Data Mining
Exploration of race time predictors

Benjamin Baxter
Daniel Westerlund

April 2016

1 Introduction

This project is focusing on the data from runners, finding interesting patterns in this data and improving on existing formulas for predicting race times given distance and time from previous races by including additional information about the heart rate and speed from these races.

1.1 Predicting race time

There are many ways of predicting the race time for a given distance, and many free race time calculators can be found online. These calculators use different information to predict the race time, such as time and distance from previous races, maximum heart rate, resting heart rate, weight, age, sex, etc. One simple way of predicting the race time is the Peter Riegel formula [Riegel, 1977] shown in Equation 1 where T_1 and D_1 is the time and distance from a previous race, and T_2 is the predicted time for distance D_2.

$$T_2 = T_1 \cdot (\frac{D_2}{D_1})^{1.06} \tag{1}$$

Another way of predicting race time is by using the runner's maximum oxygen consumption $\dot{V}O_2max$ or VDOT [Daniels, 2013], which can serve as a measurement of the cardiovascular fitness of the runner. The VDOT value for an individual can be estimated using different tests, such as the Cooper test and beep-test. Equation 2 shows one way of calculating VDOT proposed by Daniels [Daniels and Gilbert, 1979], where t is the time, in minutes, and d is the distance, in meters, from a previous race.

$$VDOT = \frac{-4.6 + 0.182258 * (\frac{d}{t}) + 0.000104 * (\frac{d}{t})^2}{0.8 + 0.1894393 * e^{-0.012778*t} + 0.2989558 * e^{-0.1932605*t}} \tag{2}$$

In another work by Daniels [Daniels, 2013] a table is provided with running times for certain VDOT values of popular distances shown in Appendix A.

When a runner is recording his or her sessions the recorded performance of the sessions could vary a lot. This could be due to the user forgetting to turn off the device and getting into a car, or resting, or just having a bad performance during that particular session and walking in combination with running. These sessions can be identified and removed as outliers, but that should not always be necessary. If a user has one low-performance session and multiple regular-performance sessions the low-performance session will have a bias and overestimate the time if it is used to predict the running times for the other, regular-performance sessions. If multiple sessions are used in a combination to predict a race time for a given distance one straight forward way of doing it is to simply take the mean value of the predictions. Some races will be better suitable for predicting than others, so every session will have a prediction uncertainty. If this bias and uncertainty can be identified using the speed and heart rate measurements the bias can be compensated for and the uncertainty can be used in a weighted average, giving a better prediction.

2 Data from runners

The data used was obtained from "ChenHou.zip" and "Danijo11-2.zip". This is data from runners with no missing values on either sex and age. A request was made to obtain the .json-files containing all runners with more than 72 logged sessions with a duration longer than 60 minutes, which corresponded to 100 users with 12000 sessions combined. Out of these sessions 3000 were provided in the "Danijo11-2.zip".

2.1 Refinement of data

A refinement of the combination of the sessions from "ChenHou.zip" and "Danijo11-2.zip" was done by filtering out:

- Sessions with no records.
- Sessions with less than 40 timestamps for speed and heart rate.
- Sessions where the difference between the duration obtained from the device and the user was larger than a factor of 4.
- Sessions where the average speed or heart rate was superhuman for the sport of running at the given distance.

After the refinement the data contained 109 users with a number of sessions ranging from 3 to 209 with a mean number of sessions of 23.5. Out of these users, 22 (20.18% of original data) users were randomly selected to form a test set, while the remaining 87 formed a training set. For the sessions in the training set the mean value and variance of the speed data and heart rate data from the time stamps were extracted.

3 Data Visualization tools

Two visualization tools, Weka and IBM Watson Analytics, were used to find interesting patterns and clusters in the data.

3.1 Weka

Based on the assumption that Session ID was generated as an auto increment as logs were submitted it was assumed that it would be a good indicator for timeline on data entry. When the sessions were plotted against the user submitting it was found that not all entries were submitted in a regular pattern and that certain users in fact would submit several dozen session all at once.

Figure 1: SessionID and PersonID.

This can be seen in Figure 1 as the gaps and solid lines on the y axis of single users. The slope of the line comes from new users joining the site and old users ceasing to upload new sessions.

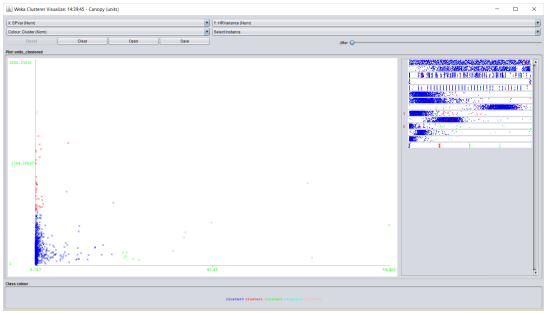

Figure 2: Heart rate variance and speed variance.

An outlier detection using a simple cluster recognition on the calculated heart rate variance and speed variance in the session is shown in Figure 2. This was possible because outside of a certain range the variance was too high to be explained by variance in users and had to be a result of extraneous or incomplete data. For example having a heart rate variance of greater than 100 beats per minute would result in a dead user, therefore we can assume that session with these variances are "noise". Similar reasoning can be made with speed.

To experiment with isolating and predicting races a single user was selected to run a series of equations and predictions on. This was done by running the VDOT equation on each race and then using the number to predict other races and generate a error variance in the VDOT vs real sessions.

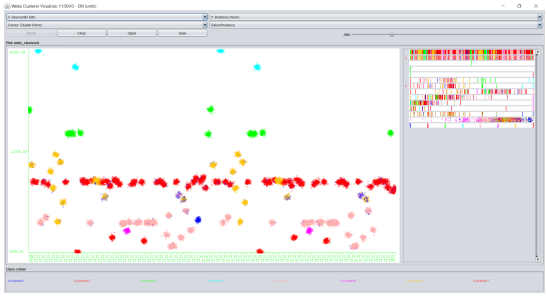

Figure 3: Distances over time (SessionID) for a single user.

Figure 3 shows what appears to be a training schedule for a marathon runner. Red is base line training of 10 km and pink is 5 km with green as 21 km. Light blue is 42 and 50 km which represents marathons over approx a year, so a summer and spring competition. Yellow and dark pink was for sessions with abnormal heart rate and speed variations as outliers were included in the data. In Figure 4 the distance is plotted against the duration and can visualize a line that closely resembles the type you would see from a VDOT plotted graph.

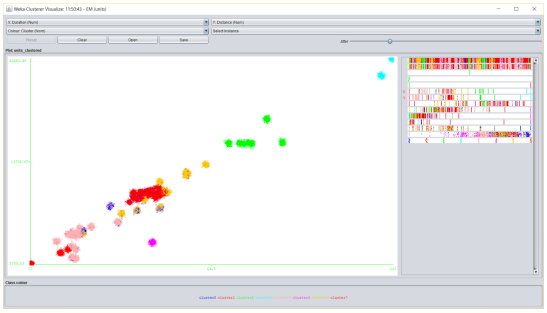

Figure 4: Distance over duration for a single user.

3.2 IBM Watson Analytics

The data was further analyzed using IBM Watson Analytics. Figure 5 and Figure 6 shows the minimum and maximum values for speed and heart rate respectively for users. These bar charts suggest that some outliers were present in the data due to the abnormal values for maximum and minimum speed and heart rate. Figure 7 shows the distribution of number of sessions for some users.

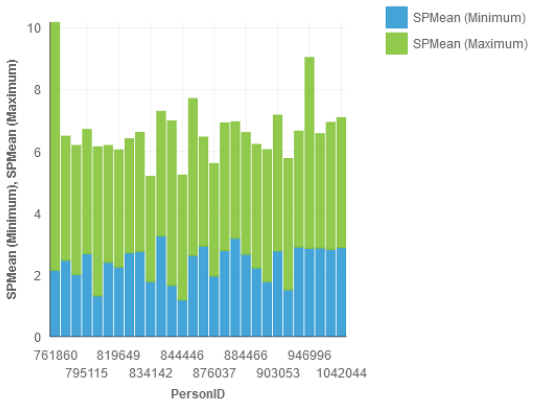

Figure 5: Min and Max of speed metrics by user.

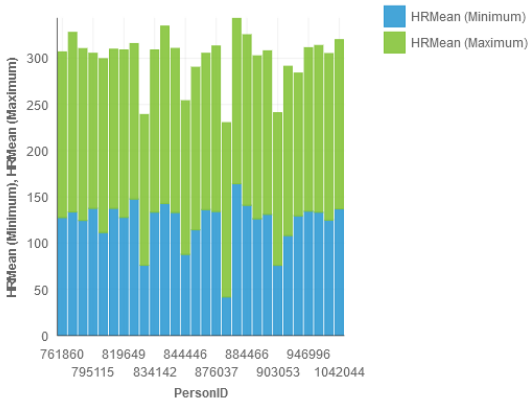

Figure 6: Min and Max of heart rates by user.

Figure 8 shows the duration by average distance for different users. This plot clearly shows a gap between two types of users. The group of users closer to the bottom left represent shorter distance or regular runners whereas the group closer to the upper right correspond to long distance and marathon runners. In figure 9 instead of using average values the distances are summated and reflects the same pattern that again match what a VDOT slope shound follow, the exeption being that this data set was representative of multiple users instead of one.

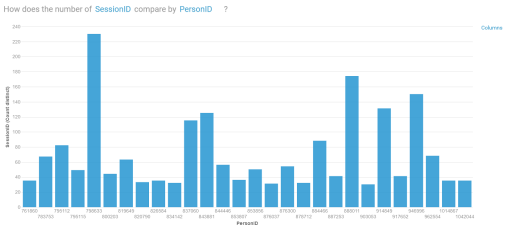

Figure 7: Distribution of session count of users.

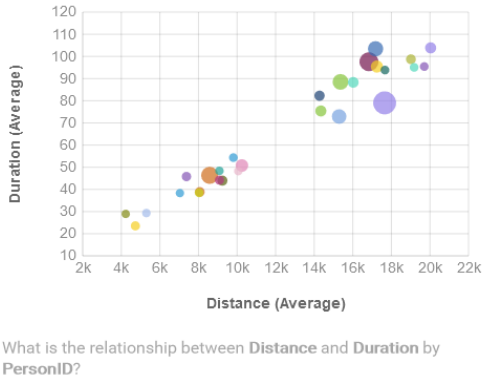

Figure 8: Duration by average distance with color representing user and size representing session count.

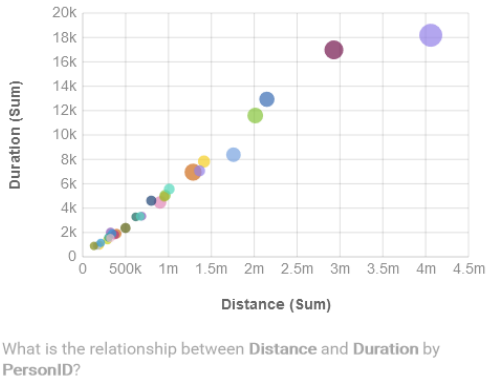

Figure 9: Duration by average distance with color representing user and size representing session count.

4 Predicting bias and uncertainty

4.1 Finding the race time

Predicting the time for a new race using the Riegel method (1.1.1) is a one-step process where a previous race is used to predict the new time. The Daniels VDOT method (1.1.2) is a two-step process, where the VDOT is calculated from a previous race and then the new time can be found using the table in Appendix A. As the VDOT values and distances in the table are discrete, and the VDOT value obtained from the formula as well as the desired distances are continuous a continuous function for predicting the race time given a distance and value of VDOT is required. An exponential function can be found for each value of VDOT by taking the natural logarithm of both the time and the distance, and applying a linear fit. Figure 10 shows an example of this exponential function for the VDOT-value of 64.

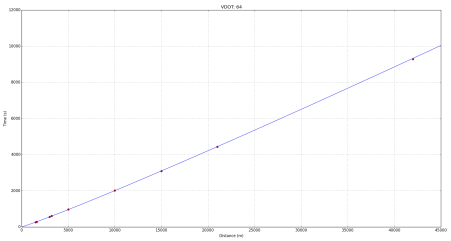

Figure 10: Exponential function of time and distance for VDOT value 64.

As the two parameters for the linear equation, Slope and y-intercept, are obtained for each equation of VDOT these can in turn be found for a continuous value of VDOT. Figure 11 and Figure 12 shows the slopes and y-intercepts of these equations plotted against the VDOT values as well as third-order polynomial fits for these points. So by finding the VDOT value from a race the slope and y-intercept can be found, which then is used in the exponential function to find the predicted time at a given distance.

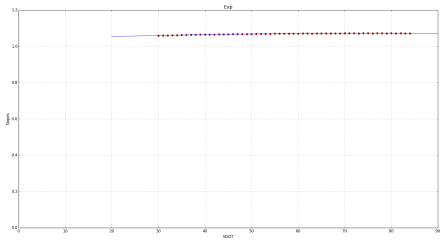

Figure 11: Slopes for the linear equation plotted against VDOT values

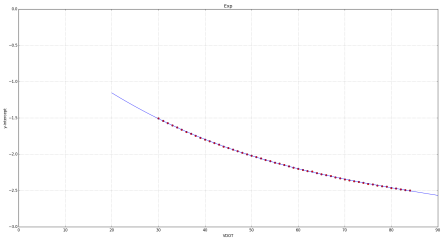

Figure 12: Y-intercepts for the linear equation plotted against VDOT values

4.2 Calculating the prediction errors

The sessions of every user in the training set are contained in two sets, S and S'. The set S contains all the sessions for this user, and is used to predict the race times of the sessions in set S', which contains all sessions where the difference in the calculated VDOT of the session and the mean value of the remaining sessions is within 10 units of VDOT. Every session in S is used to predict the time for all sessions in S' using both the Riegel method and the Daniels VDOT method, and a percentage error for each predicted session using both predictions are calculated as follows:

$$Error = \frac{PredictedTime - TrueTime}{TrueTime} \quad (3)$$

Now every session in S will have a set of prediction errors corresponding to the sessions in S'. The mean value of these predictions represent the prediction bias of the session, and the variance in the predictions represent the prediction uncertainty of this session.

4.3 Predicting the errors

For every session in S the deviation from the mean of the remaining sessions is calculated for:

- Heart rate mean value
- Variance in heart rates
- Speed mean value
- Variance in speed

Multiple linear regression is then used to predict the bias and uncertainty for the Riegel and Daniels predictions as linear functions of these deviations.

5 Results

The 22 users in the test set were used to calculate a performance measure of the bias compensation and weighted average. For every user one random session was selected to be used for time prediction using the remaining sessions from the same user. An absolute percentage prediction error was calculated as follows:

$$Error = \frac{|PredictedTime - TrueTime|}{TrueTime} \quad (4)$$

This was done for eight methods:

- R
  The predicted time is the mean value of the predictions using the Riegel formula.
- D
  The predicted time is the mean value of the predictions using the Daniels VDOT formula.
- RD
  The predicted time is the mean value of the predicted times in R and D.
- R*
  The predicted time is the weighted average of the predictions using the Riegel formula and weighted average using the predicted uncertainty.
- D*
  The predicted time is the weighted average of the predictions using the Daniels VDOT formula and weighted average using the predicted uncertainty.
- RD*
  The predicted time is the weighted average of the predicted times in R* and D* and weighted average using the propagated uncertainty.
- BC
  The predicted time is the mean value of the predictions in R and D, with bias compensation.
- BC*
  The predicted time is the mean value of the predictions in R and D, with bias compensation and weighted average using the propagated uncertainty.

The process of selecting a random sessions for all users was repeated 1000 times and the results are shown as percentages in the table below.

UserID	R	D	RD	R*	D*	RD*	BC	BC*
'876300'	6.021	5.911	5.966	5.477	5.506	5.463	7.413	5.534
'827528'	197.878	45.093	121.485	66.239	21.868	39.805	112.448	38.026
'862404'	11.483	11.488	11.485	10.364	10.504	10.443	13.63	10.694
'903053'	6.306	6.311	6.308	6.154	6.199	6.18	6.532	6.281
'917652'	4.791	4.508	4.63	4.515	4.351	4.384	4.891	4.366
'900859'	12.893	12.494	12.694	11.494	11.225	11.327	15.659	11.701
'888011'	5.673	5.702	5.687	6.299	6.132	6.194	6.06	6.295
'880467'	1.754	1.805	1.779	1.764	1.788	1.778	1.922	1.718
'841834'	26.726	22.118	24.422	13.927	11.14	12.24	37.388	13.017
'962855'	7.059	7.08	7.07	5.658	5.576	5.583	7.882	5.508
'786979'	13.041	12.456	12.748	10.068	10.026	10.03	13.839	10.137
'822461'	5.029	5.084	5.056	5.611	5.676	5.652	6.035	7.072
'783753'	9.293	9.368	9.331	9.146	9.197	9.183	9.948	9.371
'917904'	3.93	4.285	4.107	3.699	3.709	3.702	4.436	3.526
'896882'	3.405	3.535	3.47	3.266	3.407	3.35	3.659	3.422
'884466'	7.581	7.629	7.605	7.134	7.211	7.184	8.276	7.276
'939847'	3.259	3.382	3.299	2.956	2.981	2.97	3.52	3.047
'831024'	5.369	5.353	5.361	4.967	4.978	4.973	7.05	5.329
'807532'	14.898	15.098	14.998	14.817	15.01	14.934	25.892	25.265
'872623'	24.106	20.533	22.319	16.169	15.738	15.913	30.005	16.005
'908475'	7.088	6.96	7.024	7.122	7.027	7.066	7.645	7.388
'882573'	6.446	8.441	7.444	5.562	6.995	6.439	11.471	7.883
TOTAL	17.456	10.211	13.831	10.109	8.011	8.854	15.709	9.494

6 Conclusions

For most users the errors are lower when using the weighted average for predictions compared to the mean value. The performance of the bias compensation, especially without weighted average is very poor, suggesting that another approach and model may be suitable for predicting the bias for a given race.

It would be interesting to focus more closely on the heart rate and speed data from the time stamps than simply taking the mean and variance of the measurements, as this may be a better indication of the performance of the sessions. These predictions are also on all kinds of sessions, both exercise and race. One future improvement could be to look at the users taking part in the popular races in Sweden, such as Göteborgsvarvet and Stockholm Marathon, classify these sessions as actual races and look at the similarity in the performances of these sessions compared to others. It could also be interesting to look at data from a larger time span and identify performance increases and decreases for individuals, as more recent sessions would be more indicative of the timing in a new race if the performance has changed a lot during the time span.

The "1.06"-parameter in the Riegel formula, which correspond to the average speed decrease for longer distances, could also be optimized for each individual and a study of which personal attributes, e.g. age, sex, drives this optimized value can be done.

This project focused only on the data from the .json-files, which are obtained from one device type, while the .tcx-files are from another device. There might be differences in the performances of the runners using the different devices, which could be studied further.

References

[Daniels, 2013] Daniels, J. (2013). *Daniels' running formula*. Human Kinetics.

[Daniels and Gilbert, 1979] Daniels, J. and Gilbert, J. (1979). *Oxygen power: Performance tables for distance runners*. J. Daniels, J. Gilbert.

[Riegel, 1977] Riegel, P. (1977). Time predicting. *Runner's World*, 12:61–64.

Appendices

A

VDOT Values Associated With Running Times of Popular Distances

VDOT	1500m	mile	3000m	2 mile	5K	10K	15K	Half Marathon	Marathon
30	8:30	9:11	17:56	19:19	30:40	63:46	98:14	2:21:04	4:49:17
31	8:15	8:55	17:27	18:48	29:51	62:03	95:36	2:17:21	4:41:57
32	8:02	8:41	16:59	18:18	29:05	60:26	93:07	2:13:49	4:34:59
33	7:49	8:27	16:33	17:50	28:21	58:54	90:45	2:10:27	4:28:22
34	7:37	8:14	16:09	17:24	27:39	57:26	88:30	2:07:16	4:22:03
35	7:25	8:01	15:45	16:58	27:00	56:03	86:22	2:04:13	4:16:03
36	7:14	7:49	15:23	16:34	26:22	54:44	84:20	2:01:19	4:10:19
37	7:04	7:38	15:01	16:11	25:46	53:29	82:24	1:58:34	4:04:50
38	6:54	7:27	14:41	15:49	25:12	52:17	80:33	1:55:55	3:59:35
39	6:44	7:17	14:21	15:29	24:39	51:09	78:47	1:53:24	3:54:34
40	6:35	7:07	14:03	15:08	24:08	50:03	77:06	1:50:59	3:49:45
41	6:27	6:58	13:45	14:49	23:38	49:01	75:29	1:48:40	3:45:09
42	6:19	6:49	13:28	14:31	23:09	48:01	73:56	1:46:27	3:40:43
43	6:11	6:41	13:11	14:13	22:41	47:04	72:27	1:44:20	3:36:28
44	6:03	6:32	12:55	13:56	22:15	46:09	71:02	1:42:17	3:32:23
45	5:56	6:25	12:40	13:40	21:50	45:16	69:40	1:40:20	3:28:26
46	5:49	6:17	12:26	13:25	21:25	44:25	68:22	1:38:27	3:24:39
47	5:42	6:10	12:12	13:10	21:02	43:36	67:06	1:36:38	3:21:00
48	5:36	6:03	11:58	12:55	20:39	42:50	65:53	1:34:53	3:17:29
49	5:30	5:56	11:45	12:41	20:18	42:04	64:44	1:33:12	3:14:06
50	5:24	5:50	11:33	12:28	19:57	41:21	63:36	1:31:35	3:10:49
51	5:18	5:44	11:21	12:15	19:36	40:39	62:31	1:30:02	3:07:39
52	5:13	5:38	11:09	12:02	19:17	39:59	61:29	1:28:31	3:04:36
53	5:07	5:32	10:58	11:50	18:58	39:20	60:28	1:27:04	3:01:39
54	5:02	5:27	10:47	11:39	18:40	38:42	59:30	1:25:40	2:58:47
55	4:57	5:21	10:37	11:28	18:22	38:06	58:33	1:24:18	2:56:01
56	4:53	5:16	10:27	11:17	18:05	37:31	57:39	1:23:00	2:53:20
57	4:48	5:11	10:17	11:06	17:49	36:57	56:46	1:21:43	2:50:45
58	4:44	5:06	10:08	10:56	17:33	36:24	55:55	1:20:30	2:48:14
59	4:39	5:02	9:58	10:46	17:17	35:52	55:06	1:19:18	2:45:47
60	4:35	4:57	9:50	10:37	17:03	35:22	54:18	1:18:09	2:43:25
61	4:31	4:53	9:41	10:27	16:48	34:52	53:32	1:17:02	2:41:08
62	4:27	4:49	9:33	10:18	16:34	34:23	52:47	1:15:57	2:38:54
63	4:24	4:45	9:25	10:10	16:20	33:55	52:03	1:14:54	2:36:44
64	4:20	4:41	9:17	10:01	16:07	33:28	51:21	1:13:53	2:34:38
65	4:16	4:37	9:09	9:53	15:54	33:01	50:40	1:12:53	2:32:35

A

VDOT Values Associated With Running Times of Popular Distances

VDOT	1500m	mile	3000m	2 mile	5K	10K	15K	Half Marathon	Marathon
66	4:13	4:33	9:02	9:45	15:42	32:35	50:00	1:11:56	2:30:36
67	4:10	4:30	8:55	9:37	15:29	32:11	49:22	1:11:00	2:28:40
68	4:06	4:26	8:48	9:30	15:18	31:46	48:44	1:10:05	2:26:47
69	4:03	4:23	8:41	9:23	15:06	31:23	48:08	1:09:12	2:24:57
70	4:00	4:19	8:34	9:16	14:55	31:00	47:32	1:08:21	2:23:10
71	3:57	4:16	8:28	9:09	14:44	30:38	46:58	1:07:31	2:21:26
72	3:54	4:13	8:22	9:02	14:33	30:16	46:24	1:06:42	2:19:44
73	3:52	4:10	8:16	8:55	14:23	29:55	45:51	1:05:54	2:18:05
74	3:49	4:07	8:10	8:49	14:13	29:34	45:19	1:05:08	2:16:29
75	3:46	4:04	8:04	8:43	14:03	29:14	44:48	1:04:23	2:14:55
76	3:44	4:02	7:58	8:37	13:54	28:55	44:18	1:03:39	2:13:23
77	3:41+	3:58+	7:53	8:31	13:44	28:36	43:49	1:02:56	2:11:54
78	3:38.8	3:56.2	7:48	8:25	13:35	28:17	43:20	1:02:15	2:10:27
79	3:36.5	3:53.7	7:43	8:20	13:26	27:59	42:52	1:01:34	2:09:02
80	3:34.2	3:51.2	7:37.5	8:14.2	13:17.8	27:41	42:25	1:00:54	2:07:38
81	3:31.9	3:48.7	7:32.5	8:08.9	13:09.3	27:24	41:58	1:00:15	2:06:17
82	3:29.7	3:46.4	7:27.7	8:03.7	13:01.1	27:07	41:32	:59:38	2:04:57
83	3:27.6	3:44.0	7:23.0	7:58.6	12:53.0	26:51	41:06	:59:01	2:03:40
84	3:25.5	3:41.8	7:18.5	7:53.6	12:45.2	26:34	40:42	:58:25	2:02:24
85	3:23.5	3:39.6	7:14.0	7:48.8	12:37.4	26:19	40:17	:57:50	2:01:10