SMDE - 2nd Assignment

Prat Sicart, Joan
UPC Barcelona Tech
joan.prat.sicart@est.fib.upc.edu

Thursday 16th May, 2019

This document supposes the deliverable of the first assignment of the lab practices of the subject Stadistic Modelling and Design of Experiments corresponding to the Spring semester of the MIRI master degree during the course 2018/19. This deliverable is divided according to the statement provided by the lecturers and should be accompanied by the corresponding resources containing the scripts required to reproduce the output of this work.

---

1 Executive Summary

The system that it's going to be analyzed, the problem and the Solution is stated as it follows.

1.1 Analysis

The analysis consists on evaluate how the different factors influence in the partials times of a marathon runner. In this project the system analyzed it's going to be the three last years of the Boston Marathon.

1.2 Problem

The problem of this project consists on be able to predict the behaviour of a runner by getting information regarding the different partial times.

1.3 Solution

The solution is a simulation model able to predict each partial time of a runner. The different 5km segments time for a runner will be predicted with linear models.

2 System description, introduction

This document describes the different procedures followed until reach the final simulation model for the Boston Marathon Runners, to do so, the first step is describe the system model which will be splitted in three different parts, the system, the entities and finally the factors.

2.1 System

As it has been mentioned before, the system that it's going to be modelled is the Boston Marathon, which as a marathon has 42km, and is ran as a road race.

2.2 Entity

The object of interest of our study or the entity, is the runner. In concrete in this project the runners that are going to be studied are going to be the elite runners.

2.3 Factors

The system have lot of different factors that influence on the entities behaviour, the principal ones can be ambiental factors such as the temperature and the wind, or Physical factors of each runner such as the gender, the age, the VO2, the interaction with the other runners and so on

3 Problem Description

In order to predict the different partial times for the Boston Marathon runners, first of all, it's necessary to know which factors influence in the partial times, and how and how much. However in a marathon the entities could have a lot of different behaviours with a same factor, for instance, the temperature will affect more to the older runners than the elite runners, therefore it's also necessary to split the data set into the different groups depending on the runners

behaviour, and as mentioned before in this project will be studied the elite group. And finally, once the model has been obtained it will be necessary to set some randomness avoid simulation with deterministic results, and it's going to be needed to evaluate how much noise it's better to add to the simulation model in order to have a large probability of getting a validate result but at the same time don't increase too much the confidence interval since how bigger will be, less accuracy will be in the model.

As explained before, even though there are a lot of factors that might interfere with the system, however, since the study of some factors can be very time consuming and at the end be almost irrelevant for simulating the linear model, in the following subsection will be specified which factors are considered and which no, n order to establish the limits of our study.

3.1 Simplification Hypotheses

The Simplification hypothesis are those factors that won't we in studied in the project, and therefore simplify the case to study, in this project the Simplification Hypothesis are:

SH-01

Physical parameters of the runners such us the height, the weight, the hydratation... since there is not available data of the physical characteristics of the runners.

SH-02

No interaction between the candidates even though the interaction between the runners in massive races it's a reality, it's a hard parameter to be studied since it's difficult to quantify.

SH-03

Altitude Since all the three different races occur in the same city this factor doesn't influence in the final result

SH-04

Wind Usually the marathons are cyclical therefore, it has been considered that the temperature is a ambiental parameter with much influence than this one, and mainly, there wasn't time to study the influence of this parameter.

SH-05

VO2 There wasn't time and neither data available of this factor, however it could have been calculated with some parameters such as the pace but possibly it will have caused lost of independence among the data.

SH-06

Humidity There wasn't time to indroduce that paramater.

3.1 Structural and Simplification Hypotheses

The structural hypothesis are those factors that are going to be studied in the project, and therefore evaluated whether influence the linear model or not:

EH-01

Gender is a parameter that is considered on the system and therefore will evaluated.

EH-02

Temperature is a parameter that is considered on the system and therefore will evaluated.

EH-03

Country is a parameter that is considered on the system and therefore will evaluated.

EH-04

Age is a parameter that is considered on the system and therefore will evaluated.

EH-05

Elite pace regular we'll consider that the pace for the elite runners is regular during the race.

4 Model specification

The simulation of model it has done with the FlexSim simulator, by simply creating three different elements:

- **The source** where the entities are generated.
- **the processors**, one for each partial with its particular linear model created previously in R.
- **the Sink** element where the entities are removed.

5 Codification

As explained before the processors elements if FlexSim contains the linear models for each segment for the elite runners, this linear models were obtained from the three last Boston marathon datasets, The information and the explanation of how this models have been obtained are in the Markdown document attached with this document

5.1 Data

The linear models obtained will have the following coefficients one for each independent variable and the intercept:

$Time_{km} = Intercept + \beta_1 * Gender + \beta_1 * Age + \beta_1 * Country + \beta_1 * Temperature$

The linear functions obtained for each segment are:

$Time_{5km} = 893.77598 + 178.83391 * Gender + 0.06184 * Country$

$Time_{10km} = 883.86048 + 187.18809 * Gender + 0.07771 * Country$

$Time_{15km} = 960.86594 + 182.27502 * Gender + 0.09576 * Country - 1.39796 * Temperature$

$Time_{20km} = 895.62909 + 181.41385 * Gender + 0.10460 * Country$

$Time_{25km} = 898.30456 + 190.70243 * Gender + 0.10892 * Country$

$Time_{30km} = 913.76025 + 207.30784 * Gender + 0.10885 * Country$

$Time_{35km} = 955.24821 + 224.37658 * Gender + 0.08423 * Country$

$Time_{40km} = 940.73031 + 208.79888 * Gender + 0.11079 * Country$

$Time_{final} = 441.40212 + 80.43195 * Gender + 0.03912 * Country$

Once the different segments models have been obtained it's possible to introduce them in the Flexsim, and as we said before the times will be obtained from the processors elements, which will have as process time the linear models.

However, since the linear models are deterministic it will also be necessary to introduce some randomness around the deterministic values obtained in order to make sure that the real values are found in this confidence interval, to do so, in the process time of each element will be introduce a normal distribution with a mean of 0 and a variance of 200 seconds, therefore, for the deterministic value created we will add or subtract at most 200 seconds, we've considered that it's an interval big enough to validate the model and at the same time preserve a good results accuracy.

Therefore, the final time for a particular segment will be depicted as follows

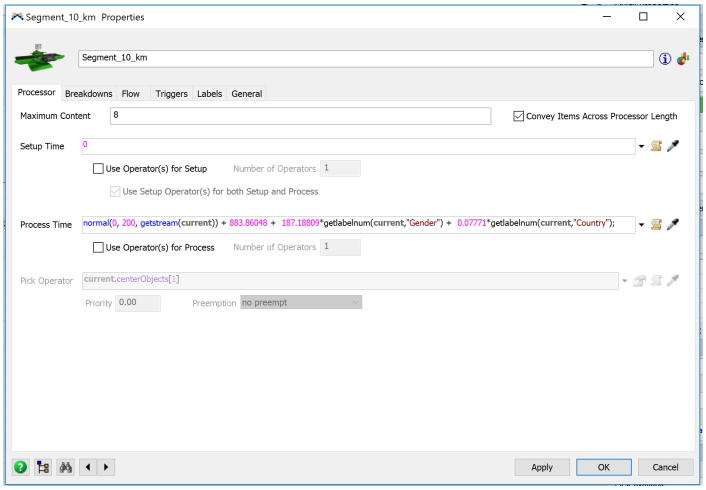

Figure 1: Settings for Flexsim processor

6 Definition of the experimental framework

Since the total number of factors considered in this project are four: Gender, Age, Temperature and Country are going to be necessary at least $2^{factors}$. Consequently the 2^k it's going to be done as it follows

Country	Gender	Age	Temperature
-	-	-	-
-	-	-	+
-	-	+	-
-	-	+	+
-	+	-	-
-	+	-	+
-	+	+	-
-	+	+	+
+	-	-	-
+	-	-	+
+	-	+	-
+	-	+	+
+	+	-	-
+	+	-	+
+	+	+	-
+	+	+	+

Table 1: Factorial design

Data	+	-
country	231	617
gender	0	1
age	20	40
Temperature	46	62

Table 2: Values for the factorial design

6.1 Factorial Design

Therefore, since the number of scenarios are $2^{factors}$ and we've got 4 factors will be needed a total amount of 16 scenarios, depicted as it follows:

where the values +/- represents the following values in the real system

Consequently, for the country the - sign is refered to the for the Gender the - sign is refered to the value 0 and therefore the males, while the + sign is refered to the females. For the the country - is refered to the runner which their nationality is from the Boston country, and + for the runners from the country of the winner, which has a country code of 231. Regarding the temperature the - it's equal to the temperature minus 54 degrees and + it's equal to the temperature with more thant 54 degrees, and finally, for the age, the minus means the runners younger than the 30 years, while the + means to the runners with more years. We are conscious that the

scenarios should be selected picking the maximun and minum values for each factor, however since the elite group marathon we are working on it's relatively small it's necessary amplify the range otherwise we won't select almost any experiment.

Therefore, once the scenarios are defined we'll begin the experimentation with 10 replications for each, and from the results we'll concrete if more replications are needed or not, so the results obtained with the model are the following ones:

Figure 2: Replication Boxplot

Summary								
	Mean (95% Confidence)					Sample Std Dev	Min	Max
Scenario 1	7708	<	7797	<	7887	125	7613	7994
Scenario 2	7685	<	7775	<	7864	125	7591	7971
Scenario 3	7708	<	7797	<	7887	125	7613	7994
Scenario 4	7685	<	7775	<	7864	125	7591	7971
Scenario 5	9351	<	9443	<	9535	129	9254	9649
Scenario 6	9328	<	9420	<	9513	129	9232	9627
Scenario 7	9351	<	9443	<	9535	129	9254	9649
Scenario 8	9328	<	9420	<	9513	129	9232	9627
Scenario 9	7686	<	7775	<	7864	125	7589	7970
Scenario 10	7663	<	7752	<	7841	125	7567	7947
Scenario 11	7686	<	7775	<	7864	125	7589	7970
Scenario 12	7663	<	7752	<	7841	125	7567	7947
Scenario 13	9329	<	9422	<	9514	129	9230	9628
Scenario 14	9306	<	9399	<	9491	129	9208	9605
Scenario 15	9329	<	9422	<	9514	129	9230	9628
Scenario 16	9306	<	9399	<	9491	129	9208	9605

Figure 3: Summary

6.2 Number of replications

To obtain the appropriate number of replications to have the desired interval of confidence, first of all, we've made an initial run with 10 replications and for them we have calculated the mean and the standard deviation, furthermore, now we'll get the half range with the formula $h =$

$t_{n-1,1-\alpha/2} * S/n^{1/2}$ where the α in our case is 0.05 to make sure that at least the 95 of 100 times the true average time is in the confidence interval. And finally, once the half range is calculated will be compared with the desired $s' = \alpha * median$ and if it's not smaller means that will be needed more replications, in concrete $n' = n * (h/h')^2$, otherwise with the 10 replications already done there's enough. Following this procedure we obtain the following table:

Scenario	response	estándar desviatio	n	expected desviatio	replications neede
1	7797	125	10	389,85	1,028075356
2	7775	125	10	388,75	1,033901635
3	7797	125	10	389,85	1,028075356
4	7775	125	10	388,75	1,033901635
5	9443	129	10	472,15	0,746482159
6	9420	129	10	471	0,750131851
7	9443	129	10	472,15	0,746482159
8	9420	129	10	471	0,750131851
9	7775	125	10	388,75	1,033901635
10	7752	125	10	387,6	1,040045859
11	7775	125	10	388,75	1,033901635
12	7752	125	10	387,6	1,040045859
13	9422	129	10	471,1	0,749813425
14	9399	129	10	469,95	0,753487605
15	9422	129	10	471,1	0,749813425
16	9399	129	10	469,95	0,753487605

Figure 4: number of replications needed for each scenario

Since in our simulation model we've just applied a variability of a normal distribution of 200, as it can be appreciated in the table above we don't need large amounts of replications to have a confidence interval of 95 per cent, however, afterwards we may have problems validating the data since the solution range will be smaller.

6.3 Factors interaction

Once the experiment it's correctly defined it's possible to get some conclusions regarding the results obtained, to do so, we are going to use the Yates algorithm to retrieve the importance of each factor and how contributes to the official.time, Therefore for the responses obtained in the previous experiment the values of each factor are:

	Response	Columna1	Columna2	Columna3	Columna4	Columna5	Effect	factor
1	7797	15572	31144	68870	137566	16	8597,875	mean
2	7775	15572	37726	68696	182	8	22,75	a
3	7797	18863	31054	90	0	8	0	b
4	7775	18863	37642	92	0	8	0	ab
5	9443	15527	44	0	-13170	8	-1646,25	c
6	9420	15527	46	0	-2	8	-0,25	ac
7	9443	18821	46	0	0	8	0	bc
8	9420	18821	46	0	0	8	0	abc
9	7775	22	0	-6582	174	8	21,75	d
10	7752	22	0	-6588	-2	8	-0,25	ad
11	7775	23	0	-2	0	8	0	bd
12	7752	23	0	0	0	8	0	abd
13	9422	23	0	0	6	8	0,75	cd
14	9399	23	0	0	-2	8	-0,25	acd
15	9422	23	0	0	0	8	0	bcd
16	9399	23	0	0	0	8	0	abcd

Figure 5: yates algorithm

Analyzing the results obtained by the Yates algorithm we can conclude that:

- **The Gender**, the factor C, as it can be appreciated in the Yates algorithm result is the factor that have a greatest importance, it can also be comproved in the Replication Boxplot, since when modifying the value of this factors the final results change completely.
- **the Age** as A, don't have impact in the final result, however in the reality we know that this isn't true, this may be because to get the model we just get the elite runners, and since almost have a similar age, the model didn't appreciate the difference between runners with different ages. However, we can conclute that for the elite runners the age is not important
- **the Temperature** as B has a lower impact compared with the ages, but also have to be considered since can modify the final results of the runners. How bigger the temperature longer official times, which has sense.
- **the Country**, the final factor, we can observe that as the temperature has lower impact compared with the gender, but can also increase the official time for those runners with a nationality of a country code bigger, which makes sense, since the African runners which are the expected faster runners have a low country code

7 Model validation

In this section, we want to compare the input vs output relationships of the model to the real system, i.e., see if the outputs of the model have the accuracy required in accordance with the problem.

To do so, first we need to obtain the real mean value for each scenario, which is depicted in the boxplot below:

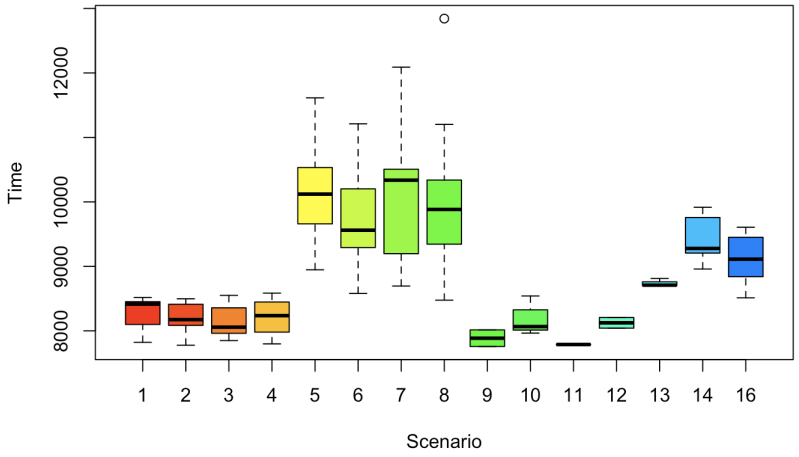

Figure 6: Boxplot of the real values for each scenario

if we now compare the real values from the simulation values we obtain the following results:

secenario	real time me	simulated me	estandar desviati	t value
1	8.284	7797	125	12,32655832
2	8.178	7775	125	10,19012353
3	8.154	7797	125	9,037308887
4	8.216	7775	125	11,16440863
5	10.134	9443	129	16,93232993
6	9.794	9420	129	9,159623041
7	10.054	9443	129	14,97274736
8	9.896	9420	129	11,6754478
9	7.886	7775	125	2,808102562
10	8.182	7752	125	10,87823515
11	7.788	7775	125	0,328876877
12	8.125	7752	125	9,423587427
13	8.740	9422	129	-16,71839817
14	9.428	9399	129	0,70545757
15		9422	129	-
16	9.106	9399	129	-7,194794521

Figure 7: Compartion between the real and the simulation responses

Comparing the real and the simulation results we obtain a large t values for few scenarios meaning that the real results are not in the simulation results with a confidence interval of 95 per cent, since the t values is larger than $t_{n-1,1-\alpha/2} = 2.262$, this may be for two reasons, the first one, as we've advanced when calculating the number of the number of replication necessary is that our model have a low variability it gives results with too much accuracy, futhermore, the other problem is that since our group elite is, the group we have analysed is very smaller (just 160 entities) to obtain values for each scenario we have had to pick values almost different from the simulation values, i.e for the country

we pick a value of 231 for the + country value, and in the real time we've selected from countries with a country code value greater 220 and smaller than 420. Therefore it's still necessary add more noise to the simulation model and moreover more real values to obtain a more accurate result for each scenario.

7.1 Individual validation

EH-01

Gender To validate this structural hyphotesis the data has been divided in two regarding the value of the gender and compare whether the official time follows the same normal distribution, to do so and Anova test has been implemented and the results were:

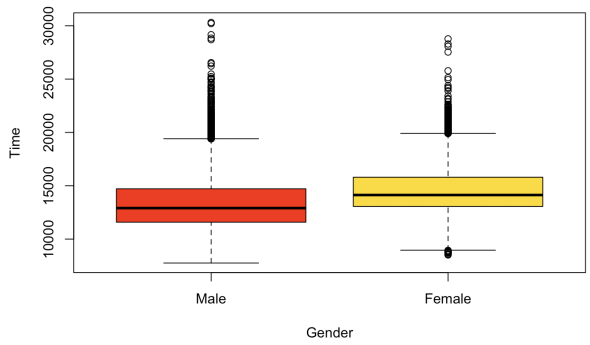

Figure 8: Annova Boxplot

```
              Df    Sum Sq   Mean Sq F value Pr(>F)
M.F            1 3.028e+10 3.028e+10    5220 <2e-16 ***
Residuals  79036 4.585e+11 5.801e+06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9: Annova statistics

As it can be appreciated in the image the normal distributions are different, since have a different mean, and the statistics confirm it, since the p-value is $< 2e-16$, smaller than 5 per cent, so we can reject the null hypothesis, meaning that the two means are different because of the gender, Therefore the official time depends on the Gender variable

EH-02

Temperature The temperature influence in the official time, this has also been validated with an Annova test.

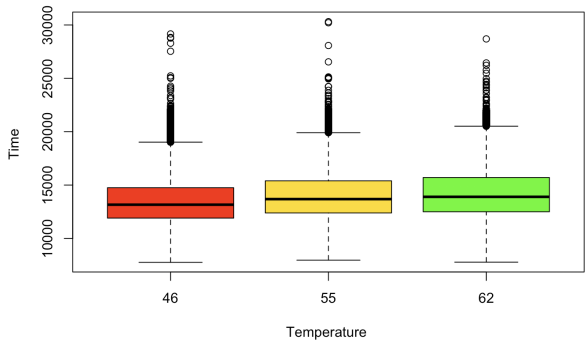

Figure 10: Annova Boxplot

```
                Df    Sum Sq   Mean Sq F value Pr(>F)
Temperature      1 6.977e+09 6.977e+09    1145 <2e-16 ***
Residuals    79036 4.818e+11 6.096e+06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: Annova statistics

Again the Annova p-value is lower than $< 2e - 16$, therefore at least there's one different normal distribution modifying the temperature, so again we can conclude that the temperature affects the official time.

EH-03

Country it's also a parameter that influence at the official time, since that the runners with African nationality tend to have better results.

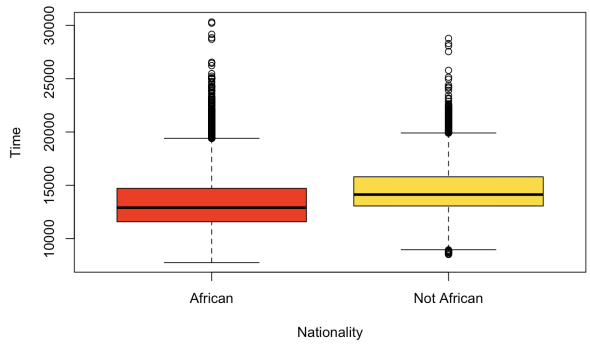

Figure 12: Annova Boxplot

Again the Annova p-value is lower than $< 2e - 16$, so we can conclude that the country affects the official time.

EH-04

Age We've also considered that the age may influence in the official time, and therefore we'll be also validated with an ANOVA test.

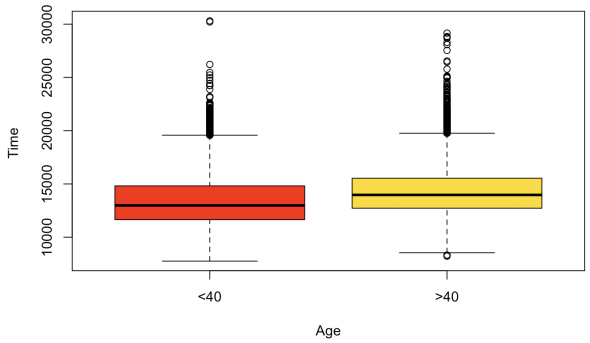

Figure 13: Annova Boxplot

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
Old           1 1.248e+10 1.248e+10    2072 <2e-16 ***
Residuals 79036 4.763e+11 6.026e+06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 14: Annova statistics

Again the Annova p-value is lower than $< 2e - 16$, so we can conclude that

the age affects the official time.

EH-05

Elite pace regular To validate this hyphotesis we'll implement two different linear regressions in function with the runner whether is in the category elite or amateur, where the regressions will be made among the partial times and the distance.

```
Residual standard error: 439.1 on 1070 degrees of freedom
Multiple R-squared:  0.9747,    Adjusted R-squared:  0.9747
F-statistic: 4.12e+04 on 1 and 1070 DF,  p-value: < 2.2e-16
```

Figure 15: Linear regression for elite runners

```
Residual standard error: 1292 on 286182 degrees of freedom
Multiple R-squared:  0.9064,    Adjusted R-squared:  0.9064
F-statistic: 2.772e+06 on 1 and 286182 DF,  p-value: < 2.2e-16
```

Figure 16: Linear regression for the amateur runners

As it can be appreciated in the statistics of the linear model for the amateur runners have a lower r square, and consequently has a minor linear behaviour, therefore we can validate that the elite pace is more regular than the amateur runners.

8 Conclusions

Finally, we can conclude that, we've considered that the majors factors that could affect the official time of a runner has been studied, however, we can assume the model as a good model for the moment, since in the section model validation, the results indicate that it doesn't correctly predicts the results with an interval of confidence of 95 per cent, furthermore, the model fails the three test, the Durbin Watson, which indicates that the data is not independent , the Shapiro test which indicates the model don't follow a normal distribution and finaly a the Breusch Pagan test which indicates that the data don't have homoscedasticity, which has sense because how bigger the distance bigger is the difference between the times of the different runners.

Even though, we've been able to obtain a important information regarding the factors, since the importance of the different factors have been revealed in the Yates algorithm, and therefore we can conclude that the factor that has a bigger impact in the final result is the gender, the second one with the biggest

impact is the country, followed by the age with a medium influence, and finally, the temperature, which had the lowest influence int the final time.