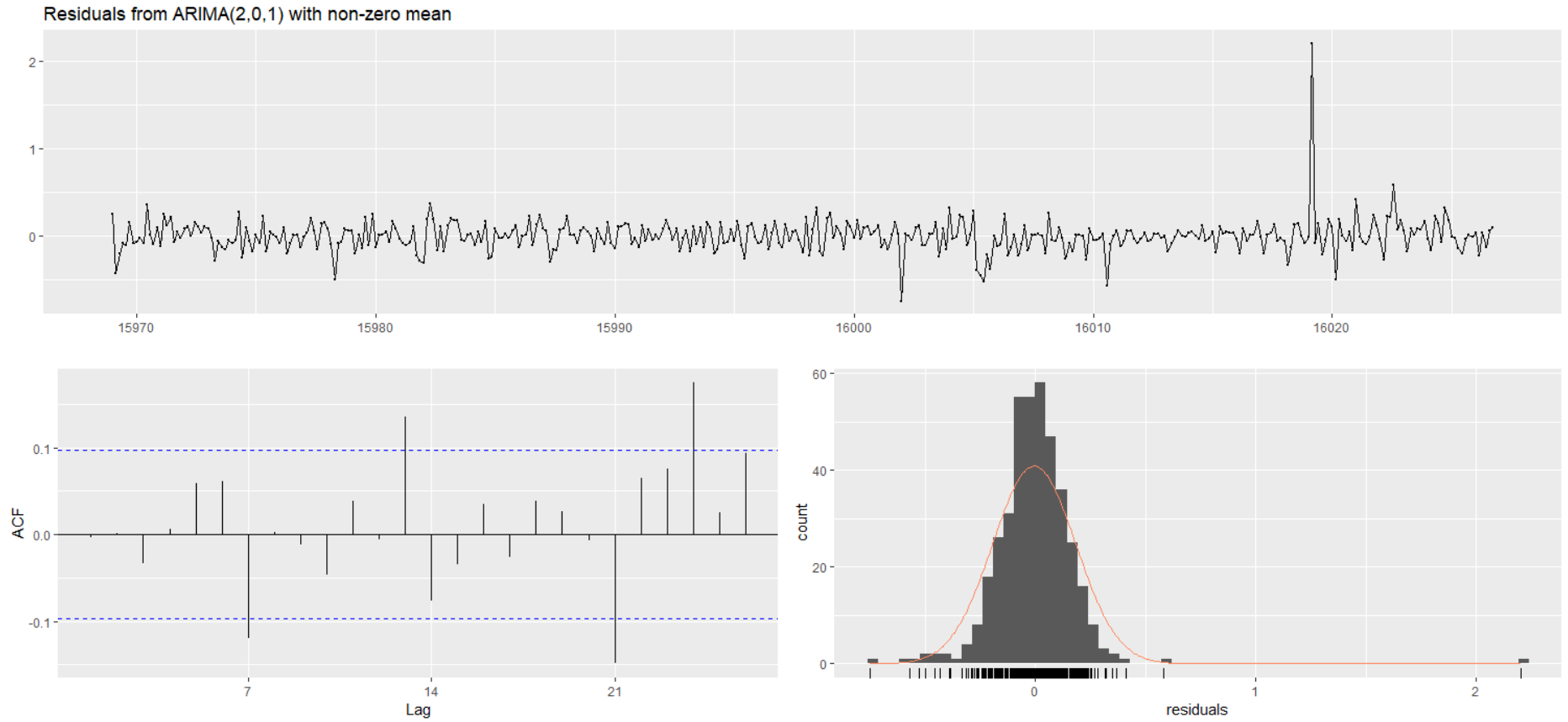


Figure 160. Resulting residuals from the best forecasting model achieved by using a de-seasonalized ARIMA model: “AAA-2013J” and “AAA-2014J” joint time series

7.5.1.5 De-seasonalized ARIMA model with regressors

ARIMA parameters	Training RMSE	Test RMSE
(0,1,0) (0,0,2)[7]	5.417	2.574
(2,0,0) (2,0,0)[7]	5.413	2.505
(2,0,1)	5.414	2.528
(2,1,1)	5.408	4.774

Table 92. Results for ARIMA (de-seasonalized data with regressors) forecasting process

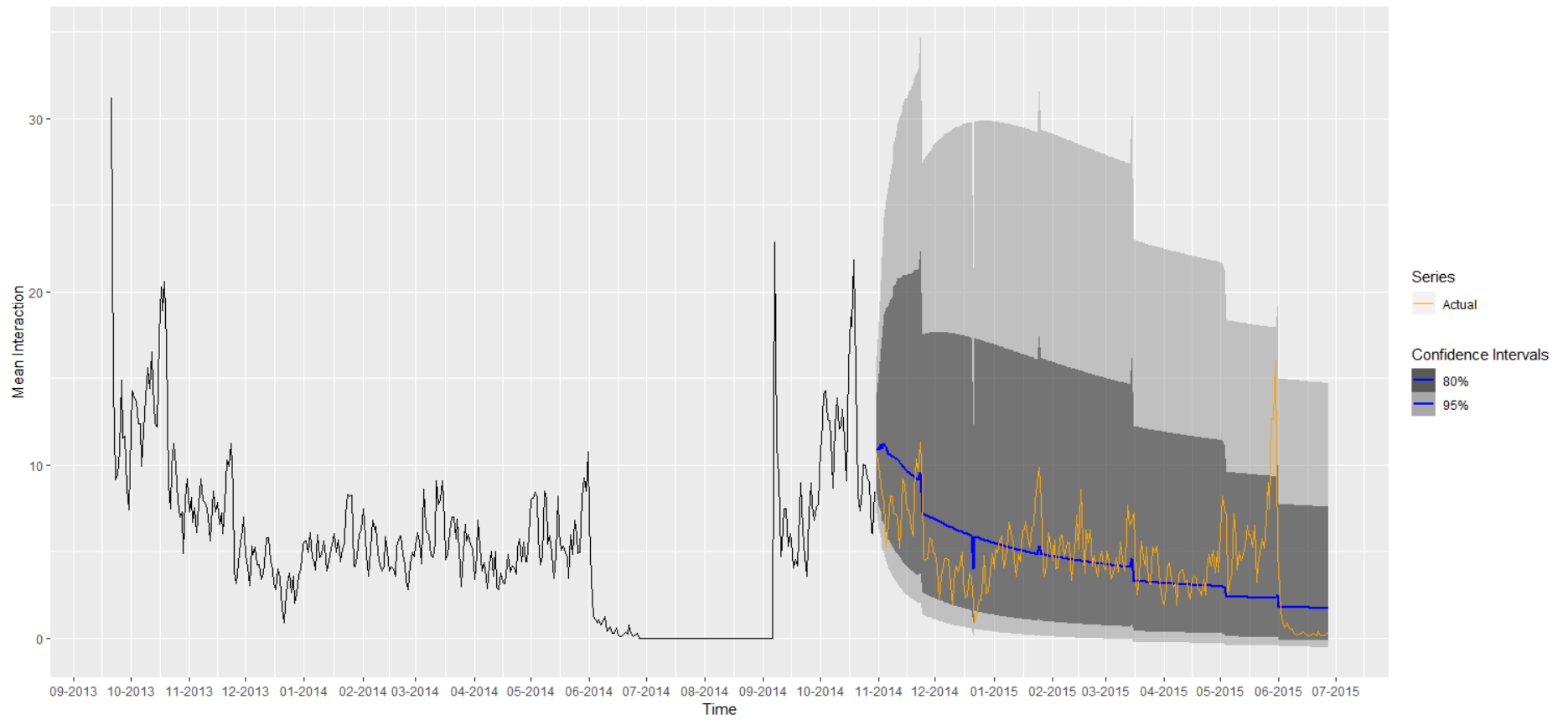


Figure 161. Best forecasting model achieved by using an ARIMA model on de-seasonalized data with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series

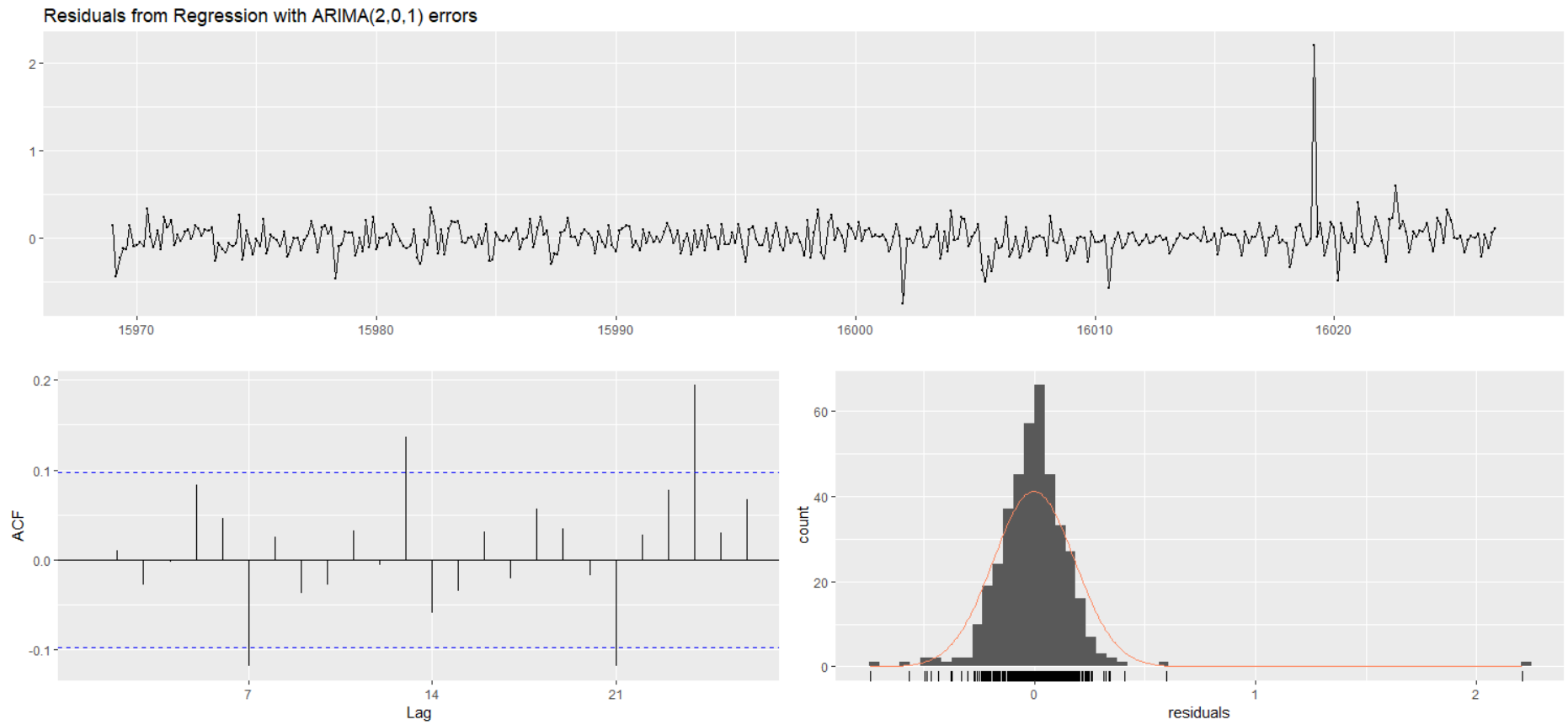


Figure 162. Resulting residuals from the best forecasting model achieved by using an ARIMA model on de-seasonalized data with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series

7.5.1.6 Results' discussion

It is significant to point out that, although non-stationarity was detected during the previous analysis, the best models don't apply any type of differencing. This may be due to some distortion on the KPSS test due to the 0-mean interaction time span between courses, that may have affected the assumption that the mean is not constant throughout the time series (while this is true for it as a whole, mean is most likely to be constant within each course, making it unnecessary to apply differencing for its modelling).

Additionally, it is significant how some well-performing models include weekly seasonal components (ARIMA only allows for the direct treatment of one seasonal component and having 7 days defined as the frequency of the time series makes the algorithm explore this seasonal component). This may be due to some uncaptured seasonality, although it cannot be objectively stated.

Taking these considerations into account, results from the best-performing model (de-seasonalized ARIMA with regressors) are recalled:

ARIMA parameters	Training RMSE	Test RMSE
(0,1,0) (0,0,2)[7]	5.417	2.574
(2,0,0) (2,0,0)[7]	5.413	2.505
(2,0,1)	5.414	2.528
(2,1,1)	5.408	4.774

Table 93. Best ARIMA forecasting results

It has been decided to take the third model as reference for the analysis of performance conducted in the following section. This is, an ARIMA model assuming stationarity of the provided de-seasonalized, thus discarding seasonality.

Although the seasonal model implies a slightly better performance, it was considered to contradict the assumptions implied in a de-seasonalized time series and thus rejected.

7.5.2 Neural Network

7.5.2.1 Neural Network model (Fourier seasonal terms)

Training RMSE	Test RMSE
5.257	2.646

Table 94. Results for regular Neural Network forecasting process

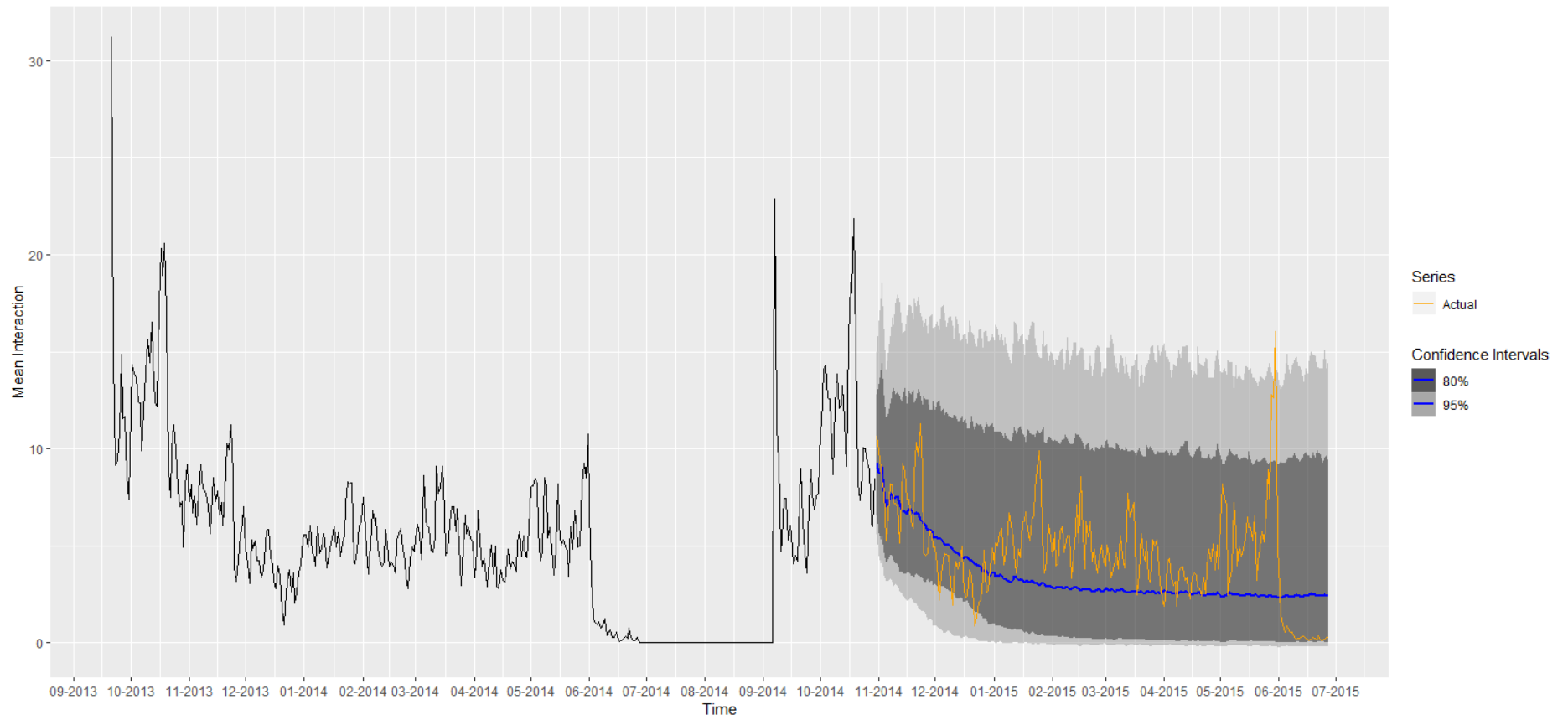


Figure 163. Best forecasting model achieved by using a Neural Network model: “AAA-2013J” and “AAA-2014J” joint time series

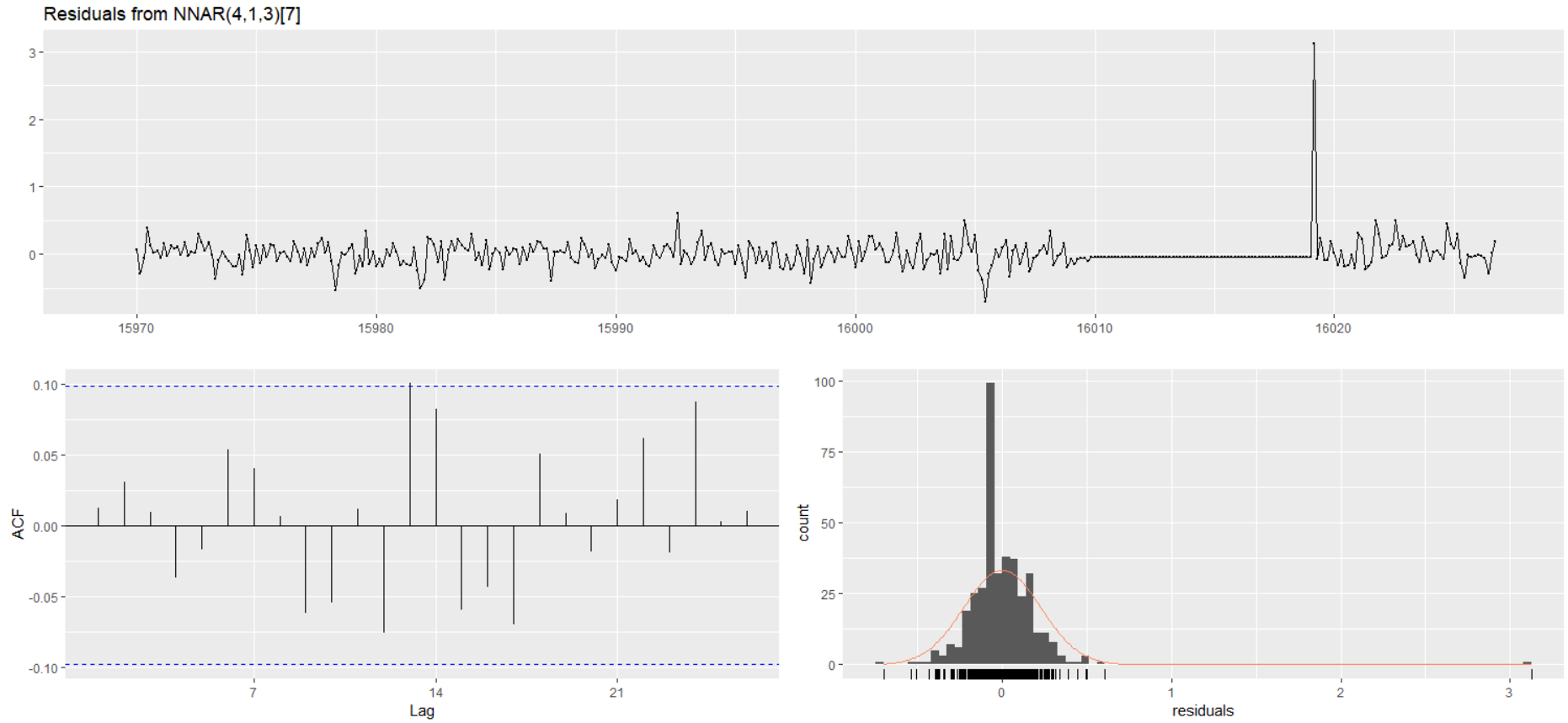


Figure 164. Resulting residuals from the best forecasting model achieved by using a Neural Network model: “AAA-2013J” and “AAA-2014J” joint time series

7.5.2.2 Neural Network model with regressors (Fourier seasonal terms)

Training RMSE	Test RMSE
5.239	2.188

Table 95. Results for Neural Network (with regressors) forecasting process

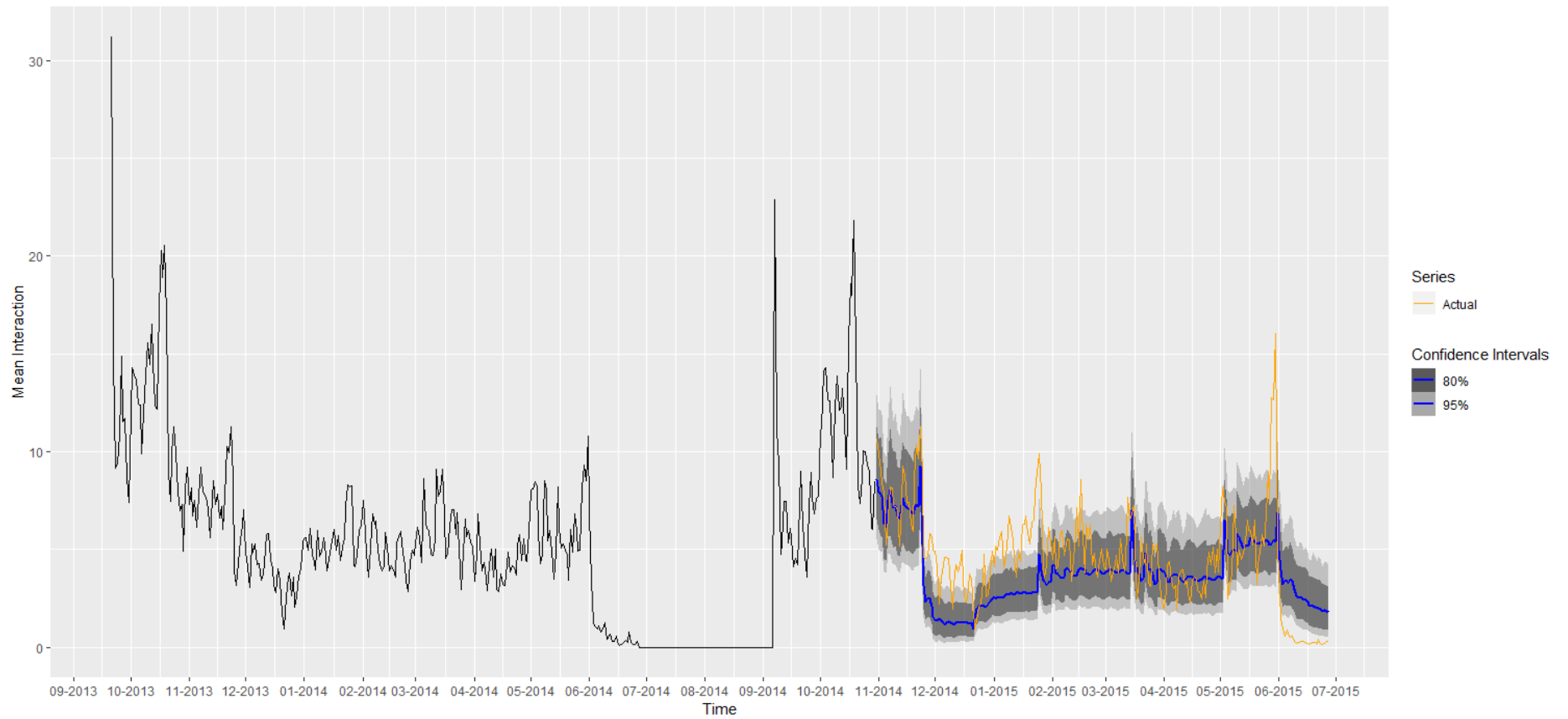


Figure 165. Best forecasting model achieved by using a Neural Network model with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series

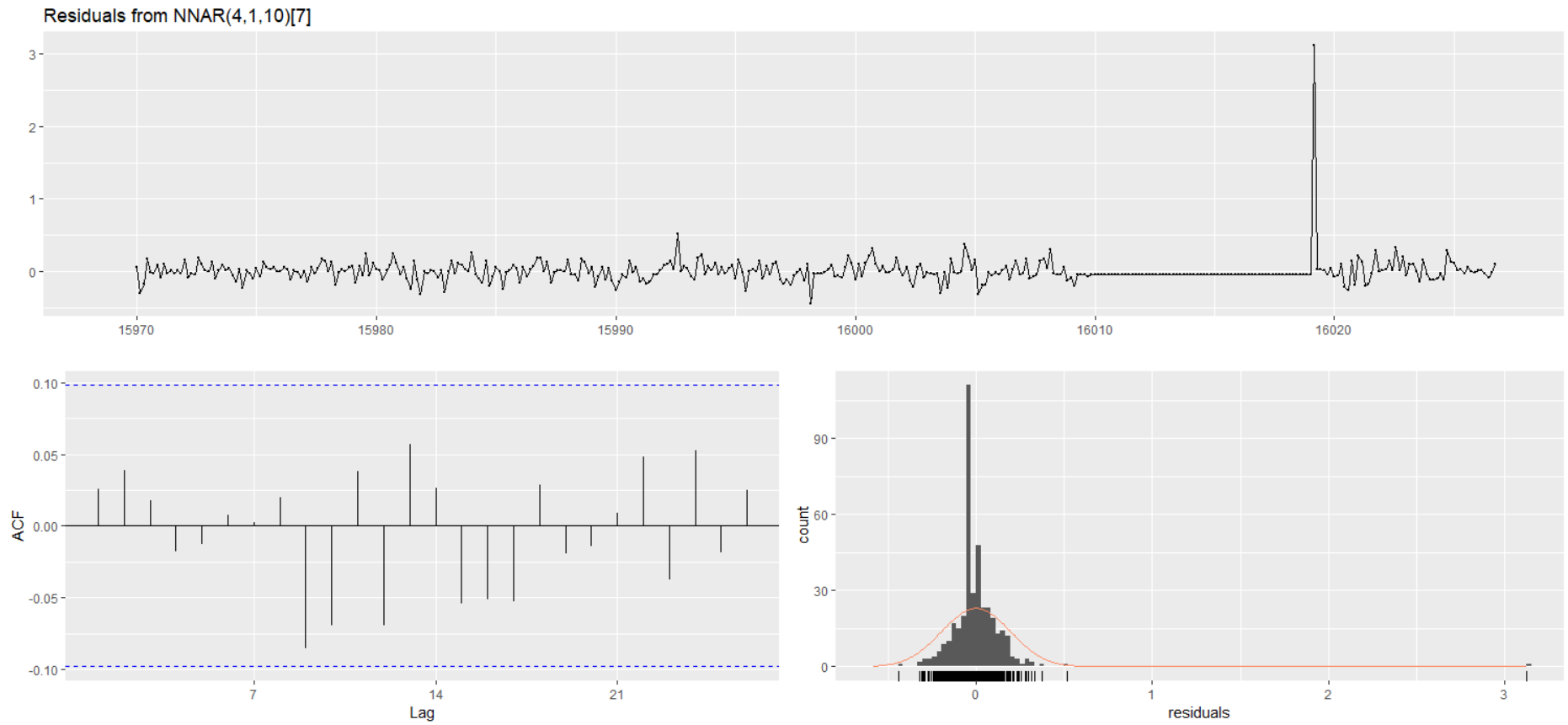


Figure 166. Resulting residuals from the best forecasting model achieved by using a Neural Network model with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series

7.5.2.3 De-seasonalized Neural Network model

Training RMSE	Test RMSE
4.928	3.547

Table 96. Results for Neural Network (de-seasonalized data) forecasting process

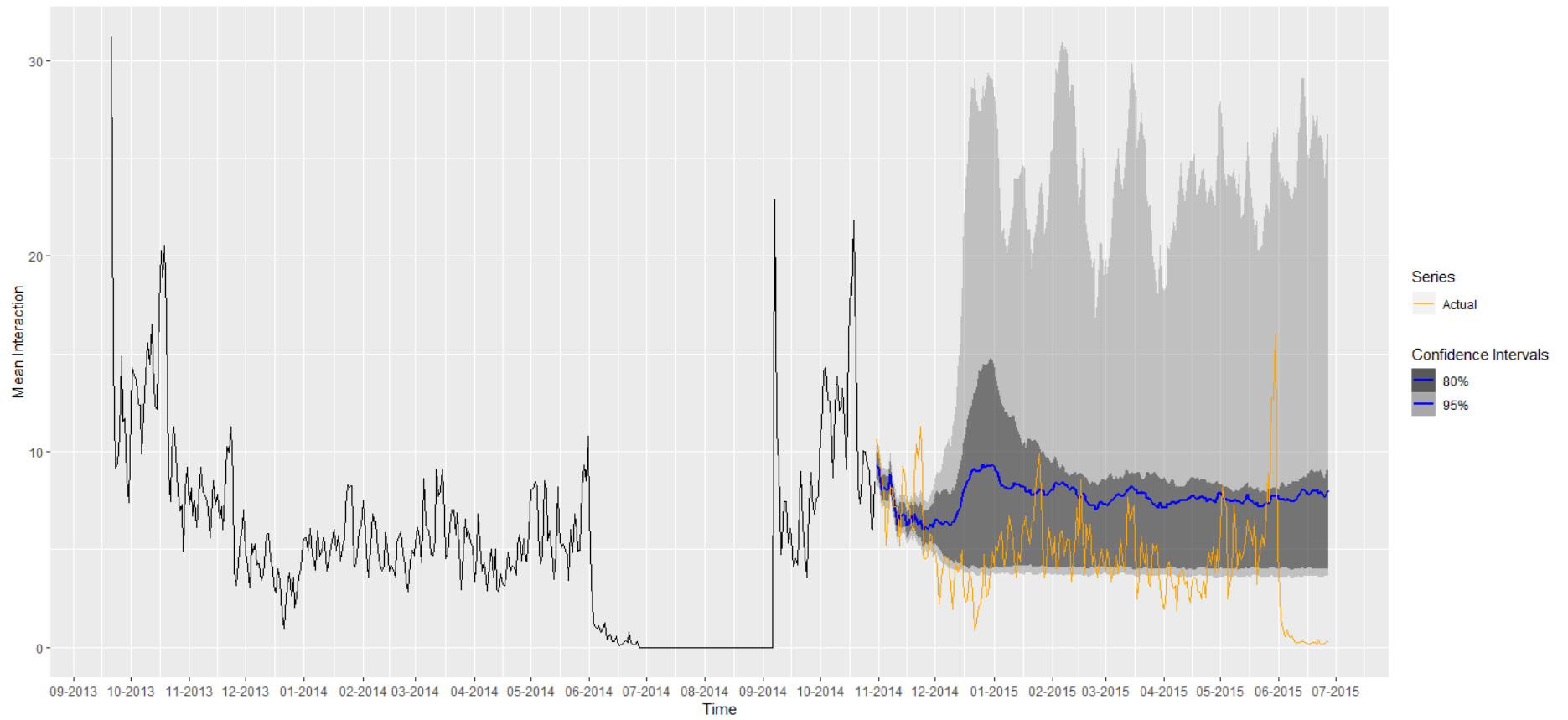


Figure 167. Best forecasting model achieved by using a Neural Network model on de-seasonalized data: “AAA-2013J” and “AAA-2014J” joint time series

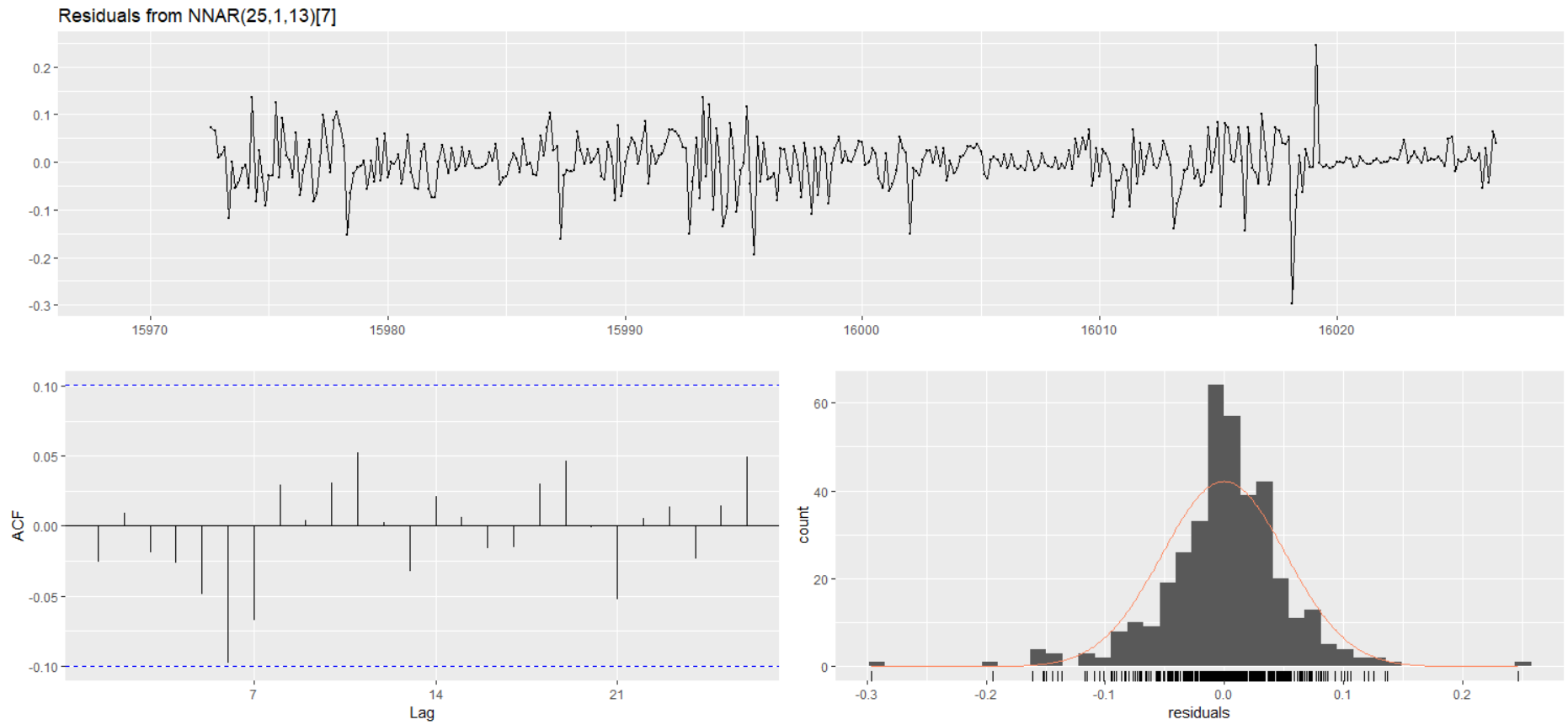


Figure 168. Best forecasting model achieved by using a Neural Network model on de-seasonalized data: “AAA-2013J” and “AAA-2014J” joint time series

7.5.2.4 De-seasonalized Neural Network model with regressors

Training RMSE	Test RMSE
4.924	3.607

Table 97. Results for Neural Network (de-seasonalized data with regressors) forecasting process

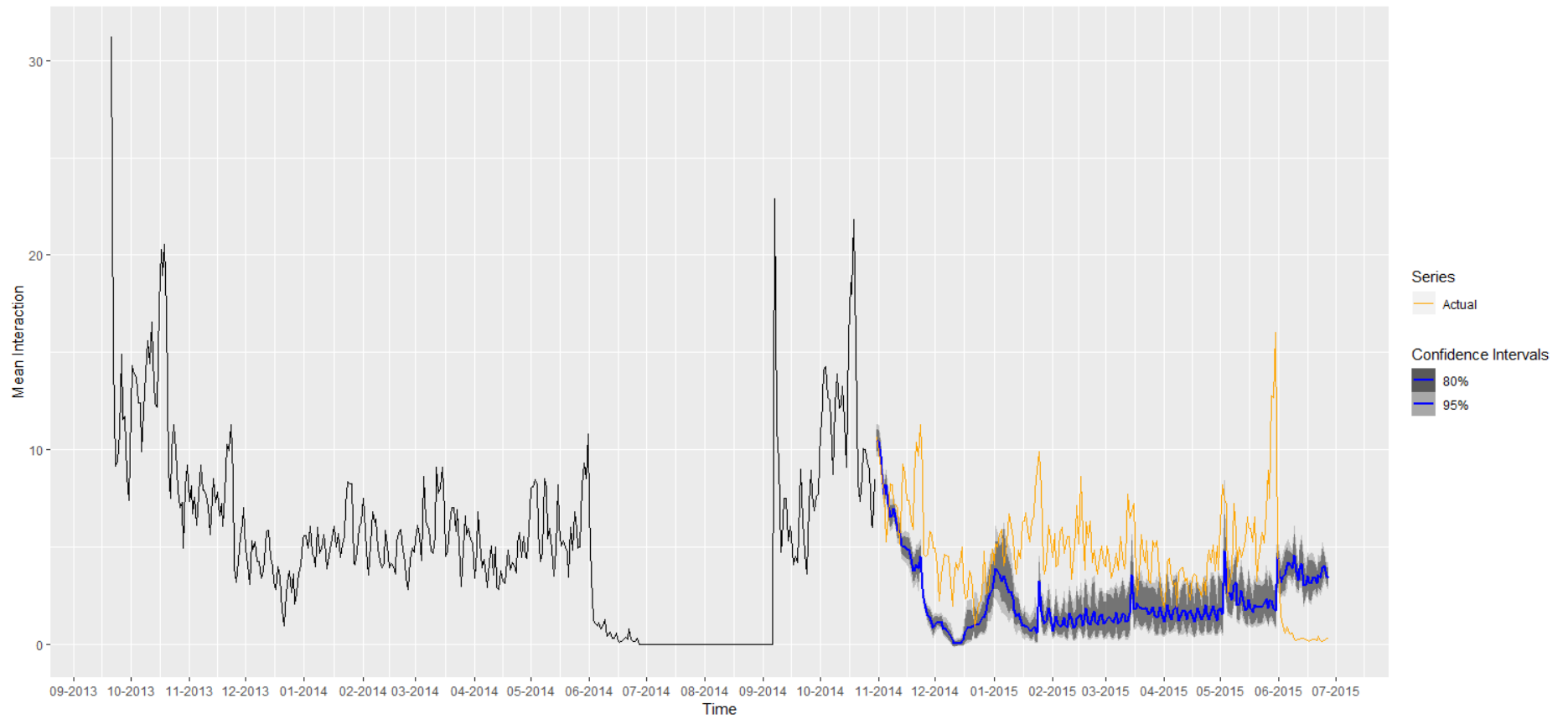


Figure 169. Best forecasting model achieved by using a Neural Network model on de-seasonalized data with regressors (outlier effects): “AAA-2013J” and “AAA-2014J” joint time series

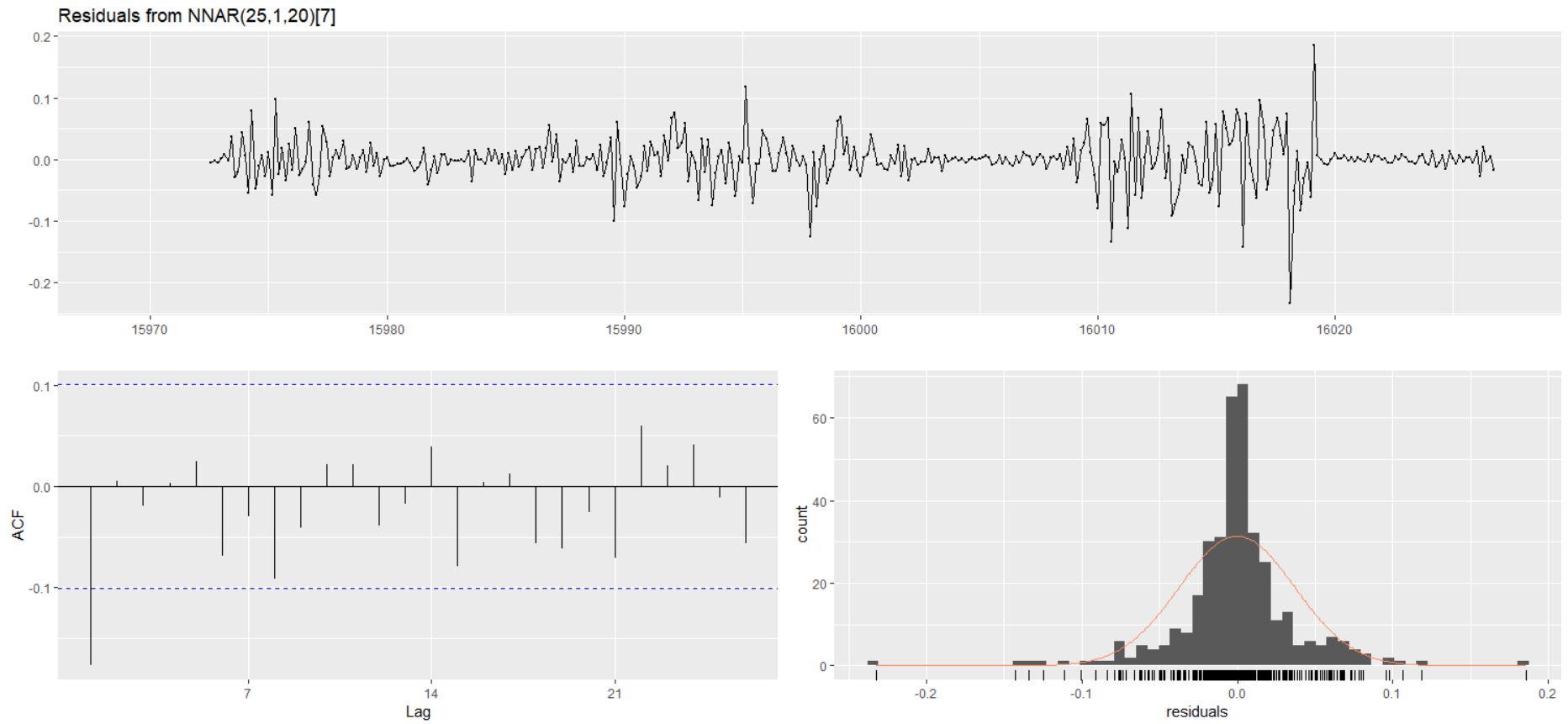


Figure 170. Resulting residuals from the best forecasting model achieved by using a Neural Network model on de-seasonalized data with regressors (outlier effects): "AAA-2013J" and "AAA-2014J" joint time series

7.5.2.5 Results' discussion

It is relevant to point out how de-seasonalized models implied a significant decrease of performance, causing the model to fail to predict irregular variations in the series' shape. This may have to do with how neural networks models are generally "parameter-hungry", in the sense that their performance is significantly improved by external regressors describing the variable to predict (as it can be observed by the significant performance increase for these cases), rather than relying in raw or transformed data with no additional information.

Results from the best-performing model (Neural Network with regressors and Fourier seasonal terms) chosen to conduct the general performance process are presented:

Training RMSE	Test RMSE
5.239	2.188

Table 98. Best Neural Network forecasting results

7.5.3 ETS

Training RMSE	Test RMSE
2.041	2.829

Table 99. Results for ETS forecasting process

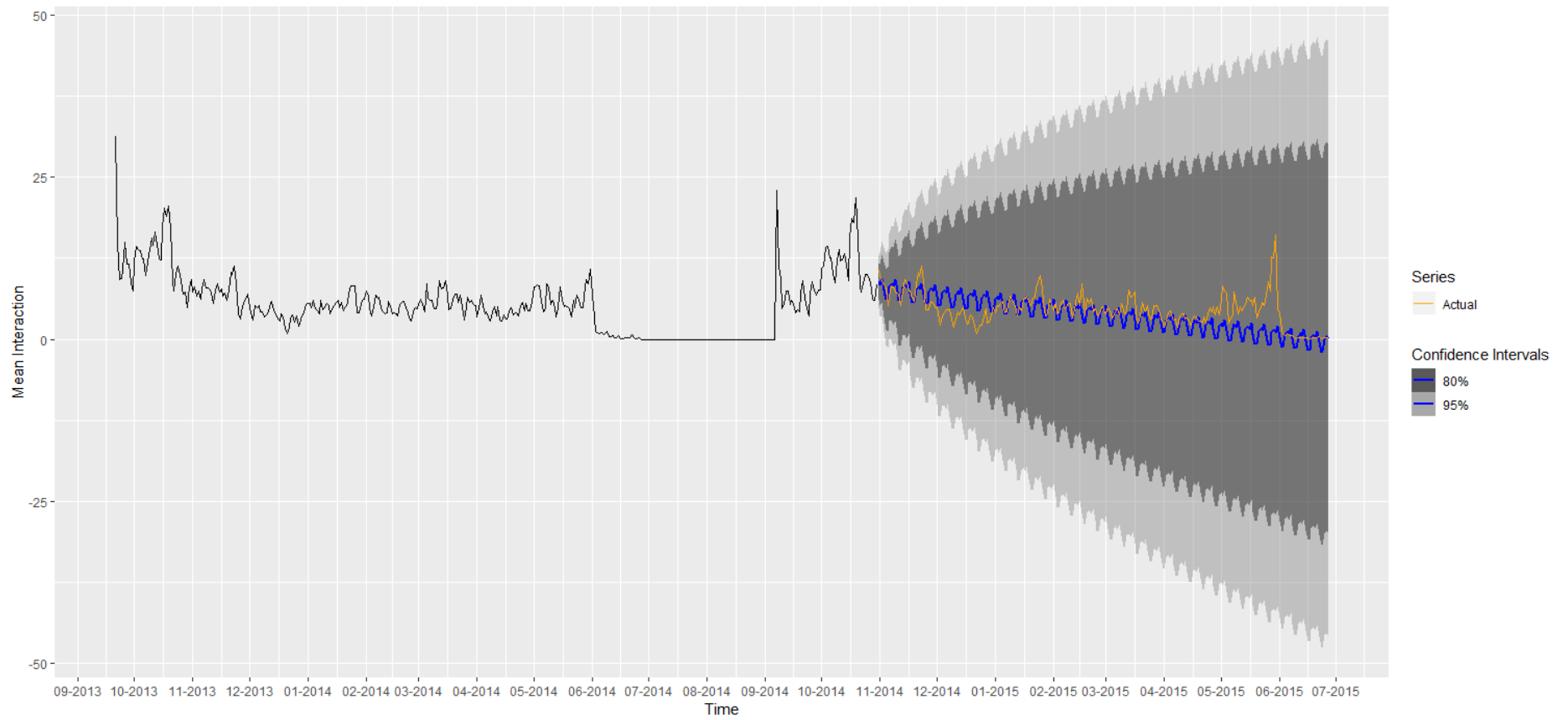


Figure 171. Overfitted forecasting model achieved by using an ETS model: “AAA-2013J” and “AAA-2014J” joint time series

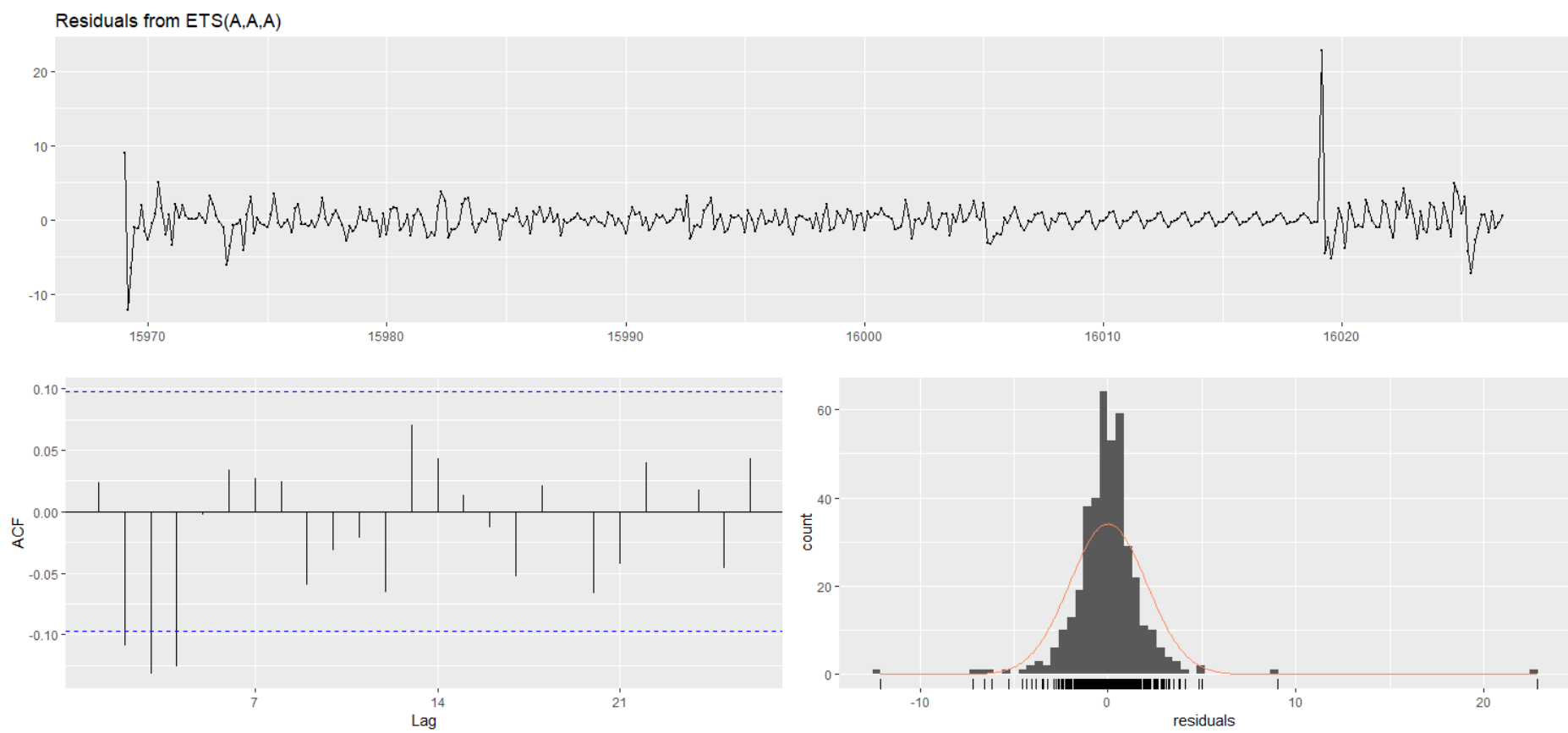


Figure 172. Resulting residuals from the overfitted forecasting model achieved by using an ETS model: “AAA-2013J” and “AAA-2014J” joint time series

7.5.3.1 Results' discussion

Clear overfitting of the model leads to this algorithm being discarded from the general-performance measurement process.

Only results referred to one experiments are presented to avoid redundancy, since every approach resulted in overfitting, which may be caused to the nature of exponential smoothing processes, which give more value to recent observations. This, in a domain in which observations form previous courses are of high relevance, may not be an appropriate approach, causing such undesirable results.

7.5.4 TBATS

Training RMSE	Test RMSE
2.019	6.295

Table 100. Results for TBATS forecasting process

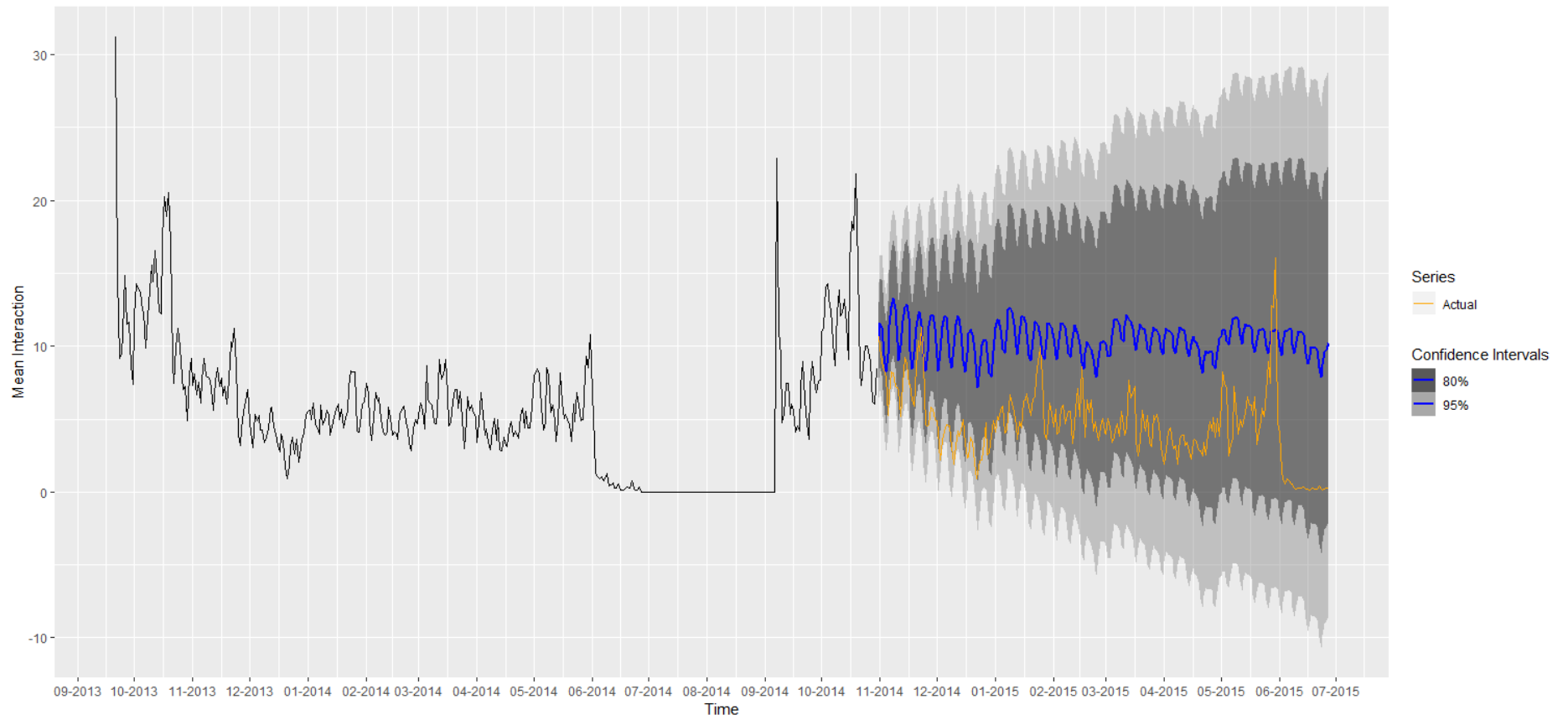


Figure 173. Overfitted forecasting model achieved by using a TBATS model: “AAA-2013J” and “AAA-2014J” joint time series

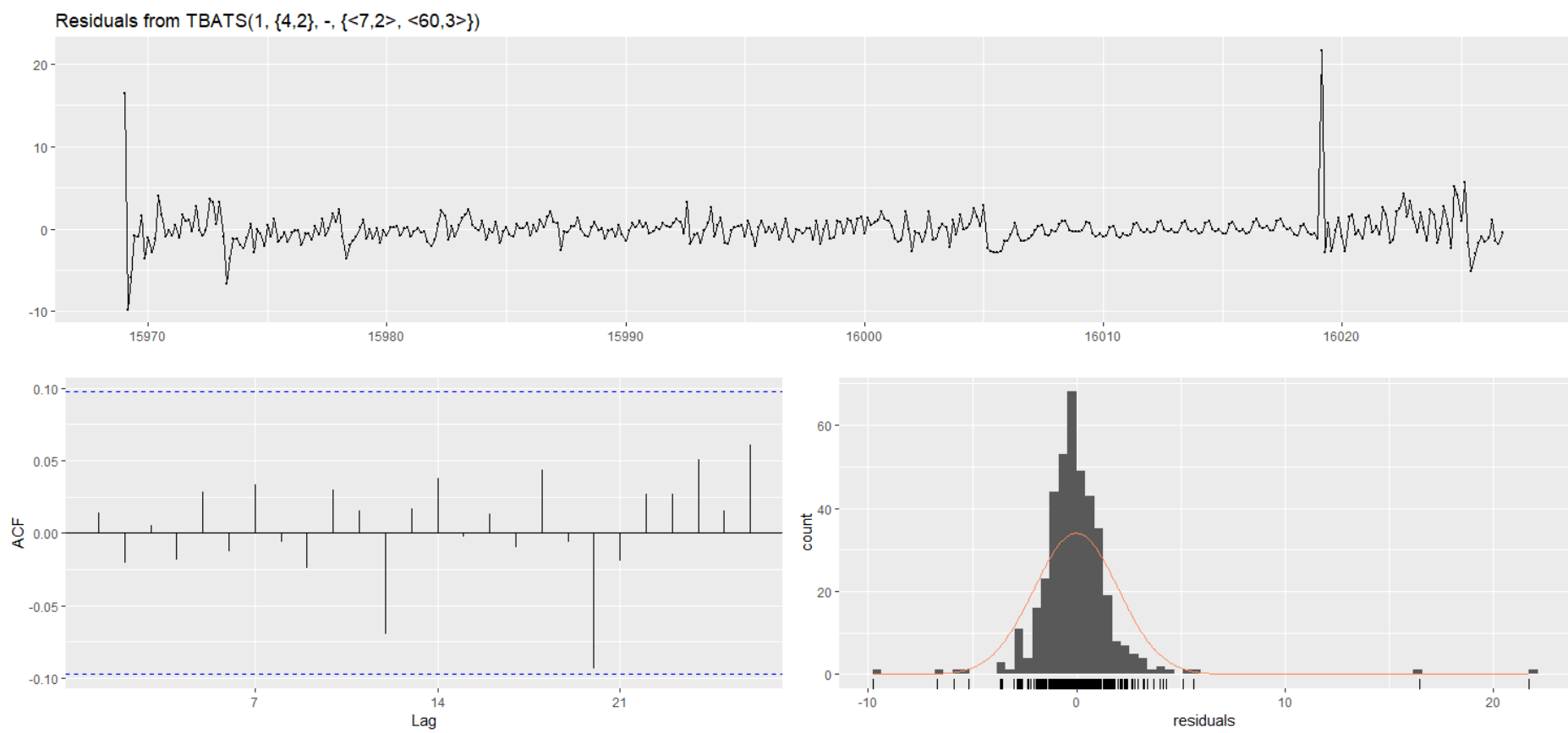


Figure 174. Resulting residuals from the overfitted forecasting model achieved by using a TBATS model: “AAA-2013J” and “AAA-2014J” joint time series

7.5.4.1 Results' discussion

Same event as observed with ETS occurred: overfitting of the model leads to this algorithm being discarded from the general-performance measurement process.

Observations previously made to the potential un-appropriateness of exponential smoothing algorithms for modelling time series of this specific domain are reinforced by this occurrence.

7.5.5 Combination

Models involved in this combination process:

- De-seasonalized ARIMA model with regressors
- Neural Network with regressors

Training RMSE	Test RMSE
5.218	1.918

Table 101. Combination of best models' forecasting results

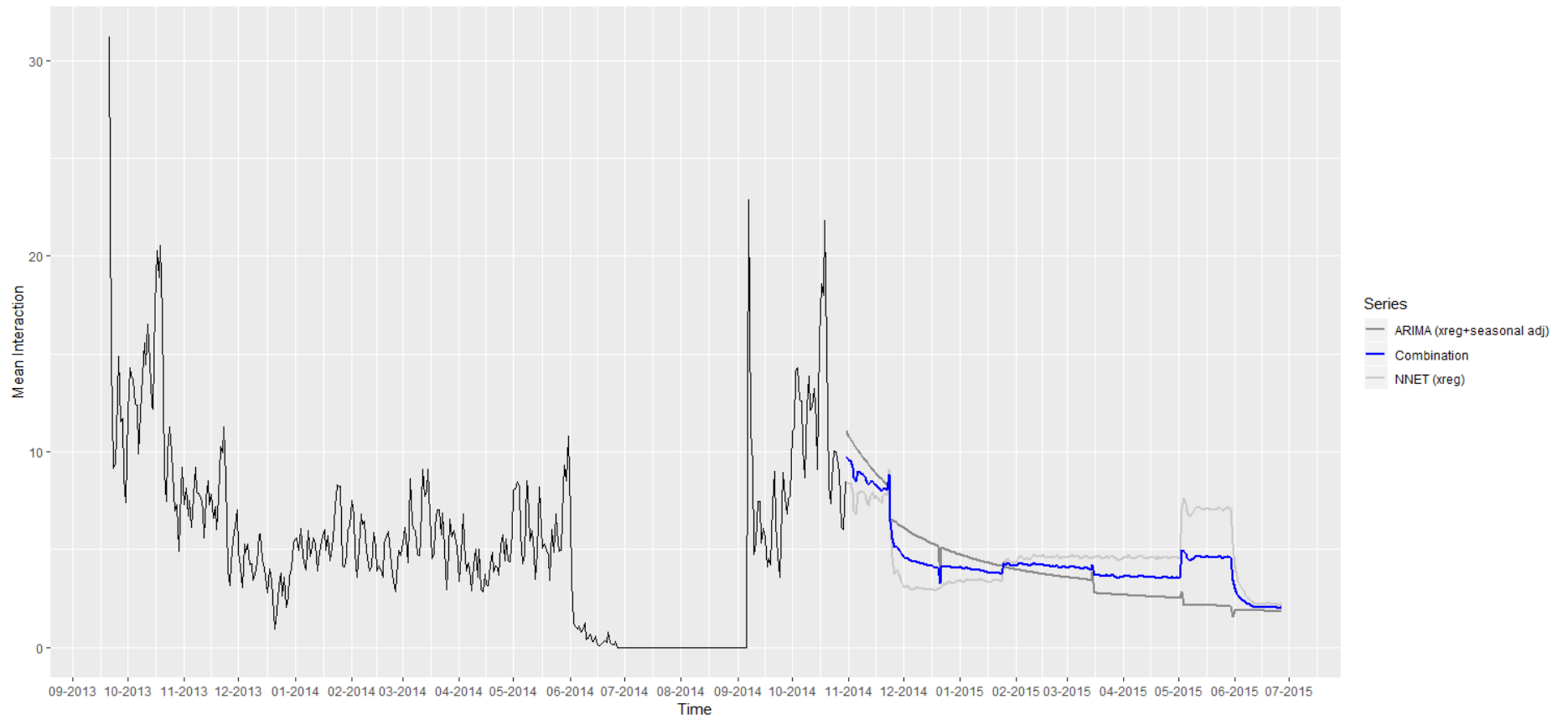


Figure 175. Forecasting model achieved by merging (mean) both best ARIMA and Neural Network models: “AAA-2013J” and “AAA-2014J” joint time series

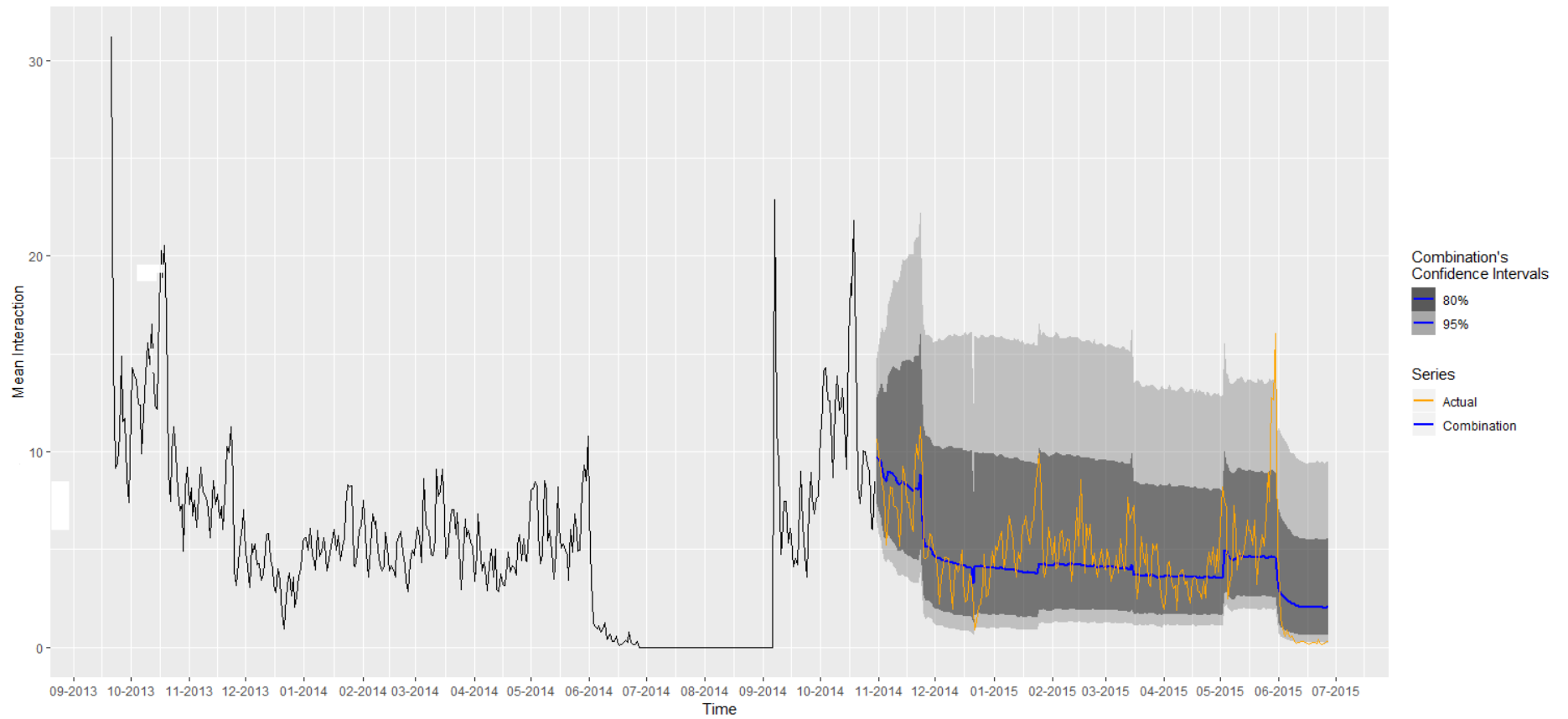


Figure 176. Residuals from the forecasting model achieved by merging (mean) both best ARIMA and Neural Network models: “AAA-2013J” and “AAA-2014J” joint time series

7.5.5.1 Results' discussion

Although this model cannot be considered as deterministic and may be volatile due to the sensitivity of the mean statistic to potential unprecise predictions, appropriateness of the results observed led to its inclusion in the general-performance measurement process.

7.5.6 Results

Course's Module	Presentation group	Years comprised
AAA	J	2013-2014
BBB	B	2013-2014
BBB	J	2013-2014
CCC	B	2014
CCC	J	2014
DDD	B	2013-2014
DDD	J	2013-2014
EEE	B	2014
EEE	J	2013-2014
FFF	B	2013-2014
FFF	J	2013-2014
GGG	B	2014
GGG	J	2013-2014

Table 102. Time series' re-arrangement (forecasting tasks)

7.5.6.1 AAA-J (2013-2014)

7.5.6.1.1 ARIMA

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	4.622966	4.622966	4.622966	4.622966	4.622966	4.622966	4.622966	4.622966
	Ts	6.411908	5.445453	4.547138	4.464047	4.336945	4.222816	4.026312	4.779231
Month 2	Tr	5.415769	5.415769	5.415769	5.415769	5.415769	5.415769	-	5.415769
	Ts	2.279354	2.154499	2.20826	2.186531	2.207926	2.103798	-	2.190062
Month 3	Tr	5.294395	5.294395	5.294395	5.294395	5.294395	-	-	5.294395
	Ts	1.033291	2.311658	2.483712	2.562784	2.438019	-	-	2.165893
Month 4	Tr	5.166084	5.166084	5.166084	5.166084	-	-	-	5.166084
	Ts	2.617952	2.54195	2.557779	2.37876	-	-	-	2.52411
Month 5	Tr	5.092491	5.092491	5.092491	-	-	-	-	5.092491
	Ts	1.803806	2.013659	1.858166	-	-	-	-	1.891877
Month 6	Tr	5.022135	5.022135	-	-	-	-	-	5.022135
	Ts	1.512042	1.297937	-	-	-	-	-	1.404989
Month 7	Tr	4.907593	-	-	-	-	-	-	4.907593
	Ts	1.272781	-	-	-	-	-	-	1.272781

Table 103. ARIMA forecasting results for course AAA-J (2013-2014)

- **Training average:** 5.079322
- **Test average:** 2.759974

7.5.6.1.2 Neural Network

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	4.274132	4.274274	4.274351	4.274358	4.274079	4.274602	4.274115	4.274273
	Ts	8.030832	7.040042	5.911615	4.65275	5.155791	4.526194	4.25749	5.653531
Month 2	Tr	5.251755	5.250896	5.250653	5.251729	5.250728	5.250821	-	5.251097
	Ts	1.923806	1.08014	2.419635	2.775535	3.730843	1.586689	-	2.252775
Month 3	Tr	4.978608	4.978281	4.978366	4.978012	4.978677	-	-	4.978389
	Ts	1.160773	1.639813	1.525492	4.006142	3.144747	-	-	2.295393
Month 4	Tr	4.837243	4.837647	4.83697	4.837494	-	-	-	4.837338
	Ts	3.020741	2.595681	2.274837	1.956267	-	-	-	2.461881
Month 5	Tr	4.789397	4.789468	4.788837	-	-	-	-	4.789234
	Ts	1.136883	1.608045	1.20051	-	-	-	-	1.315146
Month 6	Tr	4.564125	4.563998	-	-	-	-	-	4.564061
	Ts	1.160015	1.528697	-	-	-	-	-	1.344356
Month 7	Tr	4.493655	-	-	-	-	-	-	4.493655
	Ts	1.246787	-	-	-	-	-	-	1.246787

Table 104. Neural Network forecasting results for course AAA-J (2013-2014)

- **Training average:** 4.773474
- **Test average:** 2.939171

7.5.6.1.3 Combination

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	4.113309	4.113436	4.11358	4.11354	4.113272	4.113691	4.113341	4.113453
	Ts	7.145984	6.165756	5.119958	4.231926	4.626094	3.962259	3.589332	4.97733
Month 2	Tr	5.225408	5.224509	5.224231	5.225245	5.2245	5.22473	-	5.22477
	Ts	1.818775	1.510602	1.959168	2.274402	2.660436	1.592686	-	1.969345
Month 3	Tr	4.907149	4.906767	4.906875	4.906559	4.90719	-	-	4.906908
	Ts	0.972327	1.760205	1.679926	1.650216	1.480102	-	-	1.508555
Month 4	Tr	4.778261	4.778608	4.777946	4.778493	-	-	-	4.778327
	Ts	2.803172	2.512873	2.316955	1.95538	-	-	-	2.397095
Month 5	Tr	4.71833	4.718392	4.717818	-	-	-	-	4.71818
	Ts	1.255329	1.347036	1.259415	-	-	-	-	1.28726
Month 6	Tr	4.48336	4.483274	-	-	-	-	-	4.483317
	Ts	1.180635	1.082098	-	-	-	-	-	1.131367
Month 7	Tr	4.395525	-	-	-	-	-	-	4.395525
	Ts	0.812791	-	-	-	-	-	-	0.812791

Table 105. Combination forecasting results for course AAA-J (2013-2014)

- **Training average:** 4.689548
- **Test average:** 2.525923

7.5.6.2 BBB-B (2013-2014)

7.5.6.2.1 ARIMA

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	1.843458	1.843458	1.843458	1.843458	1.843458	1.843458	1.843458
	Ts	0.901344	1.043367	1.136615	1.277712	1.549629	1.651926	1.260099
Month 2	Tr	1.828891	1.828891	1.828891	1.828891	1.828891	-	1.828891
	Ts	0.932405	0.949736	1.035513	1.320913	1.397962	-	1.127306
Month 3	Tr	1.795787	1.795787	1.795787	1.795787	-	-	1.795787
	Ts	0.805288	0.72753	1.055592	1.085791	-	-	0.91855
Month 4	Tr	1.758636	1.758636	1.758636	-	-	-	1.758636
	Ts	0.801121	1.30009	1.312765	-	-	-	1.137992
Month 5	Tr	1.714086	1.714086	-	-	-	-	1.714086
	Ts	1.734733	1.532486	-	-	-	-	1.63361
Month 6	Tr	1.679069	-	-	-	-	-	1.679069
	Ts	0.994241	-	-	-	-	-	0.994241

Table 106. ARIMA forecasting results for course BBB-B (2013-2014)

- **Training average:** 1.798643
- **Test average:** 1.168893

7.5.6.2.2 Neural Network

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	1.740855	1.740939	1.740999	1.740811	1.740843	1.740825	1.740879
	Ts	0.849021	0.889626	1.445461	1.289812	1.150256	1.011026	1.105867
Month 2	Tr	1.745365	1.745164	1.745115	1.74533	1.745202	-	1.745235
	Ts	1.005137	1.24717	0.903961	1.205599	1.141896	-	1.100752
Month 3	Tr	1.717673	1.717771	1.717569	1.717535	-	-	1.717637
	Ts	1.337204	0.575867	1.193826	0.889838	-	-	0.999184
Month 4	Tr	1.682253	1.682369	1.682232	-	-	-	1.682285
	Ts	0.245945	0.40902	0.932539	-	-	-	0.529168
Month 5	Tr	1.635345	1.635453	-	-	-	-	1.635399
	Ts	0.495841	1.036739	-	-	-	-	0.76629
Month 6	Tr	1.607071	-	-	-	-	-	1.607071
	Ts	0.600971	-	-	-	-	-	0.600971

Table 107. Neural Network forecasting results for course BBB-B (2013-2014)

- **Training average:** 1.712701
- **Test average:** 0.94556

7.5.6.2.3 Combination

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	1.659866	1.659934	1.659984	1.659836	1.659842	1.659829	1.659882
	Ts	0.683572	0.67626	0.824243	0.733864	0.84326	0.877914	0.773185
Month 2	Tr	1.660029	1.659849	1.659799	1.660008	1.659925	-	1.659922
	Ts	0.894491	0.828635	0.71872	0.826461	0.851837	-	0.824029
Month 3	Tr	1.632981	1.633025	1.632861	1.632837	-	-	1.632926
	Ts	0.981603	0.578256	1.005736	0.788044	-	-	0.83841
Month 4	Tr	1.600281	1.600364	1.600261	-	-	-	1.600302
	Ts	0.418595	0.791511	0.969106	-	-	-	0.726404
Month 5	Tr	1.557548	1.557654	-	-	-	-	1.557601
	Ts	0.970687	1.109947	-	-	-	-	1.040317
Month 6	Tr	1.528774	-	-	-	-	-	1.528774
	Ts	0.703688	-	-	-	-	-	0.703688

Table 108. Combination forecasting results for course BBB-B (2013-2014)

- **Training average:** 1.630261
- **Test average:** 0.813163

7.5.6.3 BBB-J (2013-2014)

7.5.6.3.1 ARIMA

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	1.282473	1.282473	1.282473	1.282473	1.282473	1.282473	1.282473	1.282473
	Ts	1.866167	2.057745	1.794976	2.155354	2.321914	2.271898	2.283087	2.107306
Month 2	Tr	1.387092	1.387092	1.387092	1.387092	1.387092	1.387092	-	1.387092
	Ts	1.330737	1.204219	1.972761	2.260976	2.222366	2.272298	-	1.877226
Month 3	Tr	1.42378	1.42378	1.42378	1.42378	1.42378	-	-	1.42378
	Ts	1.058492	2.219915	2.507083	2.415374	2.441859	-	-	2.128545
Month 4	Tr	1.397705	1.397705	1.397705	1.397705	-	-	-	1.397705
	Ts	1.633913	2.017846	1.880257	1.9582	-	-	-	1.872554
Month 5	Tr	1.460033	1.460033	1.460033	-	-	-	-	1.460033
	Ts	1.683607	1.532135	1.72726	-	-	-	-	1.647667
Month 6	Tr	1.51144	1.51144	-	-	-	-	-	1.51144
	Ts	1.298232	1.300186	-	-	-	-	-	1.299209
Month 7	Tr	1.49321	-	-	-	-	-	-	1.49321
	Ts	1.689272	-	-	-	-	-	-	1.689272

Table 109. ARIMA forecasting results for course BBB-J (2013-2014)

- **Training average:** 1.389492
- **Test average:** 1.906362

7.5.6.3.2 Neural Network

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	1.475589	1.474856	1.474717	1.474841	1.475254	1.475316	1.473868	1.47492
	Ts	1.356493	1.081852	1.051412	1.23863	1.175674	1.347442	1.118781	1.195755
Month 2	Tr	1.521921	1.521999	1.522012	1.521826	1.521906	1.522063	-	1.521954
	Ts	1.233822	0.900197	1.406823	1.353617	1.194365	1.58265	-	1.278579
Month 3	Tr	1.550556	1.550511	1.550467	1.550579	1.550473	-	-	1.550517
	Ts	0.664603	1.389016	1.201106	1.138451	1.389657	-	-	1.156567
Month 4	Tr	1.515747	1.515818	1.515852	1.515804	-	-	-	1.515805
	Ts	1.799956	2.3101	2.026616	2.032062	-	-	-	2.042184
Month 5	Tr	1.574318	1.57433	1.574504	-	-	-	-	1.574384
	Ts	1.217409	1.074803	1.366189	-	-	-	-	1.219467
Month 6	Tr	1.608394	1.608619	-	-	-	-	-	1.608507
	Ts	1.538254	2.098579	-	-	-	-	-	1.818416
Month 7	Tr	1.597423	-	-	-	-	-	-	1.597423
	Ts	1.571373	-	-	-	-	-	-	1.571373

Table 110. Neural Network forecasting results for course BBB-J (2013-2014)

- **Training average:** 1.528913
- **Test average:** 1.387855

7.5.6.3.3 Combination

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	1.34869	1.347908	1.34774	1.34785	1.348166	1.348302	1.347115	1.347967
	Ts	1.463779	1.351252	1.292495	1.403579	1.433366	1.325458	1.416125	1.383722
Month 2	Tr	1.428219	1.428285	1.428289	1.428147	1.428207	1.428351	-	1.42825
	Ts	1.25193	0.946503	1.618085	1.755968	1.665333	1.893363	-	1.521863
Month 3	Tr	1.46333	1.463297	1.463267	1.46338	1.463243	-	-	1.463304
	Ts	0.473191	1.725872	1.774623	1.714584	1.846086	-	-	1.506871
Month 4	Tr	1.432278	1.432343	1.432351	1.432334	-	-	-	1.432327
	Ts	1.70794	2.148385	1.943949	1.984583	-	-	-	1.946214
Month 5	Tr	1.491572	1.491563	1.491783	-	-	-	-	1.49164
	Ts	1.278724	1.141765	1.489961	-	-	-	-	1.303483
Month 6	Tr	1.536909	1.537114	-	-	-	-	-	1.537012
	Ts	0.742043	1.370523	-	-	-	-	-	1.056283
Month 7	Tr	1.521	-	-	-	-	-	-	1.521
	Ts	1.626186	-	-	-	-	-	-	1.626186

Table 111. Combination forecasting results for course BBB-J (2013-2014)

- **Training average:** 1.432894
- **Test average:** 1.492345

7.5.6.4 CCC-B (2014)

7.5.6.4.1 ARIMA

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	5.247738	5.247738	5.247738	5.247738	5.247738	5.247738	5.247738
	Ts	2.431336	4.003807	3.992752	4.309308	4.847075	4.929163	4.085573
Month 2	Tr	4.350513	4.350513	4.350513	4.350513	4.350513	-	4.350513
	Ts	4.097735	3.059178	3.379151	3.07731	3.01717	-	3.326109
Month 3	Tr	4.200204	4.200204	4.200204	4.200204	-	-	4.200204
	Ts	1.135087	3.139817	2.580821	2.871832	-	-	2.431889
Month 4	Tr	2.910787	2.910787	2.910787	-	-	-	2.910787
	Ts	4.26774	3.057815	3.176502	-	-	-	3.500685
Month 5	Tr	2.774211	2.774211	-	-	-	-	2.774211
	Ts	1.076866	2.33621	-	-	-	-	1.706538
Month 6	Tr	2.566452	-	-	-	-	-	2.566452
	Ts	3.458153	-	-	-	-	-	3.458153

Table 112. ARIMA forecasting results for course CCC-B (2014)

- **Training average:** 4.137478
- **Test average:** 3.249754

7.5.6.4.2 Neural Network

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	6.143161	6.154606	6.146432	6.139077	6.149843	6.15064	6.148579
	Ts	2.299092	3.857911	2.753401	3.14508	3.246413	3.432599	3.338138
Month 2	Tr	4.919362	4.916852	4.918708	4.918197	4.917882	-	4.919393
	Ts	3.911031	2.867821	3.343565	2.906339	2.972247	-	3.172468
Month 3	Tr	4.752156	4.750552	4.75168	4.753171	-	-	4.755158
	Ts	1.267557	3.132582	2.837664	3.379779	-	-	2.488896
Month 4	Tr	4.274967	4.276919	4.272388	-	-	-	4.274613
	Ts	4.135695	2.978784	3.454489	-	-	-	3.45116
Month 5	Tr	4.308197	4.313248	-	-	-	-	4.307109
	Ts	1.953161	2.247225	-	-	-	-	2.341022
Month 6	Tr	3.978067	-	-	-	-	-	3.980653
	Ts	2.812685	-	-	-	-	-	2.908284

Table 113. Neural Network forecasting results for course CCC-B (2014)

- **Training average:** 5.04418
- **Test average:** 3.037646

7.5.6.4.3 Combination

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	5.82548	5.836292	5.828516	5.821476	5.832004	5.832564	5.830516
	Ts	2.350081	3.928914	3.251803	3.208793	3.878114	4.081492	3.659653
Month 2	Tr	4.682708	4.680237	4.682019	4.681402	4.681242	-	4.682601
	Ts	3.979894	2.926545	3.347546	2.945255	2.909273	-	3.223393
Month 3	Tr	4.495635	4.494317	4.495019	4.496745	-	-	4.498522
	Ts	3.979894	2.926545	3.347546	2.945255	2.909273	-	2.408918
Month 4	Tr	3.446026	3.447586	3.444319	-	-	-	3.445801
	Ts	4.196425	3.012426	3.304375	-	-	-	3.468923
Month 5	Tr	3.387348	3.39147	-	-	-	-	3.38726
	Ts	1.487602	2.177366	-	-	-	-	1.90061
Month 6	Tr	3.124265	-	-	-	-	-	3.126577
	Ts	3.118349	-	-	-	-	-	3.167826

Table 114. Combination forecasting results for course CCC-B (2014)

- **Training average:** 4.601366
- **Test average:** 3.099351

7.5.6.5 CCC-J (2014)

7.5.6.5.1 ARIMA

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	5.599771	5.599771	5.599771	5.599771	5.599771	5.599771	5.599771	5.599771
	Ts	2.248045	4.13635	3.909433	4.085007	4.36024	4.399216	4.426851	3.937877
Month 2	Tr	4.670537	4.670537	4.670537	4.670537	4.670537	4.670537	-	4.670537
	Ts	4.855063	3.869456	3.947315	3.995981	3.893743	3.975447	-	4.089501
Month 3	Tr	4.473964	4.473964	4.473964	4.473964	4.473964	-	-	4.473964
	Ts	1.07617	3.18677	2.644719	2.497327	3.304902	-	-	2.541978
Month 4	Tr	3.356739	3.356739	3.356739	3.356739	-	-	-	3.356739
	Ts	4.306045	3.241973	2.821228	3.281913	-	-	-	3.412789
Month 5	Tr	3.484428	3.484428	3.484428	-	-	-	-	3.484428
	Ts	2.383183	2.244785	2.906936	-	-	-	-	2.511635
Month 6	Tr	3.221679	3.221679	-	-	-	-	-	3.221679
	Ts	1.499444	3.396281	-	-	-	-	-	2.447862
Month 7	Tr	3.153147	-	-	-	-	-	-	3.153147
	Ts	4.094425	-	-	-	-	-	-	4.094425

Table 115. ARIMA forecasting results for course CCC-J (2014)

- **Training average:** 4.395292
- **Test average:** 3.392437

7.5.6.5.2 Neural Network

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	6.385615	6.384721	6.383398	6.38372	6.386883	6.384127	6.385195	6.384808
	Ts	0.980667	3.130803	2.623314	3.127031	3.003664	2.740714	3.247946	2.693448
Month 2	Tr	5.1289	5.126817	5.12831	5.128406	5.129512	5.128475	-	5.128403
	Ts	3.890333	3.17488	3.677311	3.372471	3.523945	3.467824	-	3.517794
Month 3	Tr	4.955594	4.953006	4.952698	4.951486	4.95148	-	-	4.952853
	Ts	2.178771	3.953681	3.238428	3.039772	4.070818	-	-	3.296294
Month 4	Tr	4.475847	4.475833	4.473377	4.474417	-	-	-	4.474868
	Ts	4.568947	3.479224	3.04537	4.157426	-	-	-	3.812742
Month 5	Tr	4.52742	4.524548	4.526445	-	-	-	-	4.526138
	Ts	0.754378	1.488925	3.779455	-	-	-	-	2.007586
Month 6	Tr	4.190219	4.192393	-	-	-	-	-	4.191306
	Ts	2.224708	4.329047	-	-	-	-	-	3.276877
Month 7	Tr	3.991852	-	-	-	-	-	-	3.991852
	Ts	5.503655	-	-	-	-	-	-	5.503655

Table 116. Neural Network forecasting results for course CCC-J (2014)

- **Training average:** 5.145739
- **Test average:** 3.206197

7.5.6.5.3 Combination

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	6.141328	6.140588	6.139235	6.139628	6.142468	6.139577	6.140999	6.140546
	Ts	1.45879	3.457368	3.062222	3.239253	3.125863	3.001412	3.072698	2.916801
Month 2	Tr	4.952948	4.951037	4.952353	4.952737	4.953644	4.952794	-	4.952585
	Ts	4.210033	3.099129	3.415082	3.087827	2.812162	3.055423	-	3.279943
Month 3	Tr	4.746146	4.743593	4.743327	4.742132	4.742398	-	-	4.743519
	Ts	1.53092	3.51124	2.890295	2.722611	3.642697	-	-	2.859553
Month 4	Tr	3.787092	3.78667	3.784978	3.785804	-	-	-	3.786136
	Ts	4.427531	3.254856	2.841697	3.623232	-	-	-	3.536829
Month 5	Tr	3.941597	3.938852	3.940869	-	-	-	-	3.940439
	Ts	1.25282	1.594518	3.004364	-	-	-	-	1.950567
Month 6	Tr	3.637737	3.639622	-	-	-	-	-	3.638679
	Ts	1.818389	3.821779	-	-	-	-	-	2.820084
Month 7	Tr	3.520907	-	-	-	-	-	-	3.520907
	Ts	4.076325	-	-	-	-	-	-	4.076325

Table 117. Combination forecasting results for course CCC-J (2014)

- **Training average:** 4.792181
- **Test average:** 3.003948

7.5.6.6 DDD-B (2013-2014)

7.5.6.6.1 ARIMA

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	2.820822	2.820822	2.820822	2.820822	2.820822	2.820822	2.820822
	Ts	2.101282	2.300456	2.284323	2.266826	2.494973	2.628827	2.346114
Month 2	Tr	2.813479	2.813479	2.813479	2.813479	2.813479	-	2.813479
	Ts	1.051751	1.125217	1.040609	1.245269	1.352232	-	1.163016
Month 3	Tr	2.742358	2.742358	2.742358	2.742358	-	-	2.742358
	Ts	1.043768	0.955001	1.090904	1.157087	-	-	1.06169
Month 4	Tr	2.679397	2.679397	2.679397	-	-	-	2.679397
	Ts	0.754366	1.175347	1.26242	-	-	-	1.064044
Month 5	Tr	2.631984	2.631984	-	-	-	-	2.631984
	Ts	1.26157	1.244842	-	-	-	-	1.253206
Month 6	Tr	2.574676	-	-	-	-	-	2.574676
	Ts	0.738006	-	-	-	-	-	0.738006

Table 118. ARIMA forecasting results for course DDD-B (2013-2014)

- **Training average:** 2.754219
- **Test average:** 1.455956

7.5.6.6.2 Neural Network

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	3.094517	3.094892	3.094342	3.094078	3.094808	3.094362	3.0945
	Ts	1.52557	1.411272	1.232809	1.976916	1.142486	1.087835	1.396148
Month 2	Tr	2.959862	2.959843	2.959903	2.959676	2.959431	-	2.959743
	Ts	1.056424	1.173655	1.415886	1.216683	1.444041	-	1.261338
Month 3	Tr	2.879899	2.879587	2.879372	2.879479	-	-	2.879584
	Ts	1.004611	1.127555	1.049494	1.21636	-	-	1.099505
Month 4	Tr	2.809039	2.809367	2.809155	-	-	-	2.809187
	Ts	0.515678	0.787666	1.083162	-	-	-	0.795502
Month 5	Tr	2.761448	2.76137	-	-	-	-	2.761409
	Ts	1.409829	1.398877	-	-	-	-	1.404353
Month 6	Tr	2.696605	-	-	-	-	-	2.696605
	Ts	0.626891	-	-	-	-	-	0.626891

Table 119. Neural Network forecasting results for course DDD-B (2013-2014)

- **Training average:** 2.930049
- **Test average:** 1.18589

7.5.6.6.3 Combination

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	2.929267	2.929437	2.92901	2.928647	2.929441	2.929002	2.929134
	Ts	1.44839	1.475495	1.442164	2.066869	1.404632	1.592049	1.5716
Month 2	Tr	2.778763	2.778789	2.77881	2.778601	2.778362	-	2.778665
	Ts	1.011453	0.912247	0.961178	1.158797	1.377672	-	1.084269
Month 3	Tr	2.705865	2.705593	2.705348	2.705451	-	-	2.705564
	Ts	1.002168	0.995877	1.010281	1.156264	-	-	1.041148
Month 4	Tr	2.643444	2.64374	2.643521	-	-	-	2.643568
	Ts	0.608422	0.961159	1.146365	-	-	-	0.905315
Month 5	Tr	2.599099	2.59906	-	-	-	-	2.59908
	Ts	1.289719	1.255926	-	-	-	-	1.272823
Month 6	Tr	2.539982	-	-	-	-	-	2.539982
	Ts	0.641175	-	-	-	-	-	0.641175

Table 120. Combination forecasting results for course DDD-B (2013-2014)

- **Training average:** 2.759963
- **Test average:** 1.186586

7.5.6.7 DDD-J (2013-2014)

7.5.6.7.1 ARIMA

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	2.10379	2.10379	2.10379	2.10379	2.10379	2.10379	2.10379	2.10379
	Ts	1.269496	1.395698	1.287793	1.249743	1.246225	1.158826	1.118297	1.246583
Month 2	Tr	2.322053	2.322053	2.322053	2.322053	2.322053	2.322053	-	2.322053
	Ts	1.065456	1.43998	1.297099	1.228396	1.23602	1.318617	-	1.264261
Month 3	Tr	2.310518	2.310518	2.310518	2.310518	2.310518	-	-	2.310518
	Ts	2.406651	1.946086	1.747697	1.758953	1.858017	-	-	1.943481
Month 4	Tr	2.250039	2.250039	2.250039	2.250039	-	-	-	2.250039
	Ts	1.012418	1.04705	0.965595	1.012346	-	-	-	1.009352
Month 5	Tr	2.210533	2.210533	2.210533	-	-	-	-	2.210533
	Ts	1.211808	0.946702	0.943482	-	-	-	-	1.033997
Month 6	Tr	2.188014	2.188014	-	-	-	-	-	2.188014
	Ts	0.901511	1.071559	-	-	-	-	-	0.986535
Month 7	Tr	2.131719	-	-	-	-	-	-	2.131719
	Ts	0.804554	-	-	-	-	-	-	0.804554

Table 121. ARIMA forecasting results for course DDD-J (2013-2014)

- **Training average:** 2.226819
- **Test average:** 1.283788

7.5.6.7.2 Neural Network

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	2.171363	2.171337	2.171239	2.171234	2.171525	2.17142	2.171282	2.171343
	Ts	4.436601	3.693021	3.406094	3.059442	2.871198	2.640685	2.483525	3.227224
Month 2	Tr	2.368284	2.368142	2.368182	2.368289	2.368169	2.368123	-	2.368198
	Ts	1.794886	1.544681	1.367303	1.593954	1.558779	1.697867	-	1.592912
Month 3	Tr	2.380283	2.38026	2.38041	2.38037	2.379948	-	-	2.380254
	Ts	1.347783	1.117033	2.303404	1.687867	1.029908	-	-	1.497199
Month 4	Tr	2.319999	2.32007	2.319992	2.319914	-	-	-	2.319994
	Ts	1.487911	1.33102	1.083359	1.09808	-	-	-	1.250092
Month 5	Tr	2.27831	2.278068	2.27804	-	-	-	-	2.278139
	Ts	1.657214	1.76483	1.370125	-	-	-	-	1.59739
Month 6	Tr	2.22095	2.220963	-	-	-	-	-	2.220956
	Ts	1.202548	1.188959	-	-	-	-	-	1.195754
Month 7	Tr	2.164572	-	-	-	-	-	-	2.164572
	Ts	1.310918	-	-	-	-	-	-	1.310918

Table 122. Neural Network forecasting results for course DDD-J (2013-2014)

- **Training average:** 2.286812
- **Test average:** 1.897464

7.5.6.7.3 Combination

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	2.059525	2.059498	2.059343	2.059344	2.05966	2.059562	2.059407	2.059477
	Ts	2.390937	2.252823	2.072132	1.909426	1.890232	1.713452	1.591947	1.974421
Month 2	Tr	2.287646	2.287459	2.287518	2.287628	2.287501	2.287491	-	2.28754
	Ts	1.110985	1.000645	0.983405	1.08878	1.00903	1.010922	-	1.033961
Month 3	Tr	2.286158	2.286189	2.286332	2.286258	2.285886	-	-	2.286164
	Ts	1.757815	1.491057	1.170164	1.316759	1.276685	-	-	1.402496
Month 4	Tr	2.226252	2.226374	2.226244	2.226133	-	-	-	2.226251
	Ts	1.211785	1.156853	1.007335	0.979054	-	-	-	1.088757
Month 5	Tr	2.186313	2.186106	2.186084	-	-	-	-	2.186167
	Ts	1.316174	1.246846	1.056773	-	-	-	-	1.206598
Month 6	Tr	2.12789	2.127904	-	-	-	-	-	2.127897
	Ts	1.008017	1.071452	-	-	-	-	-	1.039734
Month 7	Tr	2.071328	-	-	-	-	-	-	2.071328
	Ts	0.979244	-	-	-	-	-	-	0.979244

Table 123. Combination forecasting results for course DDD-J (2013-2014)

- **Training average:** 2.191537
- **Test average:** 1.359669

7.5.6.8 *EEE-B (2014)*

7.5.6.8.1 ARIMA

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	5.41942	5.41942	5.41942	5.41942	5.41942	5.41942	5.41942
	Ts	2.010717	1.876504	2.696225	3.778798	3.417929	3.182029	2.827034
Month 2	Tr	4.884803	4.884803	4.884803	4.884803	4.884803	-	4.884803
	Ts	0.930561	3.263466	4.61251	4.073106	3.75306	-	3.326541
Month 3	Tr	3.920576	3.920576	3.920576	3.920576	-	-	3.920576
	Ts	4.35998	5.484502	4.55437	4.054817	-	-	4.613417
Month 4	Tr	3.819199	3.819199	3.819199	-	-	-	3.819199
	Ts	4.768619	3.528038	3.0189	-	-	-	3.771852
Month 5	Tr	4.05683	4.05683	-	-	-	-	4.05683
	Ts	2.241488	2.200416	-	-	-	-	2.220952
Month 6	Tr	3.773336	-	-	-	-	-	3.773336
	Ts	1.349702	-	-	-	-	-	1.349702

Table 124. ARIMA forecasting results for course *EEE-B (2014)*

- **Training average:** 4.569878
- **Test average:** 3.29313

7.5.6.8.2 Neural Network

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	6.120843	6.120421	6.119311	6.120182	6.126631	6.121264	6.121442
	Ts	2.116336	2.027403	2.389401	3.236389	3.175686	2.878676	2.637315
Month 2	Tr	5.406334	5.405476	5.408688	5.406936	5.407801	-	5.407047
	Ts	1.044934	3.969625	5.194957	4.606261	4.358452	-	3.834846
Month 3	Tr	4.724793	4.72462	4.724506	4.724374	-	-	4.724573
	Ts	4.85051	5.842153	4.9376	4.415702	-	-	5.011491
Month 4	Tr	4.749061	4.750124	4.748348	-	-	-	4.749178
	Ts	4.215151	3.432103	3.115237	-	-	-	3.587497
Month 5	Tr	5.045158	5.04749	-	-	-	-	5.046324
	Ts	1.307931	1.444354	-	-	-	-	1.376143
Month 6	Tr	4.715644	-	-	-	-	-	4.715644
	Ts	1.379121	-	-	-	-	-	1.379121

Table 125. Neural Network forecasting results for course *EEE-B (2014)*

- **Training average:** 5.319905
- **Test average:** 3.33038

7.5.6.8.3 Combination

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	5.837344	5.836948	5.836206	5.836757	5.841845	5.83763	5.837788
	Ts	1.990817	1.937741	2.499945	3.439992	3.271829	2.942892	2.680536
Month 2	Tr	5.164791	5.164279	5.166926	5.16554	5.166466	-	5.1656
	Ts	0.709953	3.575134	4.879337	4.317937	4.03164	-	3.5028
Month 3	Tr	4.261902	4.261522	4.261932	4.261478	-	-	4.261708
	Ts	4.60026	5.65818	4.741376	4.230647	-	-	4.807616
Month 4	Tr	4.203688	4.203999	4.203095	-	-	-	4.203594
	Ts	4.388317	3.39017	2.955173	-	-	-	3.577887
Month 5	Tr	4.453352	4.454938	-	-	-	-	4.454145
	Ts	1.330169	1.365292	-	-	-	-	1.34773
Month 6	Tr	4.16003	-	-	-	-	-	4.16003
	Ts	1.335889	-	-	-	-	-	1.335889

Table 126. Combination forecasting results for course EEE-B (2014)

- **Training average:** 4.932413
- **Test average:** 3.2187

7.5.6.9 EEE-J (2013-2014)

7.5.6.9.1 ARIMA

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	4.136763	4.136763	4.136763	4.136763	4.136763	4.136763	4.136763	4.136763
	Ts	7.299948	6.306107	5.761875	5.207609	5.568498	6.397769	6.061816	6.086232
Month 2	Tr	4.528538	4.528538	4.528538	4.528538	4.528538	4.528538	-	4.528538
	Ts	1.837776	2.674988	2.485023	3.735448	5.237968	4.929872	-	3.483512
Month 3	Tr	4.405866	4.405866	4.405866	4.405866	4.405866	-	-	4.405866
	Ts	2.866002	2.32701	3.836541	5.521824	5.09132	-	-	3.928539
Month 4	Tr	4.353322	4.353322	4.353322	4.353322	-	-	-	4.353322
	Ts	1.672546	4.303036	6.215879	5.565623	-	-	-	4.439271
Month 5	Tr	4.270214	4.270214	4.270214	-	-	-	-	4.270214
	Ts	4.652423	6.743268	5.693863	-	-	-	-	5.696518
Month 6	Tr	4.313556	4.313556	-	-	-	-	-	4.313556
	Ts	5.576224	4.115223	-	-	-	-	-	4.845724
Month 7	Tr	4.474499	-	-	-	-	-	-	4.474499
	Ts	2.157623	-	-	-	-	-	-	2.157623

Table 127. ARIMA forecasting results for course EEE-J (2013-2014)

- **Training average:** 4.338694
- **Test average:** 4.637254

7.5.6.9.2 Neural Network

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	4.359979	4.360305	4.360231	4.360354	4.360274	4.360339	4.360151	4.360233
	Ts	8.118783	6.079743	5.984388	4.833431	5.265819	6.003181	5.820407	6.015107
Month 2	Tr	4.774177	4.77422	4.774735	4.774619	4.774444	4.774531	-	4.774454
	Ts	4.605071	3.872969	3.997488	4.548611	5.915379	5.37882	-	4.719723
Month 3	Tr	4.744422	4.744461	4.744319	4.743995	4.744159	-	-	4.744271
	Ts	1.734219	2.555865	2.294137	4.390492	4.810363	-	-	3.157015
Month 4	Tr	4.63422	4.634451	4.634317	4.634174	-	-	-	4.634291
	Ts	2.150347	4.534595	6.491951	5.448544	-	-	-	4.656359
Month 5	Tr	4.555809	4.555376	4.555113	-	-	-	-	4.555433
	Ts	2.836134	7.502536	4.850433	-	-	-	-	5.063034
Month 6	Tr	4.589581	4.59026	-	-	-	-	-	4.589921
	Ts	5.742569	3.220237	-	-	-	-	-	4.481403
Month 7	Tr	4.738168	-	-	-	-	-	-	4.738168
	Ts	2.934323	-	-	-	-	-	-	2.934323

Table 128. Neural Network forecasting results for course EEE-J (2013-2014)

- **Training average:** 4.607542
- **Test average:** 4.711458

7.5.6.9.3 Combination

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	4.134377	4.134769	4.134652	4.134744	4.134647	4.134714	4.134518	4.134632
	Ts	7.689909	6.153147	5.845549	4.968796	5.390687	6.182694	5.92077	6.02165
Month 2	Tr	4.560491	4.560476	4.561007	4.560826	4.560806	4.560796	-	4.560734
	Ts	2.912663	3.105551	3.067124	4.027234	5.519156	5.10036	-	3.955348
Month 3	Tr	4.478189	4.478268	4.478092	4.477777	4.477914	-	-	4.478048
	Ts	2.127885	1.840609	2.662947	4.818874	4.910572	-	-	3.272177
Month 4	Tr	4.373353	4.373578	4.373503	4.373332	-	-	-	4.373442
	Ts	1.900286	4.408852	6.331299	5.405542	-	-	-	4.511495
Month 5	Tr	4.316847	4.316601	4.316284	-	-	-	-	4.316577
	Ts	3.281964	7.040525	5.133351	-	-	-	-	5.151947
Month 6	Tr	4.359331	4.359907	-	-	-	-	-	4.359619
	Ts	5.14009	3.383618	-	-	-	-	-	4.261854
Month 7	Tr	4.469598	-	-	-	-	-	-	4.469598
	Ts	2.532998	-	-	-	-	-	-	2.532998

Table 129. Combination forecasting results for course EEE-J (2013-2014)

- **Training average:** 4.368907
- **Test average:** 4.52868

7.5.6.10 FFF-B (2013-2014)

7.5.6.10.1 ARIMA

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	7.339589	7.339589	7.339589	7.339589	7.339589	7.339589	7.339589
	Ts	3.222774	2.704142	2.354714	2.866877	3.000854	3.117106	2.877744
Month 2	Tr	7.407904	7.407904	7.407904	7.407904	7.407904	-	7.407904
	Ts	2.071695	1.782105	2.810359	3.041216	3.208228	-	2.582721
Month 3	Tr	7.257936	7.257936	7.257936	7.257936	-	-	7.257936
	Ts	1.608671	3.388683	3.611873	3.76282	-	-	3.093012
Month 4	Tr	7.122684	7.122684	7.122684	-	-	-	7.122684
	Ts	4.202623	4.01515	4.023884	-	-	-	4.080553
Month 5	Tr	7.017004	7.017004	-	-	-	-	7.017004
	Ts	1.907375	2.050565	-	-	-	-	1.97897
Month 6	Tr	6.916005	-	-	-	-	-	6.916005
	Ts	1.981748	-	-	-	-	-	1.981748

Table 130. ARIMA forecasting results for course FFF-B (2013-2014)

- **Training average:** 7.258422
- **Test average:** 2.89207

7.5.6.10.2 Neural Network

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	7.891098	7.891316	7.891206	7.890803	7.891061	7.890623	7.891018
	Ts	6.467103	7.5909	6.080596	5.056916	3.944739	3.670709	5.468494
Month 2	Tr	7.894734	7.894863	7.895272	7.894593	7.894775	-	7.894847
	Ts	3.688379	6.060877	8.441296	5.439047	6.089145	-	5.943749
Month 3	Tr	7.751021	7.751192	7.750931	7.751141	-	-	7.751071
	Ts	3.670467	5.623142	4.148487	4.129814	-	-	4.392977
Month 4	Tr	7.574173	7.573912	7.57464	-	-	-	7.574242
	Ts	2.270372	2.641655	3.866539	-	-	-	2.926189
Month 5	Tr	7.477959	7.477823	-	-	-	-	7.477891
	Ts	3.095346	3.788371	-	-	-	-	3.441858
Month 6	Tr	7.355716	-	-	-	-	-	7.355716
	Ts	1.851289	-	-	-	-	-	1.851289

Table 131. Neural Network forecasting results for course FFF-B (2013-2014)

- **Training average:** 7.755183
- **Test average:** 4.648342

Combination

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	7.464081	7.464322	7.464122	7.463776	7.464012	7.46361	7.463987
	Ts	4.645189	4.741489	3.915479	3.628105	3.148596	3.128523	3.867897
Month 2	Tr	7.500926	7.50097	7.501378	7.500782	7.500958	-	7.501002
	Ts	2.211973	3.205139	4.006408	2.731391	3.35723	-	3.102428
Month 3	Tr	7.361187	7.361365	7.361136	7.361368	-	-	7.361264
	Ts	2.025258	3.345789	3.161154	3.554248	-	-	3.021612
Month 4	Tr	7.208996	7.208761	7.209488	-	-	-	7.209082
	Ts	2.799106	2.66	3.599497	-	-	-	3.019534
Month 5	Tr	7.108278	7.108185	-	-	-	-	7.108232
	Ts	2.205274	2.709042	-	-	-	-	2.457158
Month 6	Tr	7.000902	-	-	-	-	-	7.000902
	Ts	1.839893	-	-	-	-	-	1.839893

Table 132. Combination forecasting results for course FFF-B (2013-2014)

- **Training average:** 7.360886
- **Test average:** 3.172323

7.5.6.11 FFF-J (2013-2014)

7.5.6.11.1 ARIMA

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	6.030562	6.030562	6.030562	6.030562	6.030562	6.030562	6.030562	6.030562
	Ts	5.537367	5.512437	4.757616	4.916581	5.357975	5.287347	5.241418	5.230106
Month 2	Tr	6.72147	6.72147	6.72147	6.72147	6.72147	6.72147	-	6.72147
	Ts	2.714709	2.222054	3.140037	4.177604	4.250778	4.315756	-	3.470156
Month 3	Tr	6.678737	6.678737	6.678737	6.678737	6.678737	-	-	6.678737
	Ts	1.817069	2.750599	3.897003	3.876571	3.911382	-	-	3.250525
Month 4	Tr	6.570344	6.570344	6.570344	6.570344	-	-	-	6.570344
	Ts	1.845588	3.558455	3.476065	3.561598	-	-	-	3.110426
Month 5	Tr	6.545718	6.545718	6.545718	-	-	-	-	6.545718
	Ts	5.519936	4.823407	4.655854	-	-	-	-	4.999732
Month 6	Tr	6.556549	6.556549	-	-	-	-	-	6.556549
	Ts	1.311435	2.235352	-	-	-	-	-	1.773393
Month 7	Tr	6.460415	-	-	-	-	-	-	6.460415
	Ts	3.2136	-	-	-	-	-	-	3.2136

Table 133. ARIMA forecasting results for course FFF-J (2013-2014)

- **Training average:** 6.479589
- **Test average:** 3.853057

7.5.6.11.2 Neural Network

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	6.249946	6.249826	6.250054	6.250274	6.250232	6.250194	6.249727	6.250036
	Ts	12.88893	11.5662	9.629372	8.661956	9.00745	8.552806	8.089341	9.770866
Month 2	Tr	6.938568	6.93831	6.938549	6.939096	6.938917	6.939027	-	6.938744
	Ts	6.021503	3.530911	3.899694	5.058871	3.382899	3.62918	-	4.253843
Month 3	Tr	6.906084	6.905728	6.905937	6.905924	6.905832	-	-	6.905901
	Ts	3.377026	3.258481	4.145005	3.646369	3.744187	-	-	3.634213
Month 4	Tr	6.77026	6.769714	6.770168	6.769904	-	-	-	6.770011
	Ts	4.514999	4.256772	4.147417	3.756738	-	-	-	4.168981
Month 5	Tr	6.735234	6.735195	6.735458	-	-	-	-	6.735296
	Ts	3.318989	2.526655	2.697055	-	-	-	-	2.847566
Month 6	Tr	6.730719	6.731263	-	-	-	-	-	6.730991
	Ts	3.063696	2.867098	-	-	-	-	-	2.965397
Month 7	Tr	6.66108	-	-	-	-	-	-	6.66108
	Ts	2.919802	-	-	-	-	-	-	2.919802

Table 134. Neural Network forecasting results for course FFF-J (2013-2014)

- **Training average:** 6.690043
- **Test average:** 5.219979

7.5.6.11.3 Combination

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	6.016658	6.016523	6.016693	6.016962	6.016948	6.016891	6.016375	6.016721
	Ts	9.045532	8.408616	7.109514	6.685354	7.073416	6.804801	6.565352	7.384655
Month 2	Tr	6.807893	6.807914	6.808072	6.808505	6.808537	6.808406	-	6.808221
	Ts	4.071161	2.712586	3.295015	4.136726	3.665021	3.640554	-	3.586844
Month 3	Tr	6.688197	6.687786	6.688007	6.688056	6.687946	-	-	6.687999
	Ts	2.132966	2.054473	2.451315	2.500277	2.504652	-	-	2.328737
Month 4	Tr	6.5778	6.577317	6.577807	6.577503	-	-	-	6.577607
	Ts	2.674164	2.912734	3.293516	2.97875	-	-	-	2.964791
Month 5	Tr	6.548283	6.548195	6.548529	-	-	-	-	6.548336
	Ts	4.292196	3.407053	3.049588	-	-	-	-	3.582946
Month 6	Tr	6.547701	6.54824	-	-	-	-	-	6.54797
	Ts	1.727484	2.067725	-	-	-	-	-	1.897605
Month 7	Tr	6.474936	-	-	-	-	-	-	6.474936
	Ts	2.751293	-	-	-	-	-	-	2.751293

Table 135. Neural Network forecasting results for course FFF-J (2013-2014)

- **Training average:** 6.497596
- **Test average:** 4.071851

7.5.6.12 GGG-B (2014)

7.5.6.12.1 ARIMA

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	0.920994	0.920994	0.920994	0.920994	0.920994	0.920994	0.920994
	Ts	0.521139	0.783599	0.931126	0.901341	1.219873	1.330249	0.947888
Month 2	Tr	0.881716	0.881716	0.881716	0.881716	0.881716	-	0.881716
	Ts	1.11147	1.243552	1.18098	1.516669	1.488839	-	1.308302
Month 3	Tr	0.805478	0.805478	0.805478	0.805478	-	-	0.805478
	Ts	1.04109	0.868607	1.284585	2.02329	-	-	1.304393
Month 4	Tr	1.001289	1.001289	1.001289	-	-	-	1.001289
	Ts	0.951386	1.844943	3.517482	-	-	-	2.104604
Month 5	Tr	1.029637	1.029637	-	-	-	-	1.029637
	Ts	1.899566	3.144726	-	-	-	-	2.522146
Month 6	Tr	1.193189	-	-	-	-	-	1.193189
	Ts	1.729259	-	-	-	-	-	1.729259

Table 136. ARIMA forecasting results for course GGG-B (2014)

- **Training average:** 0.924419
- **Test average:** 1.453989

7.5.6.12.2 Neural Network

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	0.915663	0.916799	0.915437	0.915829	0.91529	0.915506	0.915754
	Ts	0.651524	0.770214	0.890064	0.727868	1.114659	2.169697	1.054004
Month 2	Tr	0.89012	0.890033	0.890259	0.889774	0.891096	-	0.890256
	Ts	0.92865	0.885492	1.272779	1.251362	1.998954	-	1.267447
Month 3	Tr	1.002595	1.004841	1.003889	1.003007	-	-	1.003583
	Ts	0.762578	0.754421	1.258009	1.742534	-	-	1.129385
Month 4	Tr	1.135734	1.136476	1.136806	-	-	-	1.136338
	Ts	0.820223	1.519341	2.334287	-	-	-	1.55795
Month 5	Tr	1.144102	1.143404	-	-	-	-	1.143753
	Ts	1.926765	2.326297	-	-	-	-	2.126531
Month 6	Tr	1.340497	-	-	-	-	-	1.340497
	Ts	2.071879	-	-	-	-	-	2.071879

Table 137. Neural Network forecasting results for course GGG-B (2014)

- **Training average:** 0.999864
- **Test average:** 1.34179

7.5.6.12.3 Combination

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Mean
Month 1	Tr	0.882387	0.883058	0.881825	0.882139	0.881646	0.882222	0.882213
	Ts	0.570304	0.762625	0.904184	0.785094	1.157662	1.548993	0.95481
Month 2	Tr	0.855784	0.85487	0.85575	0.855213	0.856679	-	0.855659
	Ts	1.012633	1.043432	0.868363	1.295118	1.563186	-	1.156546
Month 3	Tr	0.838306	0.839903	0.839038	0.838274	-	-	0.83888
	Ts	0.870467	0.745027	1.183733	1.822359	-	-	1.155396
Month 4	Tr	1.025825	1.026584	1.027005	-	-	-	1.026471
	Ts	0.871558	1.608846	2.88575	-	-	-	1.788718
Month 5	Tr	1.039714	1.039418	-	-	-	-	1.039566
	Ts	1.822633	2.663587	-	-	-	-	2.24311
Month 6	Tr	1.208213	-	-	-	-	-	1.208213
	Ts	1.883073	-	-	-	-	-	1.883073

Table 138. Combination forecasting results for course GGG-B (2014)

- **Training average:** 0.918755
- **Test average:** 1.327077

7.5.6.13 GGG-J (2013-2014)

7.5.6.13.1 ARIMA

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	1.06767	1.06767	1.06767	1.06767	1.06767	1.06767	1.06767	1.06767
	Ts	0.90999	0.74103	0.71969	0.75238	1.15111	1.26351	1.37649	0.98774
Month 2	Tr	1.11411	1.11411	1.11411	1.11411	1.11411	1.11411	-	1.11411
	Ts	0.41352	0.52532	0.60487	1.13189	1.25108	1.36143	-	0.88135
Month 3	Tr	1.08220	1.08220	1.08220	1.08220	1.08220	-	-	1.08220
	Ts	0.59839	0.63566	1.25024	1.35209	1.45412	-	-	1.05810
Month 4	Tr	1.05475	1.05475	1.05475	1.05475	-	-	-	1.05475
	Ts	0.54580	1.41559	1.48123	1.56590	-	-	-	1.25213
Month 5	Tr	1.03748	1.03748	1.03748	-	-	-	-	1.03748
	Ts	1.61531	1.53981	1.59541	-	-	-	-	1.58351
Month 6	Tr	1.06233	1.06233	-	-	-	-	-	1.06233
	Ts	0.87775	1.07953	-	-	-	-	-	0.97864
Month 7	Tr	1.07653	-	-	-	-	-	-	1.07653
	Ts	1.60664	-	-	-	-	-	-	1.60664

Table 139. ARIMA forecasting results for course GGG-J (2014)

- **Training average:** 1.075073
- **Test average:** 1.100568

7.5.6.13.2 Neural Network

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	1.168764	1.168629	1.168729	1.168846	1.16866	1.16871	1.16869	1.168718
	Ts	1.769214	1.471478	1.407072	1.344042	1.305385	1.241439	1.401241	1.419982
Month 2	Tr	1.202557	1.202552	1.202801	1.202529	1.202702	1.202633	-	1.202629
	Ts	0.712002	0.902008	0.969348	1.096854	1.228035	1.12086	-	1.004851
Month 3	Tr	1.168876	1.168872	1.168658	1.168975	1.168751	-	-	1.168826
	Ts	0.804029	1.015352	0.996403	0.999775	1.053604	-	-	0.973833
Month 4	Tr	1.138099	1.137715	1.137853	1.138025	-	-	-	1.137923
	Ts	0.748173	1.414716	1.176645	1.482651	-	-	-	1.205546
Month 5	Tr	1.113983	1.114119	1.11366	-	-	-	-	1.113921
	Ts	1.535325	1.294721	1.064062	-	-	-	-	1.298036
Month 6	Tr	1.14605	1.146368	-	-	-	-	-	1.146209
	Ts	0.848673	0.892642	-	-	-	-	-	0.870657
Month 7	Tr	1.156701	-	-	-	-	-	-	1.156701
	Ts	1.322296	-	-	-	-	-	-	1.322296

Table 140. Neural Network forecasting results for course GGG-J (2014)

- **Training average:** 1.16493
- **Test average:** 1.163697

Combination

To From		Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Mean
Month 1	Tr	1.08212	1.0820	1.08209	1.08217	1.08201	1.08203	1.08201	1.08207
	Ts	1.25375	1.0240	1.00584	1.00681	1.15011	1.16889	1.34619	1.13652
Month 2	Tr	1.12716	1.1271	1.12734	1.12710	1.12731	1.12724	-	1.12722
	Ts	0.43118	0.6330	0.63607	0.92043	1.22585	1.19069	-	0.83954
Month 3	Tr	1.09389	1.0939	1.09372	1.09408	1.09383	-	-	1.09389
	Ts	0.63565	0.6423	1.00755	1.16207	1.15653	-	-	0.92082
Month 4	Tr	1.06407	1.0636	1.06385	1.06400	-	-	-	1.06390
	Ts	0.64363	1.4087	1.3196	1.52177	-	-	-	1.22345
Month 5	Tr	1.04398	1.0441	1.04370	-	-	-	-	1.04393
	Ts	1.56498	1.4103	1.31296	-	-	-	-	1.42943
Month 6	Tr	1.07240	1.0726	-	-	-	-	-	1.07251
	Ts	0.72987	0.9090	-	-	-	-	-	0.81943
Month 7	Tr	1.08421	-	-	-	-	-	-	1.08421
	Ts	1.45679	-	-	-	-	-	-	1.45679

Table 141. Combination forecasting results for course GGG-J (2014)

- **Training average:** 1.08657
- **Test average:** 1.066959

7.5.6.14 Summary and discussion

To count with an intuitive depiction of each algorithm's performance, a table showing training and test average RMSE (random mean squared error) for each course is presented:

Course	AAA-J (13/14)			BBB-B (13/14)			BBB-J (13/14)		
Algorithm	ARIMA	NNET	COMB	ARIMA	NNET	COMB	ARIMA	NNET	COMB
Tr	5.0793	4.7734	4.6895	1.7986	1.7127	1.6302	1.3894	1.5289	1.4328
Ts	2.7599	2.9391	2.5259	1.1688	0.9455	0.8131	1.9063	1.3878	1.4923
Course	CCC-B (14)			CCC-J (14)			DDD-B (13/14)		
Algorithm	ARIMA	NNET	COMB	ARIMA	NNET	COMB	ARIMA	NNET	COMB
Tr	4.1374	5.0441	4.6013	4.3952	5.1457	4.7921	2.7542	2.93	2.7599
Ts	3.2497	3.0376	3.0993	3.3924	3.2061	3.0039	1.4559	1.1858	1.1865
Course	DDD-J (13/14)			EEE-B (14)			EEE-J (13/14)		
Algorithm	ARIMA	NNET	COMB	ARIMA	NNET	COMB	ARIMA	NNET	COMB
Tr	2.2268	2.2868	2.1915	4.5698	5.3199	4.9324	4.3386	4.6075	4.3689
Ts	1.2837	1.8974	1.3596	3.2931	3.3303	3.2187	4.6372	4.7114	4.5286
Course	FFF-B (13/14)			FFF-J (13/14)			GGG-B (14)		
Algorithm	ARIMA	NNET	COMB	ARIMA	NNET	COMB	ARIMA	NNET	COMB
Tr	7.2584	7.7551	7.3608	6.4795	6.69	6.4975	0.9244	0.9998	0.9187
Ts	2.892	4.6483	3.1723	3.853	5.2195	4.0718	1.4539	1.3417	1.327
Course	GGG-J (13/14)								
Algorithm	ARIMA	NNET	COMB						
Tr	1.0750	1.1649	1.0867						
Ts	1.1005	1.1636	1.0669						

Table 142. Summary of forecasting models and algorithms' results

The first observation that can be made is referred to the overfitting reflected by some courses' results (signalled with a light-red colour). An in-depth inspection of these cases reveals that, while significant overfitting does take place in them, it occurs at specific scenarios, generally related with long-term forecasts.

This added to the fact that, although accurate, the results shown for each course are not uniform (the error for each course significantly varies from one course to other), indicates that the generalist approach formulated for forecasting each courses' time series may be weak in its foundation and more emphasis should be placed into assessing each courses' patterns and particularities independently.

Additionally, it is remarkable how overfitting appears at some courses' forecasts when the algorithms count with only 1 month of previous information. This may imply the need for setting a new threshold of previous information for this type of forecasting scenarios.

However, as a first approach, results invite to develop an optimistic view of the potentiality of these methodologies for predicting student's behaviour over time.

Since there does not appear to be enough difference between the performance of each algorithm defined for stating the appropriateness of any of them to suit this specific problem, it becomes an open-ended question which may be responded with further experimentation, addressing the benefits and disadvantages of each of them (e.g. the non-deterministic nature of the combinative approach against its possibility of smoothing its comprised algorithms' potential bias).

8 CONCLUSIONS

As a concluding point, and recalling the objects stated for this project's development, a summary of the main processes conducted along with the interpretation of their outcomes is presented:

- Assessment of current literature's approach to Learning Analytics: the descriptive capabilities of a purely performative approach with respect to courses' outcomes has been statistically tested.

Results from this analysis concluded that, in order to capture a broader and effective picture of students' engagement to learning processes, more sources of information need to be considered, including qualitative ones.

Additionally, a model addressing these points is proposed as the basis for the development of further analytics tasks.

- Development of scalable predictive and analytics processes: students' interaction with the virtual environment and outcomes were set as the cornerstones of this project's analytics tasks.

Considering the model defined during the assessment of the approach given to Learning Analytics, a broad set of features is confectioned in order to capture as much valuable information as possible.

The regression tasks performed offered high precision in its outcomes, and the previous rejection of purely performative models is reinforced by comparing the results obtained by using this model and our proposed one.

A general forecasting model for student's interaction was made with promising results, although arising the need for a more in-depth assessment of each course's particularities.

Additionally, as a support tool for these processes, the design and deployment of a database with which to operate with data was conducted.

8.1 Further Work

In coherence with the delineation of objectives for Learning Analytics as part of an institution's Business Intelligence agenda made in the introduction of this document, this project's main purpose is to set a first step into the inclusion of the predictive tasks conducted into a real operating system.

Along with room for the improvement of the processes detailed, there are many ambits and disciplines (e.g. software development, pedagogy, etc) from which to draw inspiration and develop guidelines for the confection of a robust and perdurable system.

Finally, and accounting for the need of alignment of this discipline's approaches and methodologies in order for it to leave its nascent stage towards a widespread adoption of its functionalities, effort should be put into the development of robust and statistically proven assumptions from which to construct a unified approach to the fulfilment of Learning Analytics' objectives.

9 APPENDIX: ORGANIZATION

9.1 Project's planification

The development of this project has implied a methodology based with three distinct cornerstones:

- Bibliographical research: consisting in the search and review of previous studies related with the project's domain, as well as content aimed obtain the required basis for the comprehension and conduction of the different processes involved in its development.
- Development: coding of the analytical tasks to be conducted using R language and its associated development environment RStudio (apart from the install and upload of the correspondent packages).
- Assessment of results: once experiments have been conducted their results are evaluated to extract conclusions with respect to their causes, while addressing the accomplishment of the project's objectives.

Following, a schedule detailing the terms assigned to each different task is shown:

Schedule	2018					
	May	June	July	August	September	October
Bibliographical research						
Project's objectives						
Development						
Test of Analytics tasks						
Project's documentation						
Project's presentation						

Table 143. Schedule of the project's tasks

9.2 Project's budget

The project's conduction has comprised the utilization of the following main resources:

HARDWARE			
Item	Model	Cost	Amortization
Computer	HP Envy 700-400NS	1000 €	3 years 333 €
Office supplies	Various (pens, sheets, etc)	100 €	100 €

Table 144. Hardware required for the development of the project

WORKFORCE		
Rank	Dedication (hours per day)	Cost (€/hour)
Junior engineer	12 hours	10 €

Table 145. Workforce required for the development of the project

With respect to the software employed, since its usage has implied no cost, a description of its functions is presented:

- Internet: for accessing the different bibliographical resources required. Access permissions granted by Carlos III University has also been used to access different online sources.
- Oracle Database Express Edition 11g: to develop the required database in which the data collection used has been allocated. Operation with the database has been conducted by using Window's CMD.
- RStudio: R language statistical computing environment for the development of the required coding tasks.

An additional expenditure has come from the computer's electric consumption, which has been estimated to be of 3.6kWh per month, which is equivalent to 21.6 total kWh along the duration of this project's development.

Given a default cost of 0.13 €/kWh ([56]) a total amount of 2.8 € referred to electric consumption is calculated.

Finally, the total project's budget can be summarized (assuming a time span of 6 months and round out of electricity cost) as:

Source	Cost
Hardware	433 €
Software	0 €
Workforce	21.600 €
Electric consumption	3 €
Total	22.036 €

Table 146. Final project's budget

10 REFERENCES

- [1] “Advanced Analytics & Big Data Adoption Report,” International Institute for Analytics, 2104.
- [2] R. v. Meulen and T. McCall, “Gartner,” 5 2 2018. [Online]. Available: <https://www.gartner.com/newsroom/id/3851963>. [Accessed 20 6 2018].
- [3] N. Henke, J. Brughin, M. Chui, J. Manyika, T. Saleh, B. Wiseman and G. Sethupathy, “The age of Analytics: Competing in a data-driven world,” McKinsey & Company, Inc., 2016.
- [4] “Learning and Knowledge Analytics,” 5 8 2011. [Online]. Available: <http://www.learninganalytics.net/uncategorized/learning-and-academic-analytics/>. [Accessed 15 6 2018].
- [5] R. Ferguson, A. Brasher, D. Clow, A. Cooper, G. Hillaire, J. Mittelmeier, B. Rienties, T. Ullmann and R. Vuorikari, “JRC Science for policy report: Research evidence on the use of Learning Analytics,” European Commission's Joint Research Centre, 2016.
- [6] P. Arroway, G. Morgan, M. O'Keefe and R. Yanosky, “Learning Analytics in Higher Education. Research Report,” EDUCAUSE Center for Analysis and Research, Louisville, CO:ECAR, 2016.
- [7] “Real Decreto 1720/2007, de 21 de diciembre, por el que se aprueba el Reglamento de desarrollo de la Ley Orgánica 15/1999, de 12 de diciembre, de protección de datos de carácter personal,” in *BOE*, Boletín Oficial del Estado, 2008, pp. 17:4103-4136.
- [8] “ETEC510: Design Wiki,” 11 3 2013. [Online]. Available: http://etec.ctlt.ubc.ca/510wiki/Learning_Analytics:_An_Introduction_and_Critical_Analysis#History. [Accessed 13 6 2018].
- [9] G. Siemens, “Learning Analytics: The Emergence of a Discipline,” *American Behavioral Scientist*, vol. 57, no. 10, pp. 1380-1400, 2013.
- [10] E. R. Khu, “Framing student engagement in higher education,” *Studies in Higher Education*, vol. 38, no. 5, pp. 758-773, 2013.
- [11] M. Scheffel, H. Drachsler, S. Stoyanov and M. Specht, “Quality Indicators for Learning Analytics,” *Educational Technology & Society*, 2014.
- [12] K. L. McClarty and M. N. Gaertner, “Measuring Mastery: Best practices for assessment in competency-based education,” Center on Higher Education Reform (American Enterprise Institute), 2015.
- [13] I. Gartner, “Gartner,” [Online]. Available: <https://www.gartner.com/en/research/methodologies/gartner-hype-cycle>. [Accessed 10 6 2018].
- [14] K. Panetta, “Gartner,” 7 11 2017. [Online]. Available: <https://www.gartner.com/smarterwithgartner/top-trends-from-the-gartner-hype-cycle-for-midsize-enterprises-2017/>. [Accessed 10 6 2018].

- [15] M. D. Sarrel, *The state of Data Analytics and Visualization adoption: A survey of usage, access methods, projects, and skills*, Sebastopol, CA: O'Reilly Media, Inc., 2017.
- [16] A. Urbinati, M. Bogers, V. Chiesa and F. Frattini, "Creating and capturing value from Big Data: A multiple-case study analysis of provider companies," *Technovation*, 2018.
- [17] T. H. Davenport, "The rise of Analytics 3.0: How to compete in the Data Economy," International Institute for Analytics, 2013.
- [18] V.-H. Trieu, "Getting value from Business Intelligence systems: A review and research agenda," *Decision Support Systems*, vol. 93, pp. 111-124, 2017.
- [19] R. Meredith and P. O'Donnell, "A functional model of Social Media and its application to Business Intelligence," in *Bridging the Socio-technical Gap in Decision Support Systems*, IOS Press, 2010, pp. 129-140.
- [20] T. O'Reilly, "What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," in *Communications & Strategies*, Sebastopol, CA, O'Reilly Media, Inc., 2007, pp. 17-37.
- [21] Z. Zdrahal, D. Beran, D. Herramannova, M. Hlosta, J. Kocvara and J. Zapletal, "OU Analyse," Knowledge Media Institute, The Open University, 2018. [Online]. Available: <https://analyse.kmi.open.ac.uk>. [Accessed 15 6 2018].
- [22] "Tribal Group," 2018. [Online]. Available: <https://info.tribalgroup.com/student-insight-whitepaper>. [Accessed 15 6 2018].
- [23] L. Corrin, G. Kennedy, P. d. Barba, A. Bakharia, L. Lockyer, D. Gasevic, D. Williams, S. Dawson and S. Copeland, "Loop: A learning analytics tool to provide teachers with useful data visualisations," Ascilite, 2015.
- [24] "Skillaware," Skillaware, 2018. [Online]. Available: <http://skillaware.com/en/>. [Accessed 15 6 2018].
- [25] "Khan Academy," Khan Academy, 2018. [Online]. Available: <https://www.khanacademy.org/>. [Accessed 15 6 2018].
- [26] K. J., H. M. and Z. Z., "OU Analyse," Knowledge Media Institute, The Open University, 2017. [Online]. Available: https://analyse.kmi.open.ac.uk/open_dataset. [Accessed 20 6 2018].
- [27] "Studytonight," [Online]. Available: <https://www.studytonight.com/dbms/first-normal-form.php>. [Accessed 2018 5 20].
- [28] C. J. Date, *An Introduction to Database Systems*, Pearson Education, 2003.
- [29] A. Sen, "Redgate," 7 5 2009. [Online]. Available: <https://www.red-gate.com/simple-talk/sql/learn-sql-server/facts-and-fallacies-about-first-normal-form/>. [Accessed 2018 5 25].
- [30] C. J. Date, *The Relational Database Dictionary*, O'Reilly Media, Inc., 2006.

- [31] “The English Indices of Deprivation 2015 – Frequently Asked Questions (FAQs),” UK Government, Department for Communities and Local Government, 2015.
- [32] G. Veletsianos, A. Collier and E. Schneider, “Digging deeper into learners’ experiences in MOOCs: Participation in social networks outside of MOOCs, notetaking and contexts surrounding content consumption,” *British Journal of Educational Technology*, vol. 46, no. 3, pp. 570-587, 2015.
- [33] D. M. Department, “The Central Limit Theorem,” Dartmouth College, 2014.
- [34] D. Randall and C. Welsler, *The Irreproducibility crisis of modern science: Causes, Consequences and the Road to Reform*, National Association of Scholars, 2018.
- [35] J. Wilhelm, “ResearchGate,” 2015. [Online]. Available: https://www.researchgate.net/post/What_are_the_alternatives_for_p-values. [Accessed 27 6 2018].
- [36] T. Vorapongsathorn, S. Taejaroenkul and C. Viwatwongkasem, “A comparison of type I error and power of Bartlett’s test, Levene’s test and Cochran’s test under violation of assumptions,” *Songklanakarinn J. Sci. Technol.*, vol. 26, no. 4, pp. 537-547, 2004.
- [37] M. Héroux, “Scientifically Sound,” 13 7 2017. [Online]. Available: <https://scientificallysound.org/2017/07/13/cohens-d-standardiser/>. [Accessed 3 7 2018].
- [38] M. Héroux, 27 7 2017. [Online]. Available: <https://scientificallysound.org/2017/07/27/cohens-d-how-interpretation/>. [Accessed 3 7 2018].
- [39] C. Balow, “Illuminate Education,” 15 6 2017. [Online]. Available: <https://www.illuminateed.com/blog/2017/06/effect-size-educational-research-use/>. [Accessed 5 7 2018].
- [40] J. Moeyersoms and D. Martens, “KDnuggets,” 8 2016. [Online]. Available: <https://www.kdnuggets.com/2016/08/include-high-cardinality-attributes-predictive-model.html>. [Accessed 7 7 2018].
- [41] E. H. Jr., “Information Gain Versus Gain Ratio: A Study of Split Method Biases,” The MITRE Corporation, Washington, 2001.
- [42] S. S. Mangiafico, “Measures of Association for Nominal Variables,” in *Summary and Analysis of Extension Program Evaluation in R*, 2016, pp. 489-502.
- [43] D. Howell, “The University of Vermont,” 2018. [Online]. Available: <https://www.uvm.edu/~dhowell/StatPages/icc/icc-overall.html>. [Accessed 15 7 2018].
- [44] “Data Science Musing of Kapild,” 10 11 2015. [Online]. Available: <https://kapilddatascience.wordpress.com/2015/11/10/using-silhouette-analysis-for-selecting-the-number-of-cluster-for-k-means-clustering/>. [Accessed 15 7 2018].
- [45] A. Struyf, M. Hubert and P. J. Rousseeuw, “Clustering in an Object-Oriented Environment,” *Journal of Statistical Software*, vol. 1, no. 4, 1997.

- [46] M. Fernández-Delgado, E. Cernadas and S. Barro, “Do we need hundreds of classifiers to solve real world classification problems?,” *Journal of Machine Learning Research*, vol. 15, pp. 3133-3181, 2014.
- [47] C. Chen and L.-M. Liu, “Joint Estimation of Model Parameters and Outlier Effects in Time Series,” *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 284-297, 1993.
- [48] J. López-de-Lacalle, “Package 'tsoutliers',” The Comprehensive R Archive Network (CRAN), 2017.
- [49] “IBM Knowledge Center (IBM),” [Online]. Available: https://www.ibm.com/support/knowledgecenter/es/SS3RA7_15.0.0/com.ibm.spss.modeler.help/ts_outliers_overview.htm. [Accessed 20 7 2018].
- [50] R. J. Hyndman and A. V. Kostenko, “Minimum sample size requirements for seasonal forecasting models,” *Foresight*, no. 6, pp. 12-15, 2007.
- [51] N. Davies and P. Newbold, “Some power studies of a portmanteau test of time series model specification,” *Biometrika*, vol. 66, no. 1, pp. 153-155, 1979.
- [52] M. Magakian, “Anomaly,” 6 8 2015. [Online]. Available: <https://anomaly.io/detect-seasonality-using-fourier-transform-r/>. [Accessed 2 8 2018].
- [53] T. P. S. University, “PennState Eberly College of Science,” 2018. [Online]. Available: <https://onlinecourses.science.psu.edu/stat510/node/71/>. [Accessed 20 7 2018].
- [54] R. Williams, “Heteroskedasticity,” University of Notre Dame, 2015.
- [55] D. Kwiatkowski, P. C. Phillips, P. Schmidt and Y. Shin, “Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that econometric time series have a unit root?,” *Journal of Econometrics*, vol. 54, pp. 159-178, 1992.
- [56] “Red Eléctrica de España, S.A.,” 20 09 2018. [Online]. Available: <https://www.esios.ree.es/es/pvpc?date=21-09-2018>. [Accessed 20 09 2018].