Universitat Politècnica de Catalunya

Barcelona School of Informatics

Master in Innovation and Research in Informatics

Statistical Modelling and Design Of Experiments

Second Deliverable

Author
Ricard Meyerhofer Parra

Lecturer
Pau Fonseca Casas

December 22, 2018

Index

1 Executive summary

In the following work we are going to present a simulation system which shows how relevant are certain attributes for a runner in the Boston Marathon. In this particular case, we are going to analyze how relevant is to be African, be Male/Female, how age affects the result and finally if being from Boston helps in having a better result and finally how the weather affects the performance.

The outcome of running simulations in the generated models, states that the aforementioned attributes have the following impact:

- **Being African:** Being African is not an advantage. If we compare African vs each continent it might be but in Africa vs rest of the world, is not.
- **Male/Female:** Being Male is helpful for a performing a better time.
- **Being Local:** Being local does not help which is quite normal because this means that more amateur runners will participate.
- **Age:** The best range of Age is the one compressed between 18-28.
- **Clime:** Sunny is better than rainy day.

2 System description, introduction

The system we are going to analyze is the one composed by a marathon and their participants. This system depends on many possible attributes such as the physical state of the participant, race of the participant, if it is a professional runner or not, how (s)he will manage the fatigue and if it is able to follow a good pace during the race, age, etc.

In our system we have that more specifically, our entities are the runners of the Boston Marathon from 2015, 2016 and 2017. The runners movement advancing to the finish goal (distance remaining) and the his/her pace would represent an endogenous activity and factors as the clime would represent as exogenous activities. The system progress is measured by each of the sectors in the race.

The system we are describing is a continuous system since it is expressed in terms of continuous equations showing how the system attributes change among the time that in the case of our system, will be expressed in seconds.

3 Problem description

The problem that we are trying to solve is to know which attributes make a runner more likely to win. In order to know **some** of these attributes, we are going to evaluate gender, age, proximity to the event and race.

- Is being from Africa a key factor to get a better result?
- Is the gender a factor that affects a runner performance?
- Since Boston and in general, USA runners tend to move less to go do the marathon. Is being from Boston an advantage? Is being from USA an advantage?
- How does the age affect a runner performance?

3.1 Structural and Simplifying Hypotheses

Since we can never achieve to simulate a system that fully represents the reality, we have to make some hypotheses that simplify the reality but because of it, are making possible to simulate something in a controlled environment. In this particular case we are going to have these hypotheses:

Systemic Structural Hypotheses

- **SH1:** In average, runners from Africa are better.
- **SH2:** In average, males run faster than female
- **SH3:** Boston runners are below the average
- **SH4:** Runners of an interval of 18-28 years, have a better performance than the remaining sectors.
- **SH5:** Runners from years that rains tend to be slower because of an inconvenient weather.

Simplification Hypotheses

- **SP1:** The wind nor aerodynamics affect a runners performance.
- **SP2:** All runners come at their best shape and personal moment.
- **SP3:** The runners we are taking into consideration do not have any unfair advantage respect other runners.
- **SP4:** Both, male and female are affected equally by all the attributes that will be studied.

Systemic Data hypotheses

- **SD1:** There is a dependency between sectors in the marathon.

4 Model specification

The model entities, operations and processes that defines the behavior of the model are the ones that can be seen in the diagram below where we can see the following equations represented as a flow. Where the first one is composed by the linear model of X5K and the other variables and in each step, the previous sector is incorporated to the prediction (so that we take into consideration the previous sectors of the runner). We add a Gaussian noise at the final prediction so that the results are non-deterministic. Once the model is created, we can proceed with the normal flow where we simulate the model to obtain some results which we will have to verify them to check which properties fulfill.

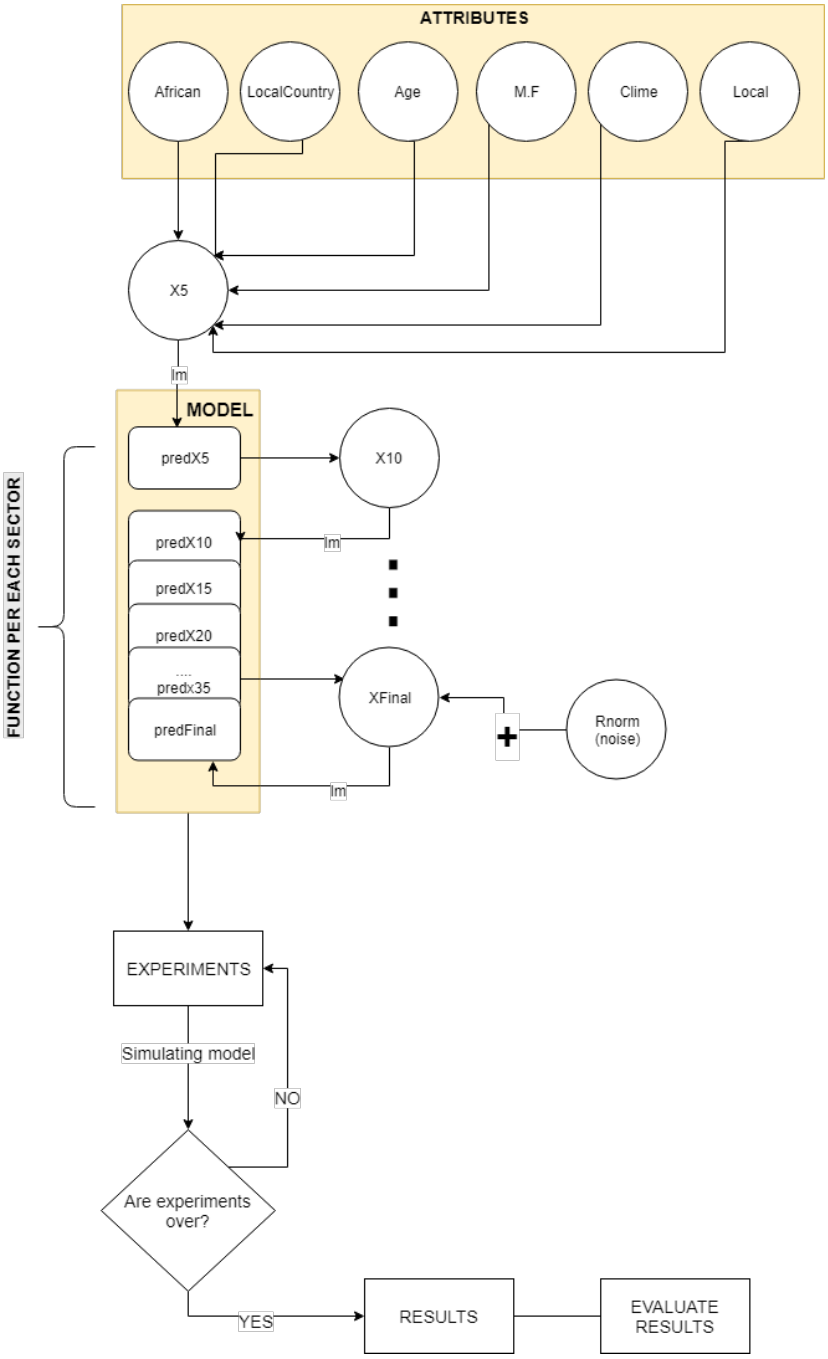

Figure 1: Model specification and its flow

5 Codification

Since I am quite comfortable with R and I want to improve my R skills I decided that instead of doing it with FlexSim or GPSS I wanted to do it in R. The code generated to read the 3 marathon datasets, joins them, generates new features, generates the model and performs the simulations is the code named "code.R" and the Boston Marathon datasets used to perform this analysis can be found in the "dataset" folder.

5.1 Data

The data that we are using comes from the Boston Marathon dataset from Kaggle which has 3 files each one corresponding to a year (2015, 2016 and 2017). Once we read this data in this case all 3 marathons are joined in a single dataset so that is easier to manipulate. The following attributes have been added or substracted from the dataset:

- All the attributes that were in a "hh:mm:ss" format have been converted into seconds (e.g X5K, X10K, etc)
- An attribute of the clime that Boston Marathon had that day has been added for each year (sun/rain)
- Age attribute was grouped according to different intervals with steps of 10.
- Binary attributes such as local, African to detect if someone is from Boston and Africa respectively (attributes that we want to study some hypotheses).
- Gender, Overall and Division are removed, Bib is not used. There are some others that could also be removed such as citizen.

Once the data is pre-processed and we do the feature extraction, it is used by the model as a linear model (lm) where we predict the final time considering each of the previous times aside from the attributes specified in the previous diagram.

6 Definition of the experimental framework

6.1 Definition of a DOE

The factors that we are using are:

- **Age:** How old is a runner. We have it classified with the following intervals: (18,28] (28,38] (38,48] (48,58] (58,68] (68,78]. To do the experiments we are going to use the minimum and the maximum since this variable is non-binary. The following ones are binary so they do not offer any problem.
- **Gender:** If the runner is male(M) or female(F).
- **Clime:** If it was rainy or sunny the day of the Marathon.
- **Local:** If it is from Boston takes the value Boston, otherwise NoBoston
- **African:** Similarly as the Local factor. African vs NoAfrican.

6.2 Detection and analysis of interactions

To detect and analyze this interactions we are going to do some experiments a measure the impact of each of this variables in order to verify our hypothesis.

Age	M.F	Clime	Boston	Africa	Total
-1	-1	-1	-1	-1	18076
-1	-1	-1	-1	1	18658
-1	-1	-1	1	-1	20093
-1	-1	-1	1	1	20678
-1	-1	1	-1	-1	15922
-1	-1	1	-1	1	16505
-1	-1	1	1	-1	17938
-1	-1	1	1	1	18528
-1	1	-1	-1	-1	16539
-1	1	-1	-1	1	17122
-1	1	-1	1	-1	18562
-1	1	-1	1	1	19141
-1	1	1	-1	-1	14382
-1	1	1	-1	1	14968
-1	1	1	1	-1	16404
-1	1	1	1	1	16988
1	-1	-1	-1	-1	13688
1	-1	-1	-1	1	14272
1	-1	-1	1	-1	15708
1	-1	-1	1	1	16296
1	-1	1	-1	-1	11529
1	-1	1	-1	1	12118
1	-1	1	1	-1	13552
1	-1	1	1	1	14139
1	1	-1	-1	-1	12152
1	1	-1	-1	1	12735
1	1	-1	1	-1	14173
1	1	-1	1	1	14759
1	1	1	-1	-1	9997
1	1	1	-1	1	10583
1	1	1	1	-1	12018
1	1	1	1	1	12603
-4386	-1536	-2155	2021	585	Effect
41,06	14,38	20,17	18,92	5,48	% Effect
(18,28]	M	Sun	NoBoston	NoAfrica	

Age	M.F	Clime	Boston	Africa	Total
(18,28]	M	Sun	Boston	Africa	
(68,78]	F	Rain	NoBoston	NoAfrica	

Figure 2: Table showing the effect of each variable and which is the most suitable

6.3 Effects of Age

Since age was not isolated in the previous experiment, now that we have everything fixed, we can isolate and work in which is the best rang of age to run.

	Time
(18,28]	12737
(28,38]	12762
(38,48]	13240
(48,58]	14030
(58,68]	15356

Table 1: Age results in seconds

As was probably expected, the best range to run is 18-28.

7 Model Validation

We must verify which properties the model has. A simulation system does not necessarily need to fulfill all the properties because it is a tool that is already working with a non-real environment. However, what is important is to know which properties the model **does** fulfill and which must fulfill for our purposes. Once said this, we are going to validate the following ones:

- **Durbin Watson:** The observations within each sample must be independent. Passed the test since we can see that p-value $<$ DL.
- **Shapiro test:** The populations from which the samples are selected must be normal. We cannot apply this test because the data is bigger than 5000 elements.
- **Breusch Pagan test:** The populations from which the samples are selected must have equal variances (homogeneity of variance). Two or more normal distributions are homoscedastic if they share a common co-variance (or are correlated). A p-value $>$ 0.05 indicates there is homogeneity (Does not pass the test, we reject the hypothesis that is homoscedastic).

- **Chi-Square:** We could have seen previously to create a model. If the variables we got initially in our data follow a Chi-Square tests in order to assure that the structure of the data is correct. I did not apply them but I think it is interesting to check it (as we did in the first deliverable).
- **R squared of the model:** We can see by doing a summary of our model that our R squared is near to 1. Therefore, it looks linear.

8 Results /Conclusions

From what we can conclude from our experiments with our data and considering the reliability of our model, the previous hypothesis:

- **SH1:** In average, runners from Africa are better. **False** this probably should be re-evaluated and check from each continent which is the one with the best times but as we did Africa vs World, Africans do not win. This surprises me and can be because of unbalanced data or an issue in the data that I do not see.
- **SH2:** In average, males run faster than female **True** it is not surprising.
- **SH3:** Boston runners are below the average. **True** it is logic if we thing that there will be more people that starts in their city to run marathons, etc.
- **SH4:** Runners of an interval of 18-28 years, have a better performance than the remaining sectors. **True** it is a bit obvious. Maybe doing chunks of 5 years would have been a better idea.
- **SH5:** Runners from years that rains tend to be slower because of an inconvenient weather. **True**. It was also obvious, the rain is inconvenient and most of times comes also with wind. Adding another parameter such as wind or temperature would probably make it more useful.