# How to Pace the London Marathon, Fuelled by Data

*Chris Dick*
*c.dick@warwick.ac.uk*
*University of Warwick*

## 1. Introduction

Inspired by the New York City Marathon, London hosted it's first mass participation road marathon on 29 March 1981. 6,255 runners completed the first race and in the following year 90,000 people decided to sign up [1]. The second London marathon accepted 18,059 runners and the race has grown in popularity ever since. As well as being one of the five original Abbot World Marathon Majors [2], the world famous course is holds the world record for the fastest ever women's marathon: Paula Radcliffe ran 2 hours, 15 minutes and 25 seconds in 2003 [3]. In addition to being a fast, flat course, the London Marathon is known world-wide for being the largest single-day charity fundraising event: in 2018 runners raised a total of £63.7 million [4]. In the same year a record 40,255 runners completed the 42.195km course [5]. While many runners wear fancy dress outfits to complete their run, the marathon is still a serious event that requires significant training and is the pinnacle of many runners' sporting careers.

In this paper we investigate the London Marathon results from the last 5 years, 2014 to 2018. We look into the distribution of results across genders and nationalities as well as consistency of pace throughout the race. Many runners place significance around particular times, such as the 'breaking 4 hours' or 'breaking 3 hours'. Achieving other set times, such as 3.05 for an under 35 year old male or 3.35 for an under 35 female, qualifies the runner for a place in the prestigious Boston Marathon. As such we investigate the distribute of times around these barriers. In the paper we look at predicting a runners full race time given a variety of factors including early race splits, gender, nationality and even weather.

## 2. Hypothesis

This paper takes a particular interest in runners' pacing strategies. At 42.195km in length, a full marathon is a long race for even an experience runner. Due to the high profile and charity aspects of the London Marathon, a number of unexperienced runners also enter the race. Beginning the race at full speed is a sure way to exhaust all energy and will likely result in the runner having to drop out of the race. Therefore, every runner must decide what pace to begin the race at, and typically a runner will decide what speed they want to keep to before they begin the race. This plan is referred to as a pacing strategy. These strategies are difficult to execute successfully. The author raises the hypothesis that actual pacing varies significantly amongst runners, with faster runners running negative splits (later sections of the race are faster than earlier sections), some runners achieving consistent splits but many runners succumbing to positive splits (later sections are slower than earlier sections). The author also hypothesises that adverse weather conditions (too hot, too wet or too cold) increase the number of runners who achieve positive splits. The author ran in both 2015 and 2018: at both times misjudged the race and ran disappointingly positive splits.

## 3. The Data

Detailed marathon completion times are not widely available. Kaggle.com has a dataset of Boston Marathon results [6] from 2015, 2016 and 2017 as well as a small dataset of results from the Hong Kong and Prague marathons. The author was not able to find a dataset of London Marathon results, however there are a number of studies that analyse London Marathon results so non-public datasets must exist. One of these studies is a blog post by Barry Smyth [7], Professor of Computer Science at University College Dublin. Professor Smyth appears to use split times (the runners elapsed time at 5km intervals) and final results from years 2011 to 2017 to conduct his analysis. Another analysis of the London Marathon was conducted by the sport social media company Strava [8]. This analysis using a far smaller sample size (around 6,000 runners compared to 200,000 runners by Professor Smyth) but looks into the training cycles of the runners in the weeks leading up to the marathon, as well as the marathon results themselves.

As far as we are aware, we are the first analysis to include 2018 data and the first analysis to combine weather data with London Marathon results.

### 3.1. Sourcing

While there is no public dataset for London Marathon results, the London Marathon website has searchable results for all runners [9]. To retrieve these results in bulk, the author built a scraping tool in python. The tool used the requests and Beautiful Soup modules: calling the results page of each participant for the last 5 years, copying the participant's data and saving the information to a local csv file. As there are nearly 200,000 unique participant results pages between 2014 and 2018, this required 200,000 calls

to the London Marathon servers. In order to not overwhelm the servers, a request was made every 0.1 seconds and so the scraping took just under 6 hours to run.

For the weather data, the public API from the website DarkSky.net was used [10]. Here historical weather data was retrieved in 10 minute intervals from the time that the first place runner crossed the finish line to the time that the last place runner crossed the finish line. This data was gathered for the 5 days when the marathons were held. The weather data was merged with the runners' results data.

### 3.2. Features

The retrieved results data was three-dimensional. That is, the dataset contains information about each runner as well as 5 fields that describe the runner's performance at each 5k point in the race. Overall there were 193,439 unique records retrieved, which is slightly lower than the number of participants in the races as the dataset does not contain elite runners. Duplicate entries were removed based on runner name, year of race and finish time. In the small possibility that two runners with the same name had the exact same finish time in the same year, one of these records will have been removed.

| | 2014 | 2015 | 2016 | 2017 | 2018 | Total |
|---|---|---|---|---|---|---|
| # Runners | 35,877 | 35,857 | 38,937 | 39,276 | 40,097 | 193,439 |
| Female % | 37.0% | 40.1% | 38.6% | 39.4% | 40.9% | 39.2% |
| Nations | 105 | 96 | 107 | 112 | 125 | 170 |

The fields describing each runner are somewhat limited: Name, Runner Number, Year of Race, Club, Age Category, Place, Gender and Nationality. It would have been ideal to have exact age, city of residence, charity they contributed to, fancy dress outfit and running experience (eg number of marathons completed) but the given fields are sufficient for our purposes. Below is a sample of the data.

| Name | Gender | Nation | Year | Category | Finish time |
|---|---|---|---|---|---|
| CHRIS DICK | M | GBR | 2018 | 18-39 | 03:16:38 |

The fields describing each runner's performance at each 5km mark are: Time of Day, Time, Split (time since last 5km mark), min/km, km/h. This data is also present for the finish line at 42.195km. The most useful fields here are overall Finish Time, min/km, Place and Time of Day. Time of Day, particularly Time of Day at the finish line, is useful to merge the weather data with the results data. In total we have 80 fields for each of the 193,439 runners. Below is a sample of the data at the 10km mark.

| Time Of Day | Time | Diff | min/km | km/h | Place |
|---|---|---|---|---|---|
| 10:46:01 | 00:42:55 | 21:38 | 04:20 | 13.88 | NaN |

For the weather data, the Dark Sky API provides the following fields: Summary, Precipitation Intensity, Precipitation Probability, Temperature, Apparent Temperature, Dew Point, Humidity, Pressure, Wind Speed, Wind Gust, Wind Bearing, Cloud Cover, UV Index, Visibility. For our purposes we keep Temperature, Wind Speed, Humidity, Precipitation Intensity and Precipitation Probability.

### 3.3. Cleaning

The results data was handled and cleaned using the pandas module in python. Each runners name came in the form "Dick, Chris (GBR)" and so the punctuation was removed, the forename was placed first and the entire name was capitalised. The runner's nationality was extracted and placed in it's own column. The retrieved data did not contain the runner's gender, however this could be inferred from the fields "Place (overall)" and "Place (M/W)". Gender was added as an additional column. When the data was retrieved, there was a column for each split (5km to 42.195km) and columns for general runner info. Each element under each split was a string describing Time of Day, Time, Split time, min/km, km/h. These strings were converted into 5 columns for each split. As the data was then three-dimensional, multi-indexing was required with the first level of column names being the split name (eg "35km") and the second level of column names being the information on each split. There was also a multi-index called "Runner" with columns containing general runner information. All time related information was converted from a string to the pandas datetime format.

The weather data arrived in a more friendly format. The following api call was used: https://api.darksky.net/forecast/⟨KEY⟩/⟨LONGITUDE⟩ ,⟨LATITUDE⟩,⟨TIME⟩?⟨FORMAT⟩?

To simplify the weather analysis, the location used for the weather was The Mall, London, which is the road on which the race finishes. Weather data was extracted in 5 miniute intervals. The API call outputs a dictionary, from which the required fields were extracted and loaded into a pandas dataframe. This dataframe was merged with the larger results data.

## 4. Analysis

### 4.1. Summary

| | Finish | 5K |
|---|---|---|
| **mean** | 04:32:36 | 00:05:45 |
| **std** | 01:00:11 | 00:01:02 |
| **min** | 02:14:48 | 00:02:53 |
| **25%** | 03:49:20 | 00:05:03 |
| **50%** | 04:27:09 | 00:05:41 |
| **75%** | 05:09:00 | 00:06:23 |
| **max** | 09:33:17 | 00:14:06 |

The first column of the table above gives the finish time statistics for the runners. The average finish time is

around 4.5 hours and 68% (1 standard deviation) of the runners finished within 3 to 5 hours. The fastest runner in the dataset is Josh Griffiths of Great Britain, who in 2017 completed the course in 2:14:48. The second column in the table gives a summary of the 5km split data, though it is hard to say if the minimum 5km split pace obtained is accurate. A first 5km pace of 2:53 is physically possible, though extremely unlikely and the runner who obtained it slowed down significantly for the rest of his race - indicating that this 5km split pace might be an error.

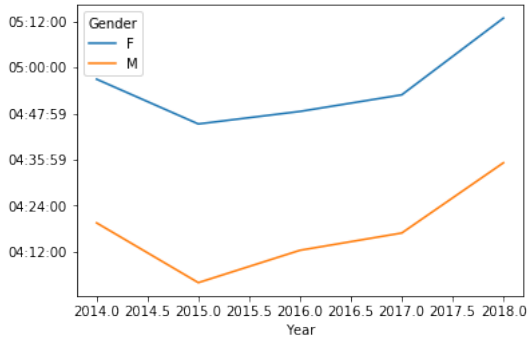


Figure 1. Finish times plotted per year, split by gender

From figure 1 we can see that average finish times are reasonably stable across the 5 years in the dataset, with men tending to finish faster than women and both genders taking longer on average to complete the course in 2018 compared to any other year.

We turn our attention now to each runner's pace throughout the race, as described by minutes per km measured at each 5km split of the race.

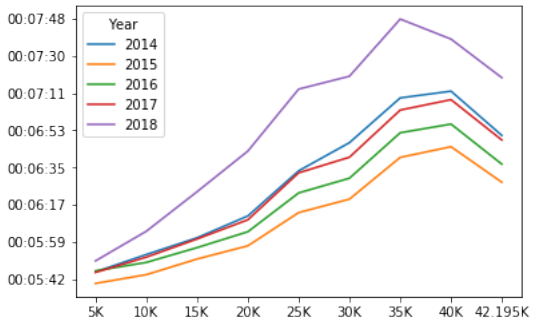


Figure 2. min/km plotted at each 5km, split by year

In figure 2 we see that average pace slows towards the end of the race every year; every year runners run positive splits on average. The slowest pace is not actually at the end of the race, as might be expected, but at the 35km or 40km mark. This can be explained as this is the point at which where many runners 'hit the wall': the runner depletes their glycogen stores and slows down significantly. What also jumps out from figure 2 is how much slower the average pace for the entire race in 2018. We look into this further, but first present average pace split by race finish time.

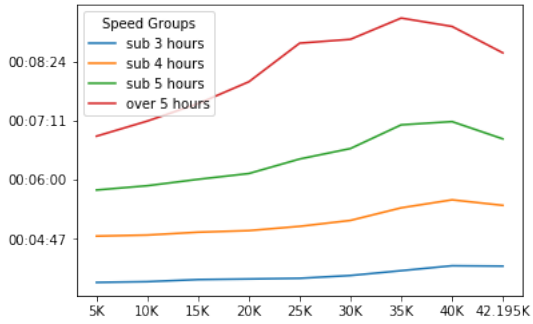


Figure 3. min/km plotted at each 5km, split by speed group

For figure 3 we split the runners into groups depending on their finish times. The categories chosen are those who finished in under 3 hours, those who finished in 3 - 4 hours, those who finished in 4 - 5 hours and those who took over 5 hours to complete the race. As expected, the faster runners held their pace better than the slower runners. It even looks like sub 3 runners held their pace constant for the last few km, indicating that fewer of these runners "hit the wall".

To look closer at the effect on runners pace in 2018, figure 4 presents the paces for only the sub 3 runners, broken down by year.

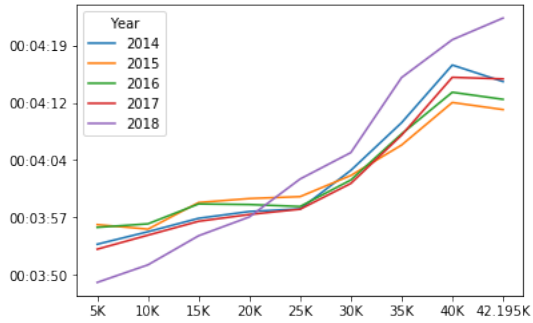


Figure 4. min/km for sub 3 runners only, split by year

Here we can see that in 2014 to 2017, average pace for sub 3 runners slowed by a maximum of 21 seconds per km (2017). In 2018 however, there is a significant difference with the average sub 3 runner actually starting out faster than normal and finishing way slower than normal, a slowing in pace of 33 seconds per km.

Bringing this together, we introduce the Slowdown Factor for each runner given by the formula:

$$100 \cdot \frac{PaceAt40km(runner) - PaceAt5km(runner)}{PaceAt5km(runner)}$$

This can be interpreted as the change in pace of a runner throughout the race as a percentage of their pace at the 5km mark, or how much they slow down throughout the race.

As expected, figure 5 shows that slower runners slow down more than faster runners. More interestingly, 2014

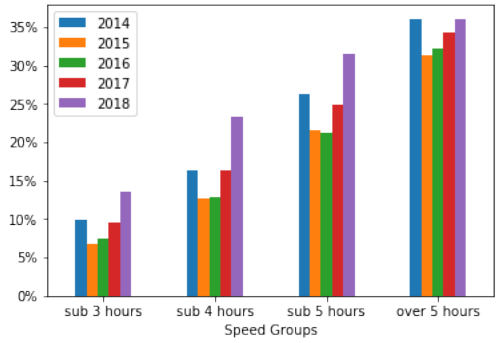


Figure 5. slowdown by speedgroup, split by year

stands out alongside 2018 as a year where the average runner slowed down significantly. Indeed, for the over 5 hour runners 2014 was the worst year for slowing down.

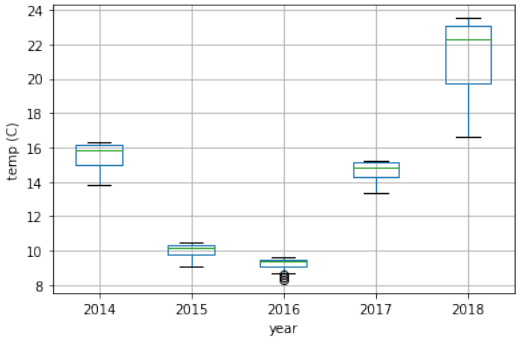


Figure 6. Boxplot of the temperature during each marathon

The range of temperatures experienced at each year of the marathon is given by figure 6. It is clear that 2018 was much hotter than previous years. 2014 is roughly speaking the second hottest year, although 2017 is a close third. Does temperature correlate to poor pacing performance?

## 4.2. Regression

To investigate causes of slowdown in runners, we ran a number of regressions and presented the results. The first regression explored was a simple linear regression of a runner's pace as a function of which point they are at in the race. The linear regression found here is described by the formula $pace = 0.00062 \cdot distance + 0.09265$ with a correlation coefficient of $0.954$. A plot of these data points with the regression line is given in figure 7.

However, the scatter plot in figure 7 does not provide a good visualisation of the results as there are 200,000 data points for each distance marker and so it is difficult to tell from the plot where the bulk of the runners are. Plotting the average pace for each distance marker gives a better visualisation in figure 8.

Although a linear regression does not capture the speed up that the average runner makes towards the end of the race,

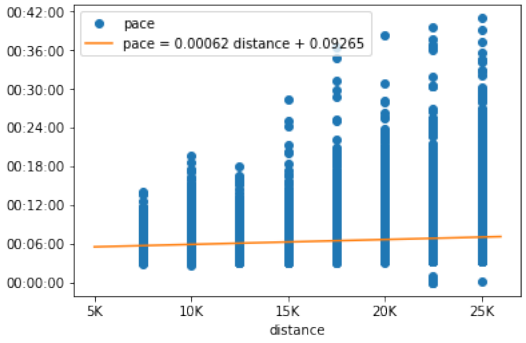


Figure 7. Pace vs distance ran with line of best fit

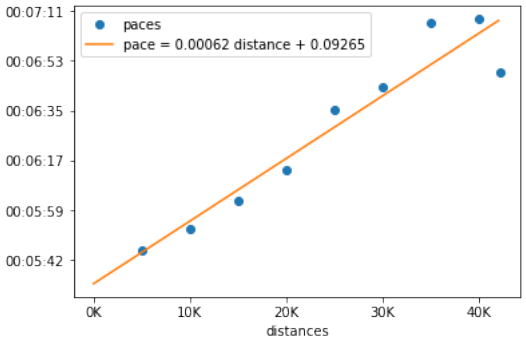


Figure 8. Average pace vs distance with line of best fit

it does have a very high correlation of $0.954$, and actually fits much better than a quadratic or cubic regression.

All this regression really tells us is that runners slow down towards the end of the race, and we already new that from the previous slowdown analysis. Is there a way we can predict the extent to which runners will slow down in a given race?

We next look at the impact that temperature has on a runner's performance. Figure 9 shows what temperature it was on The Mall, London at the time when the runner finished the race. This temperature is plotting against the total time it took the runner to complete the race.

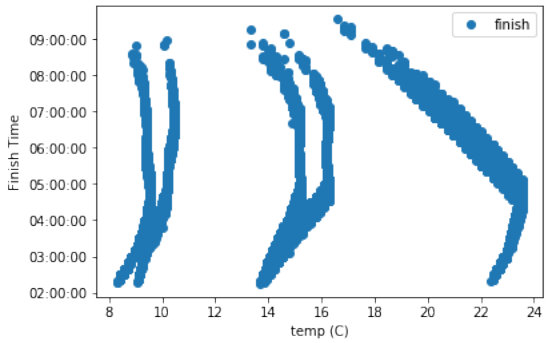


Figure 9. Temperature vs Finish Time for all runners

Figure 9 is ugly and there does not look to be any clear pattern, except that there are five apparent clusters of points. These apparent clusters are the different years of the race in the dataset, clustered in such a way because the temperature ranges for each race is different. If we separate out each year and run a linear regression we achieve strong results. Figure 10 shows the plot for 2015, with a line of best fit given by $f = 2.7t - 22.0$ (f is finish time and t is temperature) and a correlation coefficient of 0.948.

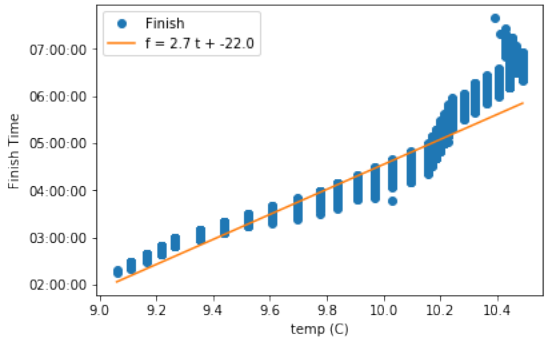


Figure 10. Temperature vs Finish Time for 2015

Results are similarly strong for the other years, however at this stage it is important to question our assumptions. Does the data really prove that temperature is a good prediction of runners performance? For this dataset that appears to be the case, but generally speaking it is more likely that temperatures increase throughout the day, and slower runners finish at a later time of day (as they take longer to complete the race).

To investigate further if temperature is a good indication of performance, we look at slowdown instead of total race time. In table 4.2 linear regressions for slowdown (s) as a function of temperature (t) are presented alongside their respective correlation coefficients.

| Year | Line | Correlation |
|---|---|---|
| 2014 | s = 12.3 t + -166.5 | 0.36 |
| 2015 | s = 22.3 t + -201.5 | 0.35 |
| 2016 | s = 23.9 t + -203.1 | 0.27 |
| 2017 | s = 18.0 t + -242.7 | 0.32 |
| 2018 | s = -1.9 t + 75.0 | 0.08 |

These correlations are quite low, and the only point worth noting is that in 2018 the correlation between finish time and temperature is exceptionally low. This might be counter intuitive as 2018 was the hottest year - but may indicate that in extreme temperatures everyone reacts differently to the heat. Quadratic regression (slowdown as a function of temperature and temperature$^2$) gives equally poor correlations, as can be seen in the following table.

So far, we are have been looking at the Slowdown factor (ie the change in pace of a runner throughout the race as a percentage of their pace at the 5km mark) as a continuous variable. Instead, we chose to categorise each runner depending on their Slowdown factor. The categories we chose were:

| Year | Correlation |
|---|---|
| 2014 | 0.13 |
| 2015 | 0.13 |
| 2016 | 0.07 |
| 2017 | 0.11 |
| 2018 | 0.01 |

- Sped up
- Pace slowed by up to 5%
- Pace slowed by between 5% and 10%
- Pace slowed by between 10% and 50%
- Pace slowed by between 50% and 100%
- Pace slowed by over 100%

We then employed a logistic regression to predict which slowdown group a runner is in, given the temperature that they ran at. This was much more successful than a linear regression, with a correlation coefficient of 0.822. The authors hypothesised that the correlation coefficient might be dependent on how experienced the runner is. That is, runners who are more experienced are likely to be less effected by temperature. To test this hypothesis, the data was split by Speed Group and a logistic regression was run for each subset of the data.

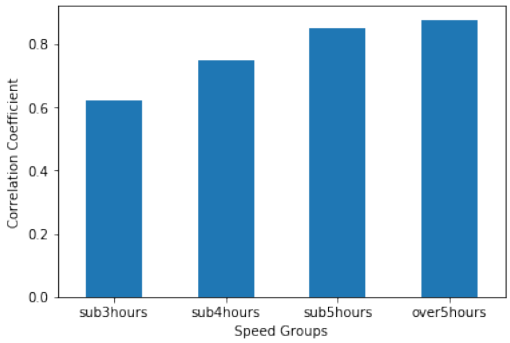


Figure 11. Correlation Coefficients for using Temperature to predict Slowdown Group, split by Speed Group

Slowdown for all speed groups have some correlation to temperature. That said, we can see from figure 11 that the slowdown of faster speed groups is less correlated to temperature. This indciates that the hypothesis is correct. Runners who completed the marathon in under 3 hours have a correlation coefficient of 0.62, whereas those runners who took over 5 hours to complete the marathon have a correlation coefficient of 0.88.

## 4.3. Classification

**4.3.1. By Finish Time.** For a classification analysis, we first look at the overall finish time of the runner. The groupings for finish time are the Speed Groups described above. We

attempted to classify which speed group a runner is in using only:

- Year (of race)
- Gender
- Temperature (at the finish line)
- Slowdown Group

We excluded all other variables for a number of different reasons. The runner's Name and Place were discarded as they contained very specific information. This caused every classifier to overfit the data: performing very well on the training data but poorly on the test data. The runner's finish time and split paces were discarded as they are used as direct inputs for the runners Speed Group. Therefore, these variables caused the classifier to predict the Speed Group too accurately without adding any additional insight.

Each classifier was evaluated using a 10-fold cross validation as a precaution against overfitting. First of all, we ran a majority classifier on the data. As the average finish time is 4 hours, 31 minutes and 4 seconds, intuitively the largest Speed Group is sub 5 hours (ie finish time between 4 and 5 hours). The majority classifier gave a 38.19% accuracy. The Kappa statistic is 0.0, as expected.

Next we ran Naive Bayes on the same variables, again to predict which Speed Group the runner is in. This had a slightly better accuracy of 45.93% but still performed poorly. This poor performance can be seen in the confusion table:

| | Classified as: | | | |
|---|---|---|---|---|
| | sub3hours | sub4hours | sub5hours | over5hours |
| sub3hours | 0 | 5315 | 2384 | 680 |
| sub4hours | 0 | 19725 | 25874 | 8814 |
| sub5hours | 0 | 11561 | 43011 | 17936 |
| over5hours | 0 | 2393 | 27699 | 24453 |

This Naive Bayes classifier does not classify any runners as sub3hours, even though many runners are in this category. Furthermore, the classifier does well by just putting most of the runners in the sub5hours class and so is really not that much better than the majority classifier. The Kappa statistic of 0.185 reinforces this.

Finally we use a J48 decision tree to classify which Speed Group a runner is in. This classifier performs significantly better, with an accuracy of 93.20% and a Kappa statistic of 0.901. This tree is large (size of 215) and it does have a large number of leaves (136), which indicates that there may be some overfitting. However, the above accuracy is once again evaluated with 10-fold cross validation. The higher branches on the tree divide the data into those with a high temperature and those with a low temperature. Temperature is used as an informative feature throughout the tree with the lower branches making decisions with the Slowdown Groups. The tree is too large to include here. The confusion table is below.

This confusion table has much higher values on the diagonal, and when it does error then the errors are close to the diagonal. This indicates that there is good logic involved, not just random guessing.

| | Classified as: | | | |
|---|---|---|---|---|
| | sub3hours | sub4hours | sub5hours | over5hours |
| sub3hours | 7736 | 642 | 0 | 1 |
| sub4hours | 0 | 53001 | 1412 | 0 |
| sub5hours | 0 | 3278 | 66689 | 2541 |
| over5hours | 0 | 69 | 4958 | 49518 |

A criticism of this classification is that temperature is heavily correlated with Time of Day, which in turn is strongly correlated with the total finish time for a given runner. For this reason, it may be the case that temperature does not *cause* runners to slow down, it is just correlated.

**4.3.2. By Pacing Ability.** Because of the above criticism, we turn out attention to classifying runners based on their ability to keep running the same pace that they started out at. The groupings for this are given by Slowdown Groups, which are groupings derived from Slowdown factor, as described above. As before we use Year, Gender and Temperature to predict the correct class. We also us the runner's Speed Group as a feature here.

Unfortunately, classification efforts for Slowdown Group were not as impressive as for Speed Group. To demonstrate this, we highlight that the majority classifier gives an accuracy of 67.60% by classifying every runner as having a Slowdown of between 10 and 50%. The frequency distribution of the Slowdown Classes is given below.

| Slowdown Group | Frequency |
|---|---|
| negative | 20,753 |
| to5 | 15,865 |
| to10 | 15,315 |
| to50 | 126,770 |
| to100 | 15,315 |
| over100 | 870 |

It is clear from this table that the vast majority of runners slowed down by between 10% and 50%.

Employing Naive Bayes to classify Slowdown Groups gives an accuracy of 67.67%, only marginly higher than the majority classifier. It has a Kappa of 0.05, so it is not really any better than guessing. This is further demonstrated by the confusion matrix below, almost all runners have been classified as the majority class.

As with Speed Group, we use a J48 decision tree to classify the Slow Down Group of a runner. Here, the decision tree classifier is just as poor as the Naive Bayes classifier, with an accuracy of 67.74% and a Kappa of 0.04. The J48

| | classified as: | | | | | |
|---|---|---|---|---|---|---|
| | to10 | to5 | to50 | negative | to100 | over100 |
| to10 | 0 | 1360 | 19527 | 0 | 0 | 0 |
| to5 | 0 | 1745 | 14188 | 0 | 0 | 0 |
| to50 | 0 | 1602 | 126729 | 0 | 0 | 0 |
| negative | 0 | 624 | 7553 | 0 | 0 | 0 |
| to100 | 0 | 16 | 15615 | 0 | 0 | 0 |
| over100 | 0 | 0 | 886 | 0 | 0 | 0 |

decision tree classifier's confusion matrix is omitted here, but looks similar to the Naive Bayes confusion matrix.

## 5. Extensions and Improvements

Given more time, there are many extensions that the authors would like to make to this work.

1) Firstly, the paper contains no analysis of the 3 hour, 4 hour or Boston Marathon Qualifier barriers. It would be interesting to see if the data reflects the authors' belief that runners finish times cluster around these points. Despite the authors' intentions, the paper contains only a small amount of analysis of the performance of runners with respect to their different nationalities. Weather data was included, but only the temperature and not any other weather factors such as precipitation or humidity.
2) In addition to including more types of weather, this work could be extended by including temperature measurements from throughout the race or potentially even the weather in the weeks leading up to each race. Naturally, the quality of the weather analysis would be improved if there was race data available from more years of the London Marathon. With only 5 years worth of data, there is hardly statistical significance to make any conclusions about the impact of temperature on a runner's performance.
3) It would be interesting to add data from marathons in other places such as New York City marathon, or perhaps a non-city marathon as a contrast. It would also be interesting to include data from races that are a shorter distance to a marathon and compare how runners' paces changes over those distances.
4) Adding race data from other races would not only increase the sample size of the dataset, it would reveal which runners have run multiple marathons before and would open the possibility of measuring a runner's experience. The author hypothesises that a runners experience is negatively correlated with their Slowdown Factor. That is, more experienced runners can hold their pace better.

## 6. Conclusion

The aim of this paper was to investigate the finish times for runners who completed the London Marathon in the last 5 years. Additionally, the paper highlighted that most runners' pace slows down throughout the marathon and aimed to investigate the causes of this slowdown. The average runner's pace at the 5km mark is 24.25% slower by the time they reach the 40km mark, but this varies by the runners total race time and, in our dataset, varied depending on the year of the race. With an average temperature of over 22 degrees celsius, 2018 was by far the hottest race in our dataset and this is reflected in the runners' results.

As far as regressions go, there is a strong correlation between finish time and temperature at the finish line, but this may be caused by time of day. Given more time, the author would like to investigate if finish time and temperature are still correlated once time of day has been accounted for. Looking at Slowdown (continuous variable) as a function of temperature shows that in general, there is a weak correlation of around 0.30. However, if the runners are divided into groups based on their Slowdown Factor then this correlation jumps significantly to 0.822. Further dividing the runners into Speed Groups highlights that the correlation between Slowdown Factor and temperature increases as a runner's finish time increases. This is intuitive.

When classifying the runners to predict their Speed Group, the J48 decision tree performed very well. It is possible that its strong performance is caused by the correlation between temperature and time of day, as with regression. For predicting a runner's Slowdown Group, the authors found no classifier that performed well.

The authors hope to continue their investigation into what causes runners to slow down throughout a race such as the London Marathon.

## References


[1] "The london marathon: In the beginning." [Online]. Available: https://www.virginmoneylondonmarathon.com/en-gb/news-media/media-resources/history-london-marathon/in-the-beginning/

[2] "How it works." [Online]. Available: https://www.worldmarathonmajors.com/about/how-it-works/

[3] "Iaaf: Radcliffe runs 2:15:25 in london!— news," Apr 2003. [Online]. Available: https://www.iaaf.org/news/news/radcliffe-runs-21525-in-london

[4] "Latest news london marathon sets another world record." [Online]. Available: https://www.virginmoneylondonmarathon.com/en-gb/news-media/latest-news/item/london-marathon-sets-another-world-record/

[5] "Latest news record finish as largest london marathon displays its spirit." [Online]. Available: https://www.virginmoneylondonmarathon.com/en-gb/news-media/latest-news/item/record-finish-as-largest-london-marathon-displays-its-spirit/

[6] "Finishers boston marathon 2015, 2016 2017," Apr 2017. [Online]. Available: https://www.kaggle.com/rojour/boston-results

[7] B. Smyth, "Perfect pacing at the london marathon – running with data – medium," Mar 2017. [Online]. Available: https://medium.com/running-with-data/perfect-pacing-at-the-london-marathon-95f09db2ac5c

[8] "Strava reveals 2016 london marathon data." [Online]. Available: https://www.runnersworld.co.uk/strava-reveals-2016-london-marathon-data

[9] "London marathon results search." [Online]. Available: http://results-2018.virginmoneylondonmarathon.com/2018/

[10] "Dark sky api - overview." [Online]. Available: https://darksky.net/dev/docs