COMPLETING THE RESULTS OF THE 2013 BOSTON MARATHON

Dorit Hammerling[1], Matthew Cefalu[2], Jessi Cisewski[3], Francesca Dominici[2], Giovanni Parmigiani[2,4], Charles Paulson[5], Richard Smith[1,6]

[1]Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, North Carolina; [2]Department of Biostatistics, Harvard School of Public Health; [3]Department of Statistics, Carnegie Mellon University; [4]Dana Farber Cancer Institute, Boston ; [5]Puffinware LLC; [6]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill.

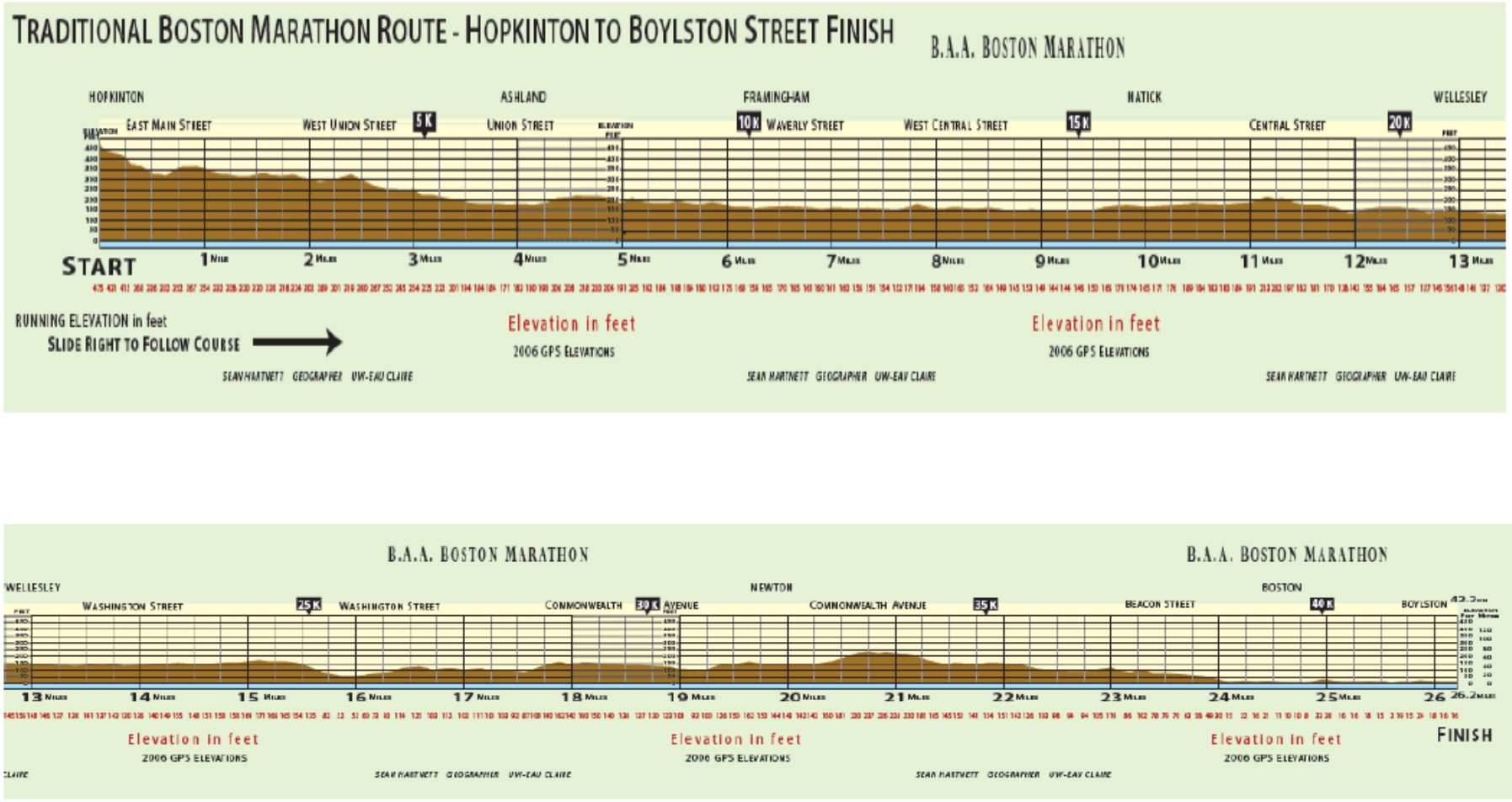


Figure 1: Profile of the Boston Marathon course (source: BAA)

26.2 miles (42.2 km.), Hopkinton to Boston

From: Michael Pieroni [Pieroni@baa.org]
Sent: Tuesday, April 23, 2013 12:26 PM
To: Smith, Richard L
Subject: Boston 2013

Richard-

Hope all is well with you.

I am writing to ask for some guidance.
We are attempting to assemble marathon results for those runners who were stopped prior to the completion of the full marathon distance.

We have split information for the 5000+ of them through 30k to 40k marks

Is there a method that we can use to accurately (as best as possible) a possible finish performance for these individuals?

Give it some thought and let me know what you think.

Michael Pieroni

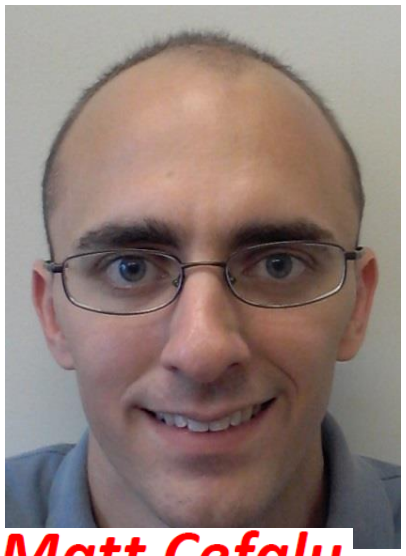
Dorit Hammerling

Francesca Dominici

Giovanni Parmigiani

Jessi Cisewski

Chuck Paulson

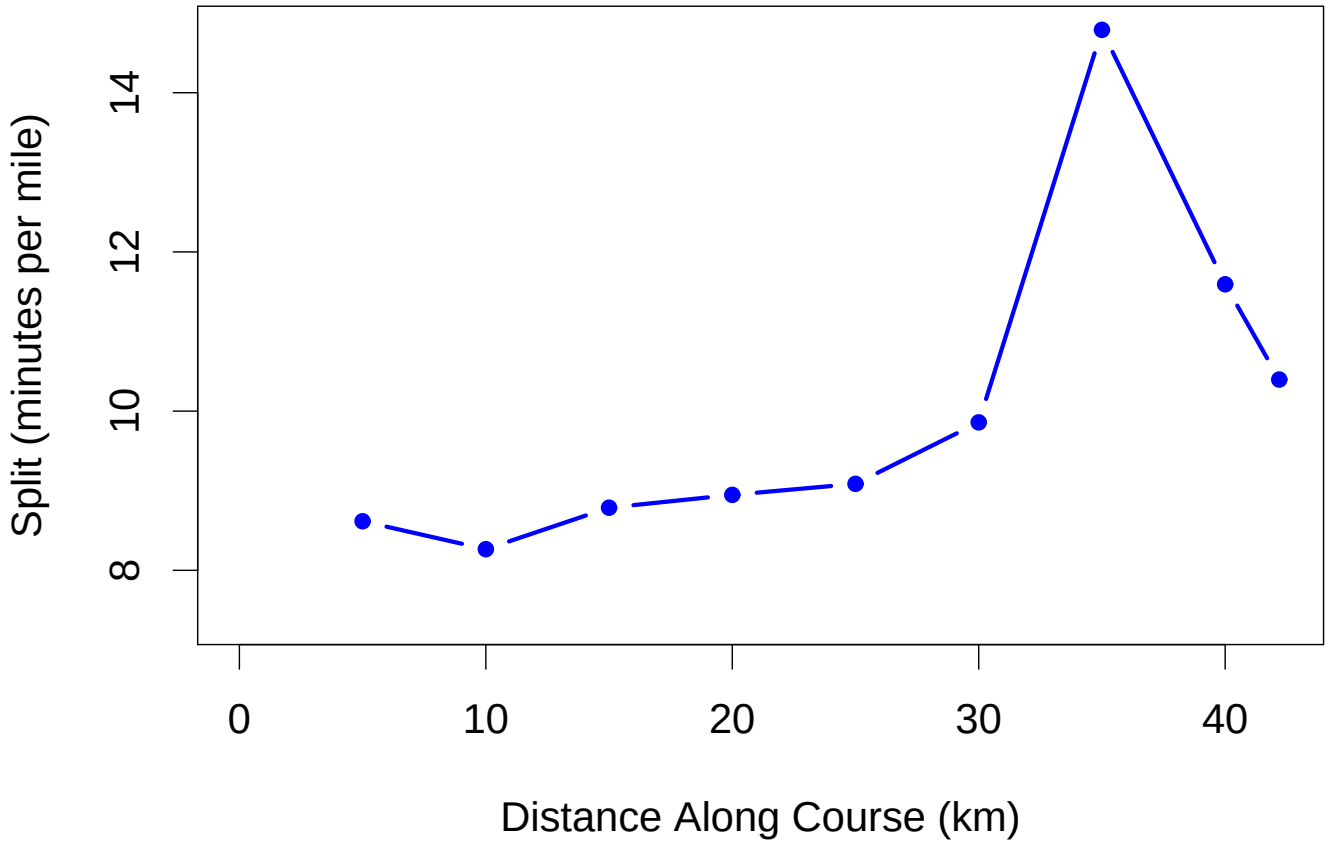


Matt Cefalu

- About 5,700 runners did not finish (DNF), most of them because of the bombs
- Data: Full data on 2013 race, plus 2011 and 2010
- "Split times" every 5 km.
- Missing values in the middle of someone's split times (before they stopped) were interpolated
- Among the DNFs,
  - 80% had complete times up to 40 km.,
  - 9.5% stopped between 35 km. and 40 km.,
  - 8.2% stopped between 30 km. and 35 km.,
  - 0.7% stopped between 25 km. and 30 km.,
  - 1.7% stopped between 20 km. and 25 km. Didn't consider earlier dropouts.

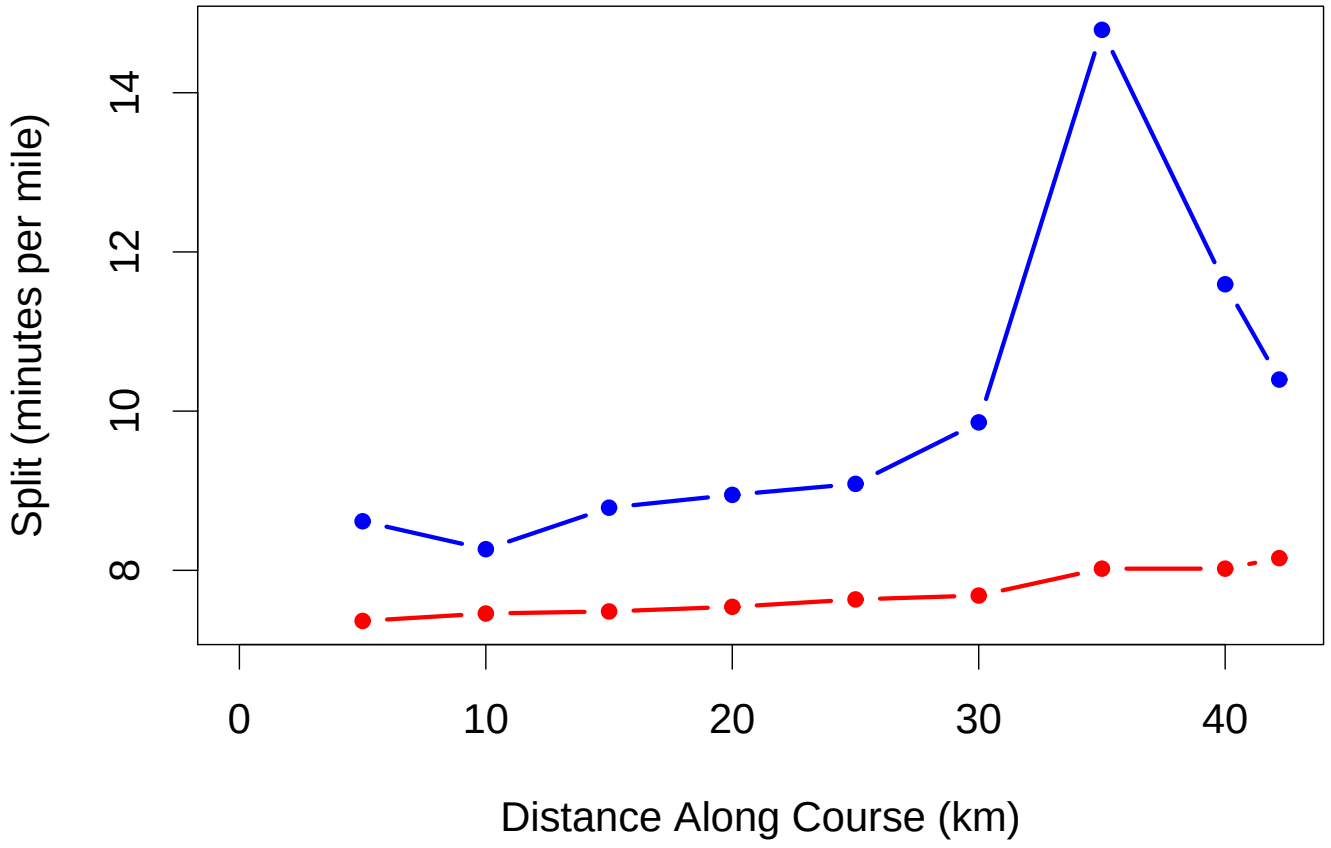
Objective: Project the finish times

Richard, 2011



Richard, 2011

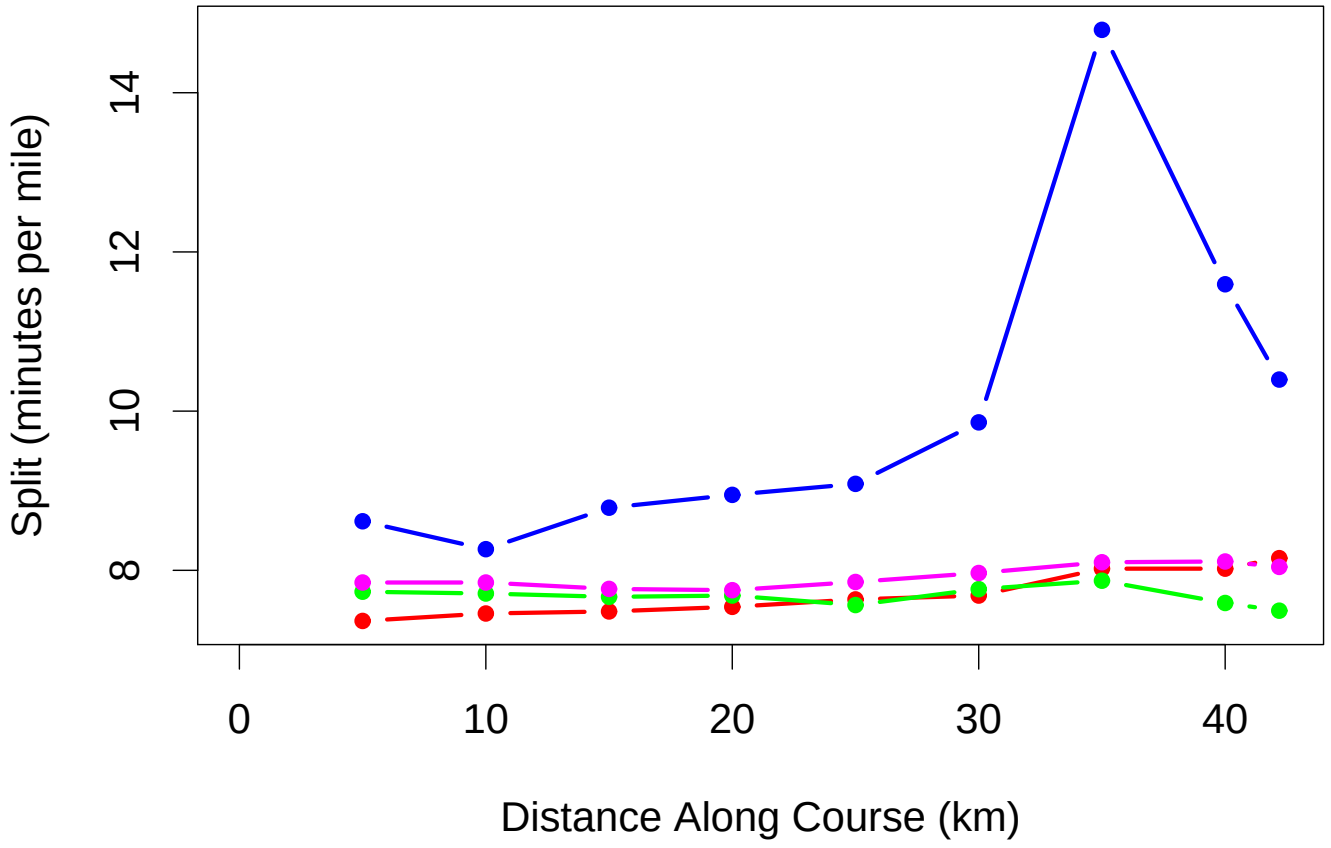
Francesca, 2013

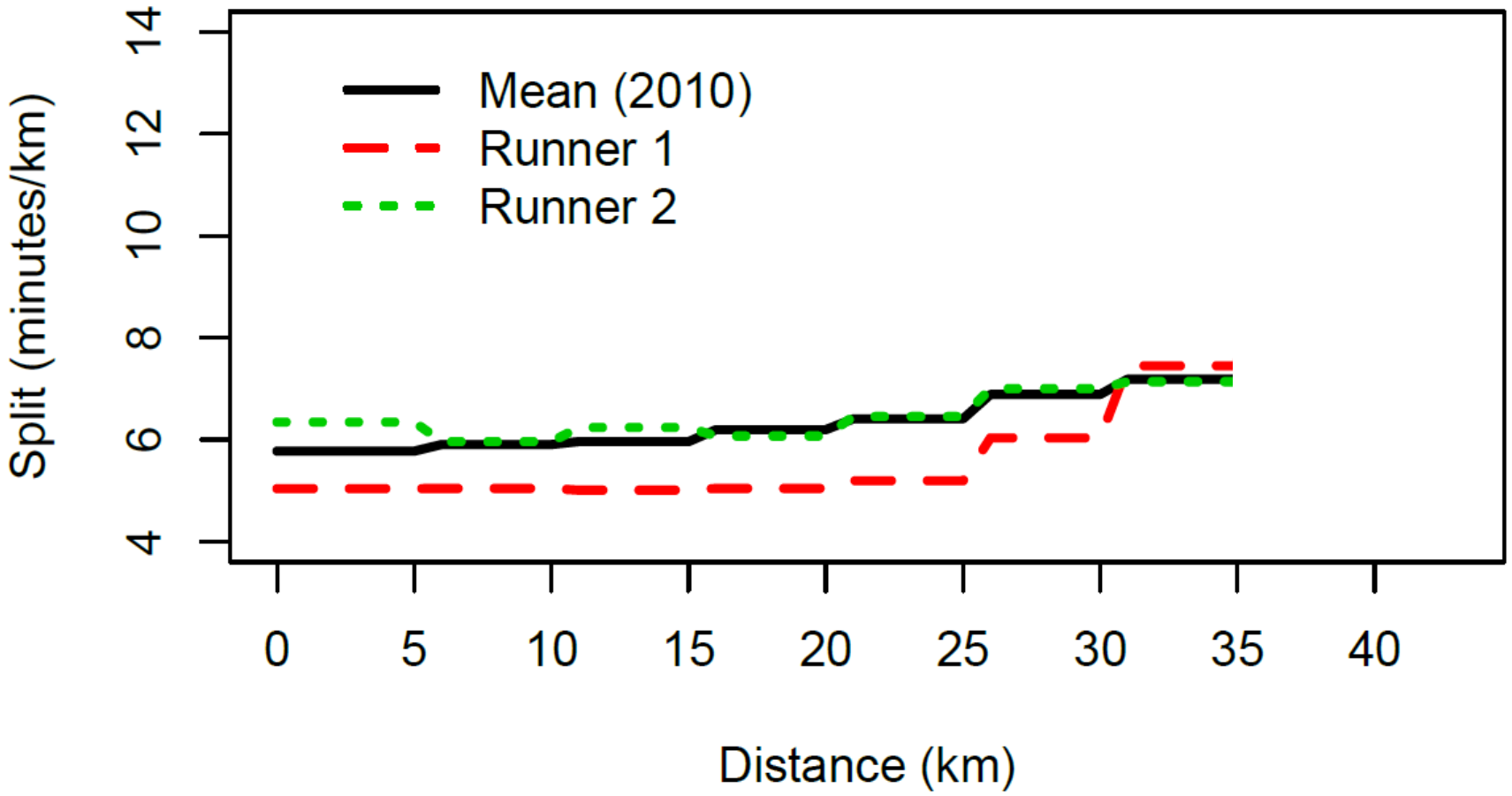


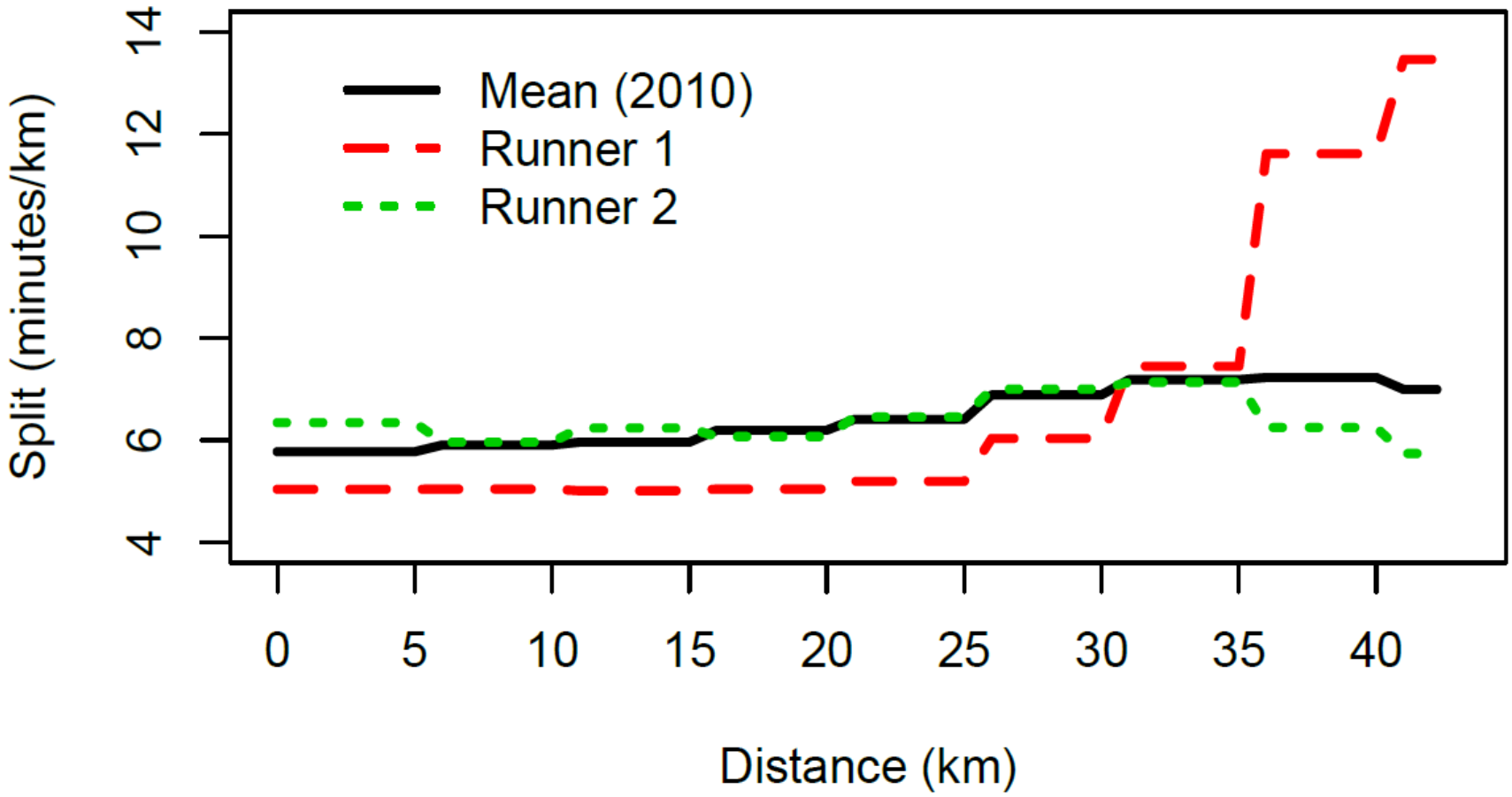
Richard, 2011

Francesca, 2013

Giovanni Dorit




Split (minutes/km)
4
6
8
10
12
14
Mean (2010)
Runner 1
Runner 2
0
5
10
15
20
25
30
35
40
Distance (km)


Split (minutes/km)
4
6
8
10
12
14
Mean (2010)
Runner 1
Runner 2
0
5
10
15
20
25
30
35
40
Distance (km)

Scientific Context

- Big Data
- Scalable Algorithms
- Reproducible Research
  - `http://www.unc.edu/~rls/boston.html`
- The Matrix Completion Problem
  - Netflix Prize Competition
    - 480,000 subscribers
    - 18,000 movies
    - $\approx$100,000,000 ratings (1.2% of all possible)
  - DNA Microarrays
  - Handicapping a race

Runners' Times in 8 Races

Name	Eno Equalizer	Geezer Pleezer	Hard Climb Hill 3M	Hard Climb Hill 7M	Hard Climb Hill 10M	Misery Run	Couch Mountain	New Year's Day
Robert Agans	33:05							
Charles Alden								53:15
Halle Amick						66:17	49:52	49:18
Lisa Anderson							42:12	
Maria Archibald							49:13	45:37
Owen Astrachan	30:05	26:11				48:22		
Jordan Baker						48:55		
Brent Baker						53:53	42:15	44:29
Bart Bechard	27:52				69:15	41:50		33:46
Karen Bell		36:51						
.....								

Runners' Times in 8 Races

Name	Eno Equalizer	Geezer Pleezer	Hard Climb Hill 3M	Hard Climb Hill 7M	Hard Climb Hill 10M	Misery Run	Couch Mountain	New Year's Day
Robert Agans	33:05 33:04	29:41	25:07	58:34	84:50	53:01	39:24	39:47
Charles Alden	44:16	39:43	33:37	78:23	113:31	70:57	52:43	53:15 53:14
Halle Amick	41:24	37:09	31:26	73:18	106:10	66:17 66:21	49:52 49:18	49:18 49:47
Lisa Anderson	35:25	31:47	26:54	62:44	90:51	56:47	42:12 42:11	42:37
Maria Archibald	39:35	35:31	30:04	70:06	101:31	63:27	49:13 47:08	45:37 47:37
Owen Astrachan	30:05 29:48	26:11 26:44	22:38	52:47	76:26	48:22 47:46	35:30	35:51
Jordan Baker	30:30	27:23	23:10	54:02	78:15	48:55 48:55	36:20	36:42
Brent Baker	35:19	31:42	26:50	62:33	90:36	53:53 56:37	42:15 42:04	44:29 42:29
Bart Bechard	27:52 27:14	24:27	20:41	48:15	69:15 69:52	41:50 43:40	32:27	33:46 32:46
Karen Bell	41:03	36:51 36:50	31:11	72:42	105:18	65:49	48:54	49:23
.....								

Solutions to the Boston Marathon Problem

- Linear Regression
- ANOVA Method
- SVD Method
- KNN Method
- Split Ratio Method

Linear Regression

$$y_i = \sum_{j=1}^{J} x_{ij}\beta_j + \epsilon_i,$$

- y_i is sum of missing split times for runner i,
- J is number of available split times,
- x_{ij} is available split time for section j for runner i,
- β_j is coefficient corresponding to split time x_j,
- ϵ_i mean 0, uncorrelated, common variance.

Doesn't allow for different subpopulations

Interesting finding: β_1 and β_2 were negative (motto: it pays to start slow!)

ANOVA Method

Let y_{ij} be log split time for runner i on section j of the course.

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

with $\sum_i \alpha_i = \sum_j \beta_j = 0$.

Fit this model by standard OLS, estimate

$$\widehat{y}_{ij} = \widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j$$

to predict times on missing segments.

- Direct application was too slow on this dataset (algorithm not scalable)
- Also doesn't allow for different subpopulations
- Solution was to divide runners into subgroups according to half marathon time and relative splits, but this still didn't work very well.

SVD Method

Motivated by Troyanskaya *et al.* (2001) method for DNA arrays

Suppose y_{ij} is split time of runner i on section j. Write

$$y_{ij} = \sum_{d=1}^{D} \alpha_{id} \cdot s_d \cdot \beta_{jd}.$$

- $D = 1$ similar to ANOVA method
- First find optimal $D = 1$ fit, calculate residuals
- Repeat algorithm on residuals, gives optimal fit for $D = 2$
- Continue to $D = 9$, chosen as best fit by cross-validation
- First apply to complete part of data matrix, then repeat to estimate missing values

K Nearest Neighbors (KNN) Method

For a runner who has completed the first J sections of the course:

- Find K nearest neighbors based on full J-dimensional vector of split times
- Repeat linear regression step but restricted to the K nearest neighbors
- Use fitted regression to predict remaining split times.

Use kd-tree algorithm to find nearest neighbors (implementation in both Matlab and R)

$K = 200$ suggested by trial and error, but $K = 100$ or $K = 300$ very similar results

Split Ratio Method

Find multiplicative constants relating each 5 km segment time to the previous 5 km segment time. Separate by gender.

Example: suppose Mary's last observed split time was 30 minutes for the 30-35 km segment, with an overall split of 3:25:00. Predict missing time for last two splits by multiplying 30 minutes by corresponding constant (1.4055). Add to existing split to get predicted time of 4:07:10.

Multiplicative constants:

	Last 5 km segment completed				
Gender	15–20	20–25	25–30	30–35	35–40
Male	5.0648	3.8765	2.5761	1.4333	0.4207
Female	4.8354	3.6965	2.4876	1.4055	0.4230
Constant Pace	4.439	3.439	2.439	1.439	0.439

Other Methods Considered

- Regularized SVD Method (Mazumder/Hastie/Tibshirani 2009)
- Two-stage hierarchical models
  - First stage: model individual outcomes in terms of latent variables
  - Second stage: prior distribution for the latent variables
  - Most models used in biostatistics employ multivariate normal priors for the second stage. Maybe that's not appropriate in this case.
- Raymond Britt's rule: (a) for runners who reached 40 km, multiply 40 km split time by 1.06; (b) for runners who reached 35 km but not 40 km, multiply 35 km split time by 1.23. No solution given for runners who failed to reach 35 km.
- Constant pace rule: assume the runner maintained the same overall pace to the finish

Evaluating the Methods

- Create a validation dataset — omit the 2013 DNFs; randomly assign runners from $>$ 4-hour finishers in 2010 and 2011 as DNFs in same proportions as original dataset
- Apply all the prediction rules to the validation dataset
- Compare performance by mean absolute error (MAE), mean squared error (MSE) and percent correct within various bounds

Results by last recorded split (20km, 25km, 30km)

		mae	mse	1min	2min	3min	4min	5min	10min
20 km ($n = 62$)	ANOVA	14.66	401.10	0.016	0.129	0.161	0.226	0.274	0.435
	SVD	9.74	161.35	0.048	0.097	0.194	0.258	0.290	0.613
	Splitratio	8.41	144.24	0.065	0.097	0.258	0.403	0.435	0.742
	LM	7.89	124.41	0.065	0.258	0.339	0.387	0.435	0.790
	KNN	8.95	198.78	0.081	0.161	0.242	0.290	0.355	0.758
25 km (n=29)	ANOVA	9.90	173.13	0.069	0.172	0.172	0.207	0.310	0.655
	SVD	8.33	121.20	0.103	0.138	0.276	0.276	0.379	0.655
	Splitratio	7.41	107.66	0.034	0.138	0.310	0.517	0.552	0.793
	LM	6.84	97.14	0.172	0.276	0.310	0.310	0.414	0.862
	KNN	7.60	108.27	0.138	0.172	0.310	0.414	0.483	0.724
30 km ($n = 314$)	ANOVA	5.60	78.76	0.143	0.296	0.411	0.538	0.631	0.857
	SVD	5.77	82.22	0.131	0.255	0.395	0.513	0.599	0.866
	Splitratio	5.60	81.32	0.162	0.309	0.436	0.529	0.627	0.860
	LM	5.37	66.98	0.140	0.258	0.401	0.519	0.608	0.873
	KNN	4.58	54.11	0.191	0.331	0.494	0.599	0.707	0.901

Results by last recorded split (35km, 40km)

		mae	mse	1min	2min	3min	4min	5min	10min
35 km ($n = 435$)	ANOVA	3.40	25.47	0.244	0.451	0.602	0.713	0.809	0.954
	SVD	3.27	25.15	0.262	0.453	0.634	0.747	0.830	0.954
	Splitratio	3.26	28.70	0.306	0.494	0.660	0.747	0.809	0.945
	LM	3.11	22.85	0.287	0.501	0.657	0.754	0.834	0.949
	KNN	2.76	17.60	0.294	0.529	0.687	0.802	0.874	0.959
	Britt	4.19	34.89	0.189	0.347	0.497	0.609	0.710	0.920
40 km ($n = 3314$)	ANOVA	1.08	2.80	0.616	0.875	0.947	0.973	0.985	0.997
	SVD	0.96	2.75	0.675	0.904	0.959	0.976	0.986	0.998
	Splitratio	1.01	3.61	0.675	0.893	0.952	0.972	0.982	0.997
	LM	0.94	2.59	0.687	0.910	0.964	0.980	0.988	0.998
	KNN	0.94	2.32	0.697	0.899	0.957	0.977	0.987	0.998
	Britt	1.28	3.55	0.524	0.825	0.929	0.969	0.981	0.996

Other Issues Identified

- Younger Runners (age ≤ 45) predicted more accurately than older runners
- Women predicted more accurately than men
- Faster Runners (finish time up to 4 h. 25 m.) predicted more accurately than slower runners

Runners Classified By Time Differential Between 2013 Projected Time and 2014 Qualifying Time - Male

Time Differential - Age Category	18-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-99
120 or more	112	67	54	38	32	26	13	5	2	0	0
60 to 119	351	185	213	183	180	119	78	37	18	3	2
20 to 59	2	23	53	95	103	119	143	95	31	11	4
10 to 19	0	0	0	0	0	0	33	38	13	2	0
5 to 9	0	0	0	0	0	0	2	23	13	3	0
3 or 4	0	0	0	0	0	0	0	3	3	1	0
exactly 2	0	0	0	0	0	0	0	3	3	0	0
exactly 1	0	0	0	0	0	0	0	1	1	1	0
exactly 0	0	0	0	0	0	0	0	3	1	0	0
exactly -1	0	0	0	0	0	0	0	0	4	0	0
exactly -2	0	0	0	0	0	0	0	0	5	0	0
-3 or -4	0	0	0	0	0	0	0	0	2	0	0
-5 to -9	0	0	0	0	0	0	0	2	6	0	0
-10 to -19	0	0	0	0	0	0	0	0	7	4	1
-20 or better	0	0	0	0	0	0	0	0	0	7	0

Runners Classified By Time Differential Between 2013 Projected Time and 2014 Qualifying Time - Female

Time Differential - Age Category	18-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-99
120 or more	100	29	31	13	10	2	1	0	0	0	0
60 to 119	502	153	150	101	79	45	20	5	0	0	0
20 to 59	319	126	207	223	164	92	48	18	4	1	1
10 to 19	0	0	3	66	98	66	16	5	1	0	0
5 to 9	0	0	0	4	17	29	14	7	1	0	1
3 or 4	0	0	0	0	1	15	3	3	1	0	0
exactly 2	0	0	0	0	0	5	6	1	0	0	0
exactly 1	0	0	0	0	0	16	3	0	0	0	0
exactly 0	0	0	0	0	0	8	2	1	0	0	0
exactly -1	0	0	0	0	0	5	3	2	0	0	0
exactly -2	0	0	0	0	0	1	6	0	0	0	0
-3 or -4	0	0	0	0	0	2	8	3	0	0	1
-5 to -9	0	0	0	0	0	0	17	2	0	0	0
-10 to -19	0	0	0	0	0	0	19	10	5	1	0
-20 or better	0	0	0	0	0	0	0	12	7	1	0

Our Report to the BAA

- We recommended a set of finish times to the Boston marathon, based on the KNN rule.

- Under our recommendations, 158 of the DNF runners would have achieved qualifying times for the 2014 race (compared with 180 under a "constant pace" projection — note that we already excluded runners who would have finished before the bombs under our projections)

Timeline

- April 15: Date of race
- April 23: Email from BAA
- April 30: BAA send all data files we requested
- May 13: We sent our report to them
- May 16: BAA announces that all runners who completed the first half of the race but did not finish will receive automatic entries to the 2014 race
- June 3: BAA posts projected times on website ...
  - using "constant pace" rule

Comments on the BAA Decision

From a number of points of view, it is fully understandable

- Everyone can understand what they did
- No need to defend the decision to bring in a group of statisticians
- Everyone who didn't finish gets to run in 2014 anyway, so why make a big deal of the projected times?
- In *most*, but not all cases, the BAA projected a lower finish time (more favorable to the runners) than we did

But I'm still going to try to convince you our solution was better!

Results by last recorded split (20km, 25km, 30km)

		mae	mse	1min	2min	3min	4min	5min	10min
20 km ($n = 62$)	ANOVA	14.66	401.10	0.016	0.129	0.161	0.226	0.274	0.435
	SVD	9.74	161.35	0.048	0.097	0.194	0.258	0.290	0.613
	Splitratio	8.41	144.24	0.065	0.097	0.258	0.403	0.435	0.742
	LM	7.89	124.41	0.065	0.258	0.339	0.387	0.435	0.790
	KNN	8.95	198.78	0.081	0.161	0.242	0.290	0.355	0.758
	ConstPace	20.83	652.38	0.000	0.000	0.032	0.048	0.097	0.226
25 km (n=29)	ANOVA	9.90	173.13	0.069	0.172	0.172	0.207	0.310	0.655
	SVD	8.33	121.20	0.103	0.138	0.276	0.276	0.379	0.655
	Splitratio	7.41	107.66	0.034	0.138	0.310	0.517	0.552	0.793
	LM	6.84	97.14	0.172	0.276	0.310	0.310	0.414	0.862
	KNN	7.60	108.27	0.138	0.172	0.310	0.414	0.483	0.724
	ConstPace	18.54	473.51	0.000	0.000	0.000	0.034	0.069	0.172
30 km ($n = 314$)	ANOVA	5.60	78.76	0.143	0.296	0.411	0.538	0.631	0.857
	SVD	5.77	82.22	0.131	0.255	0.395	0.513	0.599	0.866
	Splitratio	5.60	81.32	0.162	0.309	0.436	0.529	0.627	0.860
	LM	5.37	66.98	0.140	0.258	0.401	0.519	0.608	0.873
	KNN	4.58	54.11	0.191	0.331	0.494	0.599	0.707	0.901
	ConstPace	12.20	248.16	0.035	0.076	0.140	0.172	0.226	0.490

Results by last recorded split (35km, 40km)

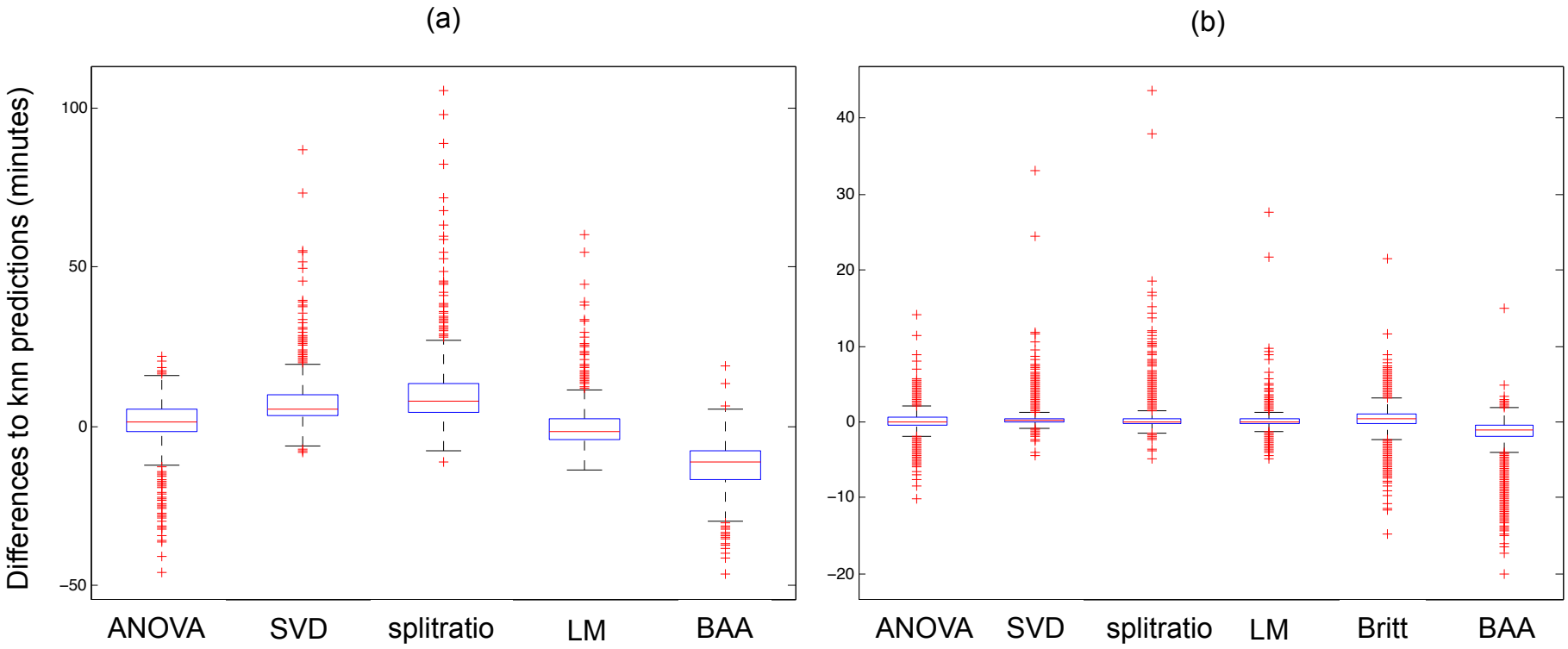
		mae	mse	1min	2min	3min	4min	5min	10min
35 km ($n = 435$)	ANOVA	3.40	25.47	0.244	0.451	0.602	0.713	0.809	0.954
	SVD	3.27	25.15	0.262	0.453	0.634	0.747	0.830	0.954
	Splitratio	3.26	28.70	0.306	0.494	0.660	0.747	0.809	0.945
	LM	3.11	22.85	0.287	0.501	0.657	0.754	0.834	0.949
	KNN	2.76	17.60	0.294	0.529	0.687	0.802	0.874	0.959
	Britt	4.19	34.89	0.189	0.347	0.497	0.609	0.710	0.920
	ConstPace	6.46	69.59	0.099	0.172	0.264	0.366	0.467	0.811
40 km ($n = 3314$)	ANOVA	1.08	2.80	0.616	0.875	0.947	0.973	0.985	0.997
	SVD	0.96	2.75	0.675	0.904	0.959	0.976	0.986	0.998
	Splitratio	1.01	3.61	0.675	0.893	0.952	0.972	0.982	0.997
	LM	0.94	2.59	0.687	0.910	0.964	0.980	0.988	0.998
	KNN	0.94	2.32	0.697	0.899	0.957	0.977	0.987	0.998
	Britt	1.28	3.55	0.524	0.825	0.929	0.969	0.981	0.996
	ConstPace	1.52	5.00	0.465	0.754	0.884	0.937	0.968	0.995

One (rather extreme) individual case

BibNum	Age	M/F	FTANOVA	FTSVD	FTSPLITRATIO
2208	25	M	4:05:10	4:10:39	4:20:01

FTLM	FTKNN	FTBRITT	Constant pace
4:17:44	4:17:35	NA	3:36:08

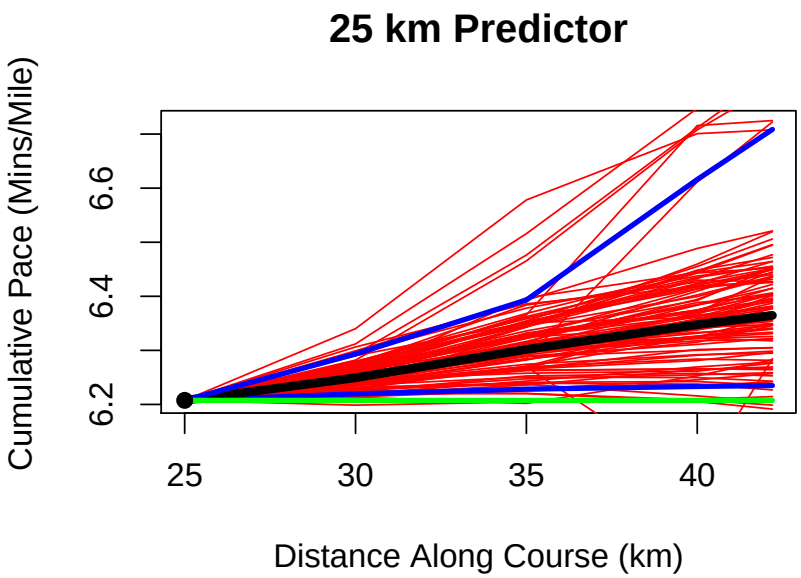
- Qualifying performance: probably about 2 hrs 55 min.
- Half-marathon time: 1:29:41
- 5 km split times for first 30 km: 20:57, 20:52, 20:49, 21:09, 28:33, 41:19
- Quit at 30 km

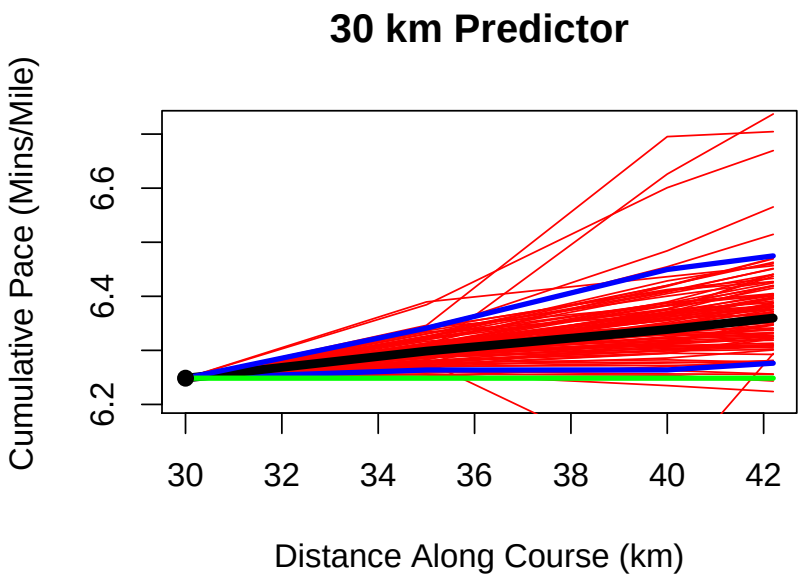


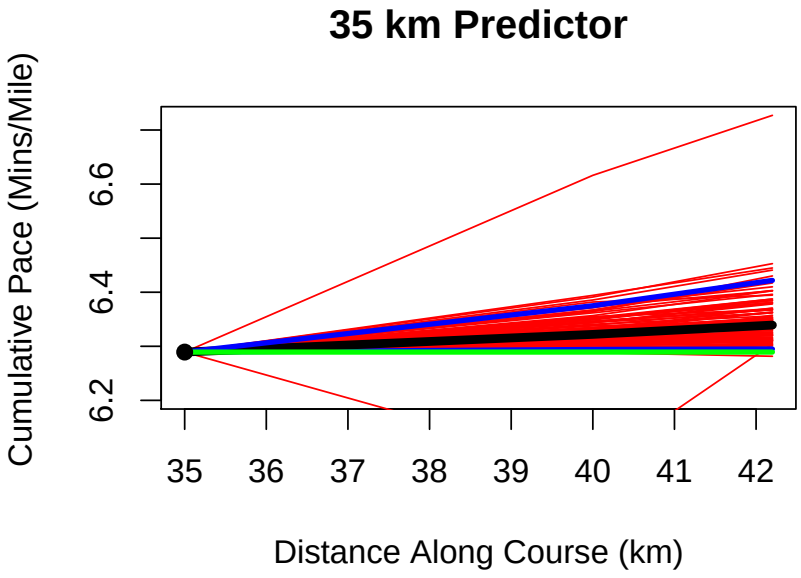
Boxplots of differences in predicted finishing times between the KNN method and other methods for participants in the 2013 Boston marathon, who passed the half-marathon mark, but did not complete the course. Predictions are based on (a) splits available up to 30k or less (n=515) and (b) splits available up to 35k or 40k (n=5009). Note the scale difference between the two plots.

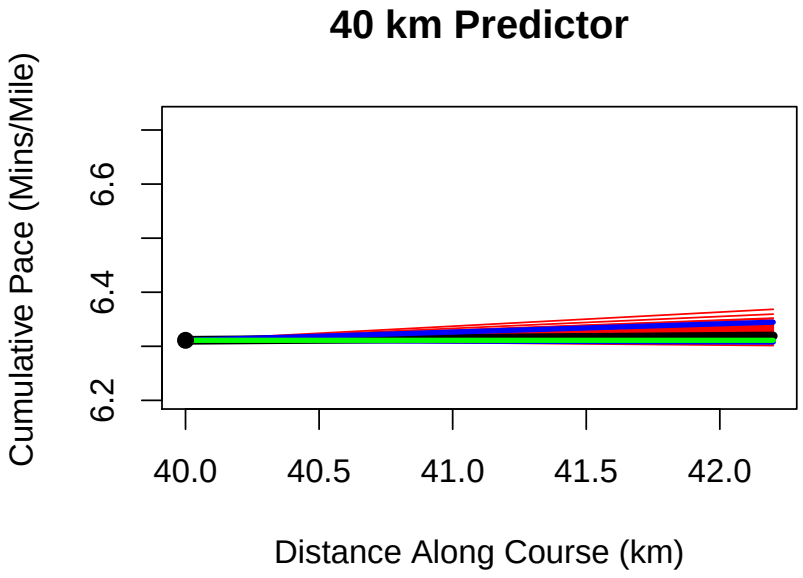
Looking Forwards: Predicting Finish Times from Intermediate Split Times

- Many races (including the Boston marathon) feature "athelete tracker" apps
- You can go online during the race or sign up to receive updates by email or text message, to receive updates on a runner's pace and projected finish times
- To the best of my knowledge, *all* such systems use the "constant pace method" to project finish times
- We illustrate a possible improvement, using "rescaled KNN"
  - Find K nearest neighbors, as before
  - Instead of performing a linear regression, simply "rescale" each of the neighbor runners to have the same split time as the target runner
  - We can construct a prediction interval as well as a point prediction by this method

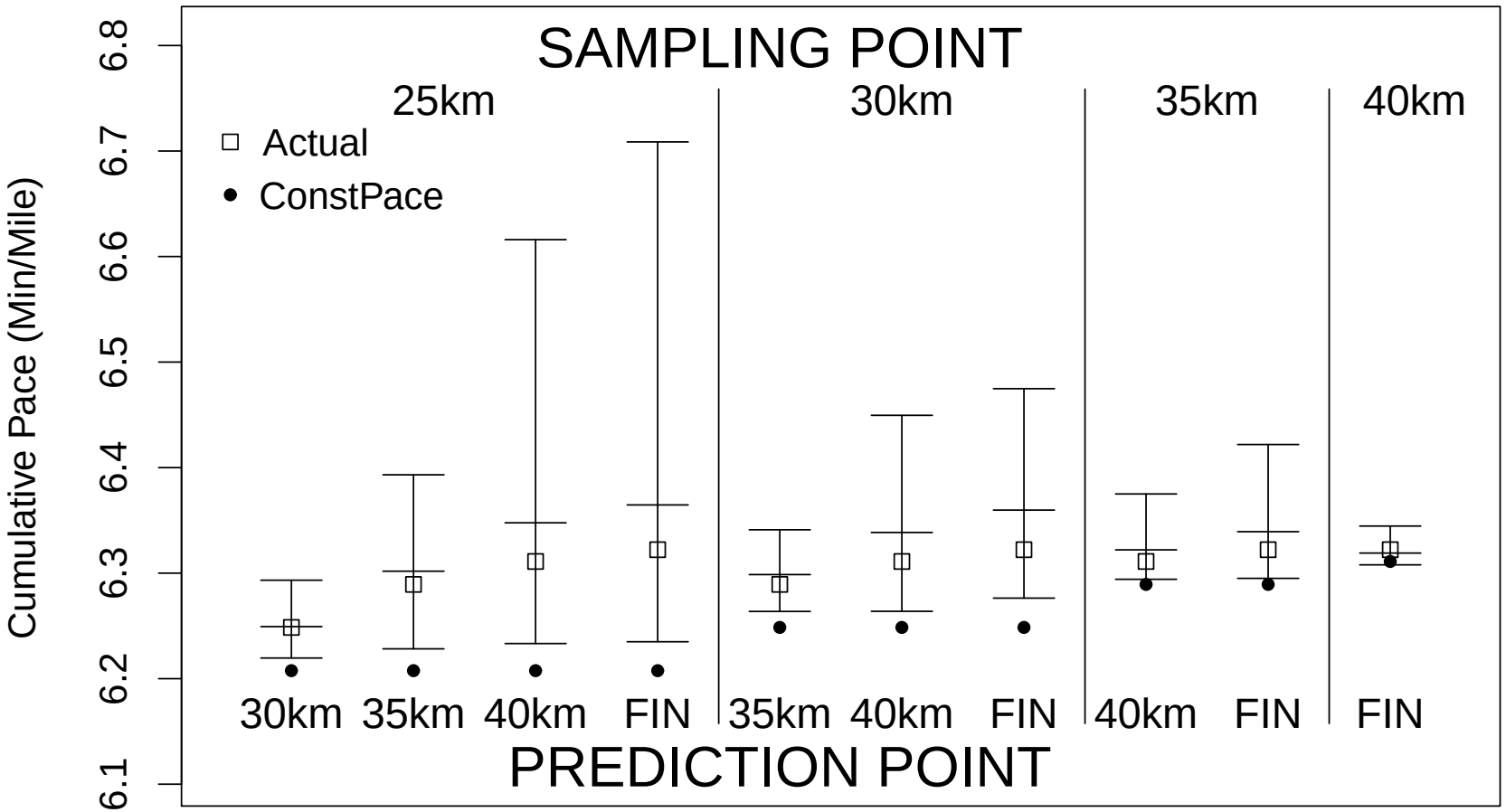

25 km Predictor
Cumulative Pace (Mins/Mile)
6.2
6.4
6.6
25
30
35
40
Distance Along Course (km)


30 km Predictor
Cumulative Pace (Mins/Mile)
6.2
6.4
6.6
30
32
34
36
38
40
42
Distance Along Course (km)

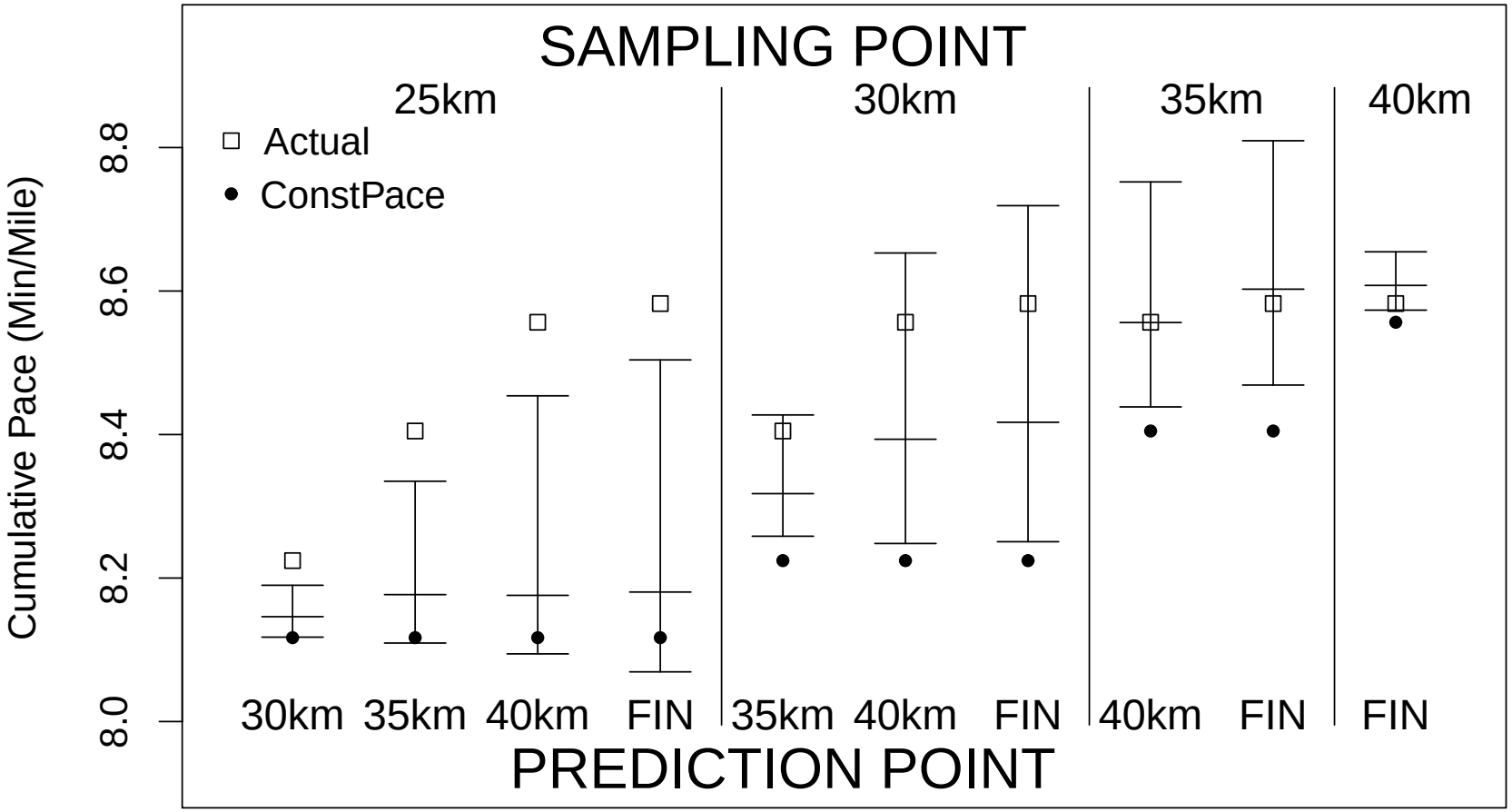

35 km Predictor
Cumulative Pace (Mins/Mile)
6.2
6.4
6.6
35
36
37
38
39
40
41
42
Distance Along Course (km)


40 km Predictor
Cumulative Pace (Mins/Mile)
6.2
6.4
6.6
40.0
40.5
41.0
41.5
42.0
Distance Along Course (km)

2:45 MARATHONER



3:45 MARATHONER



Summary and Conclusions

- Five (and more) prediction methods — calibrated on 2010/2011 data, used a validation dataset to determine which was best
- KNN method worked best — also considered "rescaled KNN method"
- Many other methods possible based on modern "big data" ideas
- The future?
  - Real-time prediction of finish times based on intermediate time points
  - Better understanding of the science of pacing
  - Could it help to detect cheating?