Data Analytics to Improve Running Form and Reduce Risk Factors in Middle to Long Distance Runners

by

Euodia Vermeulen

A dissertation submitted in fulfillment of the requirements for the degree of

Masters of Engineering (Industrial Engineering)

in the

Faculty of Engineering, Built Environment and Information Technology

University of Pretoria, Pretoria

September 2018

Executive Summary

Title	Data analytics to improve running form and reduce risk factors in middle to long distance runners
Author	Euodia Vermeulen
Study Leader	Professor V.S.S. Yadavalli
Department	Industrial and Systems Engineering
University	University of Pretoria
Degree	Masters in Engineering (Industrial Engineering)

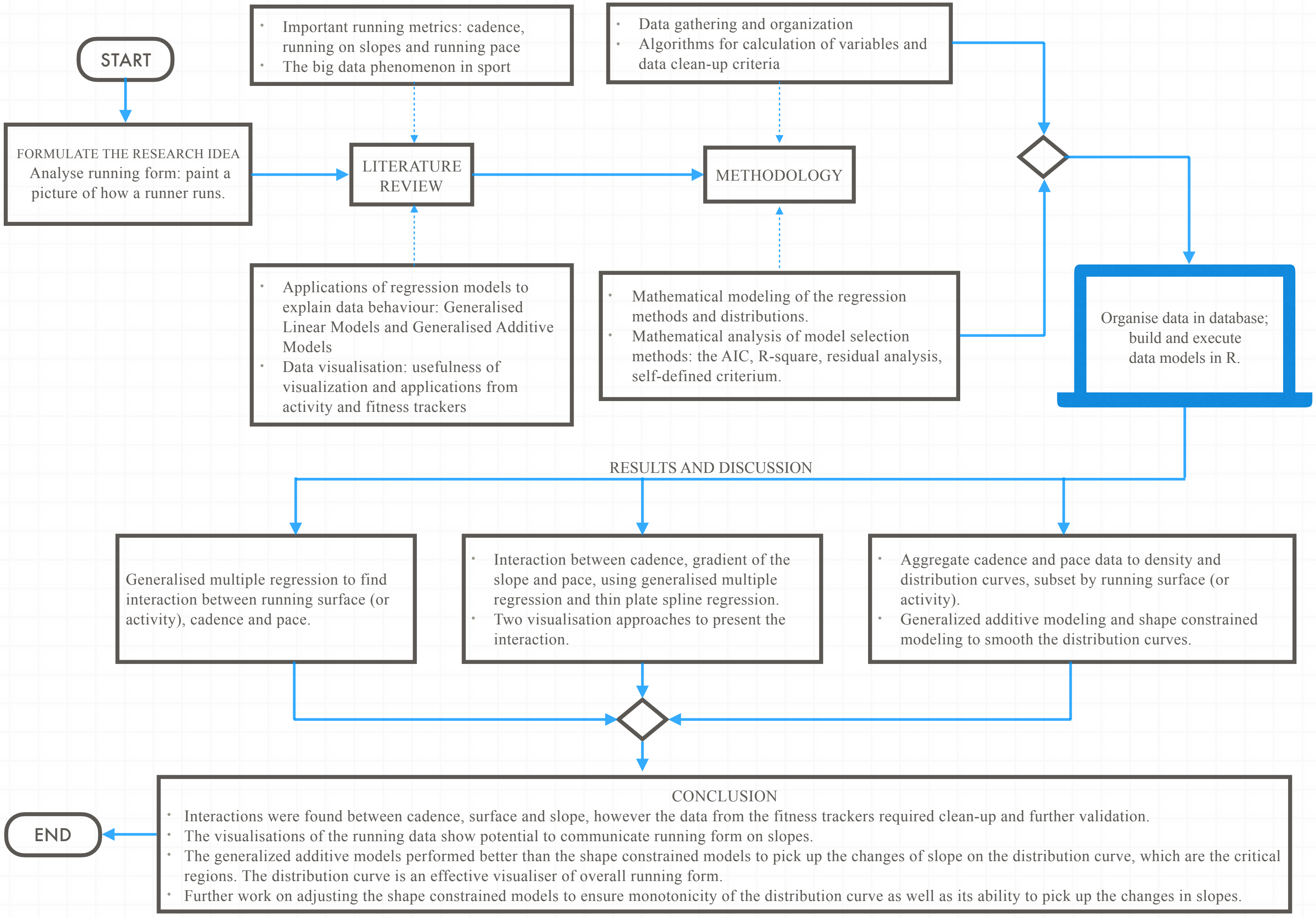
Fitness trackers equipped with accelerometers and global positioning systems are becoming more popular among the running community. These devices allow runners across the spectrum of athletic abilities to monitor their running metrics and track their performance throughout their chosen routes. The size of the data sets and the frequency at which it is generated place the tracking data from these devices into realm of big data. There are calls from research fields focused on human locomotion during running to capitalise on the data from fitness trackers, in order to evaluate athletes in the real world and outside of the sometimes unrealistic laboratory or clinical settings. Unfortunately, the real world adds noise to the data and the signal from the data becomes obscured. This dissertation explored the large tracking data sets from runners' running watches to evaluate the extent of the noise and the possibilities to extract the signal from the data. Data are cleaned and parametric as well as non-parametric regression analysis models are fitted to the data to find interactions and aggregation methods that present the athlete with a picture of his/her running form. These models may provide an athlete with a better understanding of their own capabilities, which will help them improve their running form and reduce risk factors attributed to poor form.

Results from the interaction models between running surface, cadence and pace suggest that the running surface do have an effect on cadence and running pace. However, the distribution of pace per cadence level is extensive and skew in either direction, with the R_a^2-values for the fitted models ranging between a weak 0.155 and moderate strength of 0.752 for the four case studies. The spread for road gradients (i.e. slopes) per cadence level is large and also skew in either direction. The R_a^2-values for the interaction models for slope, cadence and pace range between 0.268 and 0.681. The data visualisations for graded running is able to show the pattern of the data to a limited extent. The aggregated distribution curves for cadence and pace serve as an extension on the interaction between running surface, cadence and pace. Although all the distribution curve models had a R^2-value very close to 1, the generalised additive model outperformed the shape constrained model with lower AIC-scores to fit a smoothed line that represents the overall performance of the athlete. The shape constrained models failed to pick up segmented improvements in the running metrics, where the generalised additive models did pick up the changes in the slope of the curves where the athlete's performance improved.

The data from fitness trackers seem to hold potential to extend sport science research in running, however the data may not always be a true representation of reality. This may be due to its varying veracity and slow algorithm responses to changes in performance.

Keywords: fitness trackers, data mining, regression analysis, data visualisation, sport analytics

Data analytics to improve running form and reduce risk factors in middle to long distance runners: A pictorial presentation of the thesis



START

FORMULATE THE RESEARCH IDEA
Analyse running form: paint a picture of how a runner runs.

- Important running metrics: cadence, running on slopes and running pace
- The big data phenomenon in sport

LITERATURE REVIEW

- Applications of regression models to explain data behaviour: Generalised Linear Models and Generalised Additive Models
- Data visualisation: usefulness of visualization and applications from activity and fitness trackers

- Data gathering and organization
- Algorithms for calculation of variables and data clean-up criteria

METHODOLOGY

- Mathematical modeling of the regression methods and distributions.
- Mathematical analysis of model selection methods: the AIC, R-square, residual analysis, self-defined criterium.

Organise data in database; build and execute data models in R.

RESULTS AND DISCUSSION

Generalised multiple regression to find interaction between running surface (or activity), cadence and pace.

- Interaction between cadence, gradient of the slope and pace, using generalised multiple regression and thin plate spline regression.
- Two visualisation approaches to present the interaction.

- Aggregate cadence and pace data to density and distribution curves, subset by running surface (or activity).
- Generalized additive modeling and shape constrained modeling to smooth the distribution curves.

CONCLUSION

- Interactions were found between cadence, surface and slope, however the data from the fitness trackers required clean-up and further validation.
- The visualisations of the running data show potential to communicate running form on slopes.
- The generalized additive models performed better than the shape constrained models to pick up the changes of slope on the distribution curve, which are the critical regions. The distribution curve is an effective visualiser of overall running form.
- Further work on adjusting the shape constrained models to ensure monotonicity of the distribution curve as well as its ability to pick up the changes in slopes.

END

“We can make our plans, but the Lord determines our steps.”

Proverbs 16:9

“Run when you can, walk if you have to, crawl if you must; just never give up.”

Dean Karnazes

Acknowledgments

My sincere gratitude towards the following role players:

- God my heavenly Father who, just as He had promised, guided my steps throughout this process.
- Professor Yadavalli for your dedicated guidance throughout the completion of the thesis. Your quick response rate and enthusiasm are much appreciated.
- My mother and father for your continued support through my extensive educational journey. I am your professional-student child.
- My office mates in the Department of Industrial and Systems Engineering at the University of Pretoria for some insightful comments made on my work.
- Mr. Ronald Lagerwall, for coining the term “lifestyle athlete” and introducing me to the participating athletes.
- The athletes who willingly and wholeheartedly participated in the study and allowed me access to their tracking data.

Contents

1 Introduction **1**
1.1 Winning with Numbers 1
1.2 Problem Background 2
1.3 Problem Statement 3
1.4 Research Design 4
1.5 Research Method 4
1.6 Structure of the document 6

2 Literature review **7**
2.1 The conversion of terabytes of numbers into wisdom 7
2.2 The importance of remaining injury-free 10
2.3 Metrics for running form 12
  2.3.1 Cadence and pace 12
  2.3.2 Graded running 13
2.4 Applications of data mining in sport 14
2.5 Flaw of Averages 19
2.6 Regression methods 22
2.7 Data visualisation – painting by numbers 29
2.8 Ethics in big data and sport 32
  2.8.1 Data validity 32
  2.8.2 Data security and protection 33
  2.8.3 Athlete autonomy 33

3 Methodology: From time-series analysis to interaction and distribution modeling **35**
3.1 Data gathering 35
3.2 Data organisation 38
3.3 Variable calculations 39
3.4 Descriptive analysis 40

3.5 Data clean-up . . . 43
3.6 The transformation: time-series to distribution curve . . . 45
3.7 Regression modeling . . . 45
    3.7.1 Regression mathematics: parametric models . . . 47
    3.7.2 Interaction models . . . 52
    3.7.3 Regression mathematics: non-parametric models . . . 53
    3.7.4 Distributions considered for regression models . . . 71
    3.7.5 Associations between cadence, grade and pace . . . 75
    3.7.6 Application: the interaction between cadence and running activity . 75
    3.7.7 Application: the interaction between cadence and grade . . . 76
    3.7.8 Application: smoothing of the distribution curves . . . 77
3.8 Model selection . . . 81
    3.8.1 Model adequacy . . . 82
    3.8.2 Application: fitted negative values . . . 90
    3.8.3 Behaviour near $y = 0$ . . . 90
    3.8.4 Monotonicity of the distribution curves . . . 91

4 The interaction between cadence and running activity as explanatory variables for running pace 93
4.1 Case study A: the semi-professional all-rounder . . . 94
4.2 Case study B: the trail specialist . . . 99
4.3 Case study C: the Comrades marathoner . . . 103
4.4 Case study D: the heart-rate runner . . . 108
4.5 Conclusion . . . 113

5 Harnessing the big data from fitness trackers to visually analyse overground running 117
5.1 Case study A: the semi-professional all-rounder . . . 118
5.2 Case study B: the trail specialist . . . 124
5.3 Case study C: the Comrades marathoner . . . 130
5.4 Case study D: the heart-rate runner . . . 134
5.5 Conclusion . . . 139

6 The distribution curves to monitor running form 142
6.1 Case study A: the semi-professional all-rounder . . . 143
    6.1.1 Analyses of the Generalised Additive Model (GAM) and Shape Constrained Additive Model (SCM) on cadence . . . 143
    6.1.2 Analyses of the GAM and SCM for the distribution on pace 147

6.2 Case study B: the trail specialist 151
6.2.1 Analyses of the GAM and SCM on cadence 151
6.2.2 Analyses of the GAM and SCM for the distribution on pace 156
6.3 Case study C: the Comrades marathoner 159
6.3.1 Analyses of the GAM and SCM on cadence 160
6.3.2 Analyses of the GAM and SCM for the distribution on pace 163
6.4 Case study D: the heart-rate runner 165
6.4.1 Analyses of the GAM and SCM on cadence 165
6.4.2 Analyses of the GAM and SCM for the distribution on pace 169
6.5 Conclusion 171

7 Final remarks 176

A Profile (Euodia Vermeulen) 186
A.1 Education 186
A.2 Research interests 186
A.3 Publications 186
A.4 Awards 187

B Informed consent form 188

List of Figures

2.1 The data to wisdom maze by David Somerville (Kaushik, 2016) 9
2.2 Mary Keitany in the home straight (Monti, 2017) 11
2.3 Metabolic rate as a function of body temperature (Denny, 2017) 19
2.4 The drowning statistician (Savage, 2000) 20
2.5 Illustrated shape constrained curves 26

3.1 Simple boxplot of randomly generated variables 43
3.2 The transformation of cadence data from a time-series to the distribution curve for a single session. 46
3.3 The straight linear regression line with the distribution of y_i 48
3.4 Fitted segmented function from scatter plot data (Vaidyanathan, 2012) . . 57
3.5 The basis functions for a cubic regression spline (Smith, 2015) 62
3.6 The thin plate regression spline for two variables, x, y on z 65
3.7 The basis functions for a B-spline regression (Smith, 2015) 67
3.8 The Probability Density Funstion (PDF) curve for a Gaussian normal distribution 73
3.9 The Kullback-Leibler (K-L) distance from candidate models g_i to reality function f 83
3.10 The possible patterns exhibited in a residual plot 87
3.11 The scatter of error terms around the normalised straight line 88

4.1 The raw data for each observed data point 95
4.2 The spread of pace across cadence (median indicated by the line and mean indicated by the dot) for athlete 3 95
4.3 Fitted lines and residual plots before and after the data clean-up for athlete 3 98
4.4 The raw data for each observed data point for athlete 4 100
4.5 The spread of pace across cadence for athlete 4 101
4.6 Fitted log-transformed lines (response scale) and residual plots before and after the data clean-up for athlete 4 104
4.7 The raw data for each observed data point for athlete 2 105

4.8 The spread of pace across cadence for athlete 2 106
4.9 Fitted lines and residual plots before and after the data clean-up for athlete 2. 109
4.10 The raw data for each observed data point for athlete 5 110
4.11 The spread of pace across cadence for athlete 5 111
4.12 Fitted lines and residual plots before and after the data clean-up for athlete 5. 114

5.1 The two dimensional (2D) representations for the combinations of pace, cadence and grade for athlete 3 119
5.2 The spread of grade across cadence 121
5.3 The 3D curves as a result of the interaction between grade and cadence for athlete 3 123
5.4 The raw data for the combinations of pace, cadence and grade for athlete 4 125
5.5 The spread of grade across cadence for athlete 4 126
5.6 The 3D curves as a result of the interaction between grade and cadence for athlete 4 128
5.7 The raw data for the combinations of pace, cadence and grade for athlete 2 131
5.8 The spread of grade across cadence for athlete 2 132
5.9 The 3D curves as a result of the interaction between grade and cadence for athlete 2 133
5.10 The raw data for the combinations of pace, cadence and grade for athlete 5 135
5.11 The spread of grade across cadence for athlete 5 136
5.12 The 3D curves as a result of the interaction between grade and cadence for athlete 5 139

6.1 The distribution curves for cadence from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 3. 145
6.2 The distribution curves for cadence from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 3. 146
6.3 The distribution of cadence subset by running activity for athlete 3 147
6.4 The distribution curves for pace from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 3. 149
6.5 The distribution curves for pace from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 3. 150
6.6 The distribution of pace subset by running activity for athlete 3 151
6.7 The distribution curves for cadence from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 4. 153

6.8 The distribution curves for cadence from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 4. 154
6.9 The distribution of cadence subset by running activity for athlete 4 155
6.10 The distribution curves for pace from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 4. 157
6.11 The distribution curves for pace from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 4. 158
6.12 The distribution of pace subset by running activity for athlete 4 159
6.13 The distribution curves for cadence from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 2. 161
6.14 The distribution curves for cadence from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 2. 162
6.15 The distribution of cadence subset by running activity for athlete 2. 163
6.16 The distribution curves for pace from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 2. 164
6.17 The distribution curves for pace from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 2. 166
6.18 The distribution of pace subset by running activity for athlete 2 167
6.19 The distribution curves for cadence from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 5. 168
6.20 The distribution curves for cadence from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 5. 170
6.21 The distribution of cadence subset by running activity for athlete 5. 171
6.22 The distribution curves for pace from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 5. 172
6.23 The distribution curves for pace from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 5. 173
6.24 The distribution of pace subset by running activity for athlete 5 174

List of Tables

2.1 Data mining in sport 16

2.2 Flaw of averages 21

2.3 The application of interaction models 23

2.4 Generalised Additive Modeling 25

2.5 The SCM in application 28

2.6 Data visualisation in medicine and sport 30

4.1 Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 3 (original data) 97

4.2 Estimated parameters for the log-transformed interaction model from cadence and running activity on pace (cleaned data) 97

4.3 Summary from the GLMs for the effect of cadence on pace for athlete 3 . . 99

4.4 Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 4 (original data) 102

4.5 Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 4 (cleaned data) 102

4.6 Summary from the GLMs for the effect of cadence and running activity on pace for athlete 2 103

4.7 Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 2 (original data) 107

4.8 Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 2 (cleaned data) 107

4.9 Summary from the GLM for the effect of cadence and running activity on pace for athlete 2 108

4.10 Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 5 (original data) 112

4.11 Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 5 (cleaned data) 112

4.12 Summary from the GLMs for the effect of cadence and running activity on pace for athlete 5 113

5.1 Correlation coefficients between cadence, grade and pace for athlete 3 . . . 118
5.2 Descriptive statistics for the effect of gradient on pace for athlete 3 (pace in min/km) 121
5.3 Estimated parameters for the log-transformed interaction model from cadence and slopes on pace for athlete 3 (cleaned data) 122
5.4 Summary from the thin plate and polynomial interaction models for cadence, grade and pace for athlete 3 122
5.5 Actual and fitted values from the thin plate model for athlete 3 123
5.6 Correlation coefficients between cadence, grade and pace for athlete 4 . . . 124
5.7 Descriptive statistics for the effect of gradient on pace for athlete 4 127
5.8 Estimated parameters for the log-transformed interaction model from cadence and slopes on pace for athlete 4 (cleaned data) 128
5.9 Summary from the thin plate and polynomial interaction models for cadence, grade and pace for athlete 4 129
5.10 Actual and fitted values from the thin plate model for athlete 4 129
5.11 Correlation coefficients between cadence, grade and pace for athlete 2 . . . 130
5.12 Descriptive statistics for the effect of gradient on pace for athlete 2 132
5.13 Estimated parameters for the log-transformed interaction model from cadence and slopes on pace for athlete 2 (cleaned data) 133
5.14 Summary from the thin plate and polynomial interaction models for cadence, grade and pace for athlete 2 134
5.15 Actual and fitted values from the thin plate model for athlete 2 134
5.16 Correlation coefficients between cadence, grade and pace for athlete 5 . . . 135
5.17 Descriptive statistics for the effect of gradient on pace for athlete 5 137
5.18 Estimated parameters for the log-transformed interaction model from cadence and slopes on pace for athlete 5 (cleaned data) 137
5.19 Summary from the thin plate and the polynomial interaction models for cadence, grade and pace for athlete 5 138
5.20 Actual and fitted values from the thin plate model for athlete 5 140

6.1 Model adequacy measures for the distribution regressions for athlete 3 . . . 143
6.2 Model adequacy measures for the distribution regressions for athlete 4 . . . 152
6.3 Model adequacy measures for the distribution regressions for athlete 2 . . . 160
6.4 Model adequacy measures for the distribution regressions for athlete 5 . . . 167

Acronyms

2D two dimensional. v, 117, 118, 123, 124, 127, 130, 133, 135, 137, 139, 140, 176

3D three dimensional. 5, 76, 77, 118, 122, 123, 127, 128, 133, 137, 139, 140, 176

AFL Australian Football League. 16

AIC Akaike Information Criteria. 28, 77, 81, 82, 84, 85, 86, 89, 91, 122, 128, 133, 137, 143, 144, 148, 152, 156, 160, 165, 169, 174

ANOVA Analysis of Variance. 93, 118

CDF Cumulative Distribution Function. 73, 74

DR downhill running. 13, 75, 117, 118, 120, 121, 127, 130, 132, 133, 135, 137, 138, 176

FoA Flaw of Averages. 4, 7, 19, 21

GAM Generalised Additive Model. ii, iii, v, vi, 4, 5, 6, 24, 25, 26, 28, 58, 62, 68, 76, 77, 78, 81, 141, 143, 144, 147, 148, 151, 152, 155, 156, 159, 160, 163, 165, 169, 174, 177

GCV General Cross Validation. 70, 71, 91

GLM Generalised Linear Model. 4, 22, 24, 25, 50, 52, 58, 59

GPS Global Positioning System. 1, 3, 4, 8, 9, 14, 16, 30, 33, 37, 38, 39, 43, 99, 116, 141, 176

ICM Interaction Model. 22, 52, 96, 101, 102, 106, 107, 111, 113, 115, 118, 121, 122, 127, 128, 132, 133, 137

IMU Inertial Measurement Units. 16

IQR Inter-quartile Range. 42, 96, 100, 103, 110, 115, 116, 120, 126, 130, 136, 139

K-L Kullback-Leibler. iv, 83, 82, 83, 84, 85, 86

LOESS Locally Weighted Regression Modeling. 58, 60, 76

LR level running. 13, 75, 117, 118, 121, 127, 132, 137

MEM Microelectromechanical Systems. 1

MLE Maximum Likelihood Estimate. 82, 84

MLR Multiple Linear Regression. 49, 51, 59

PDF Probability Density Funstion. iv, 71, 72, 73

SCM Shape Constrained Additive Model. ii, iii, v, vi, vii, 4, 5, 6, 26, 28, 58, 65, 67, 68, 77, 78, 81, 143, 144, 147, 148, 151, 152, 155, 156, 159, 160, 163, 165, 163, 167, 169, 167, 169, 174, 175, 177

SCOP Shape Constrained P-splines. 67

SCS Shape Constrained Splines. 28

SLR Simple Linear Regression. 47, 49

TPS Thin Plate Spline. 5, 64, 65, 77, 118, 122, 123, 127, 128, 129, 133, 137, 138, 139, 140, 176

UR uphill running. 13, 75, 117, 118, 120, 121, 127, 132, 133, 135, 137, 138, 176

USA United States of America. 2

WLS Weighted Least-Squares. 58

Chapter 1

Introduction

The human body was designed to move, irrespective of athletic capabilities. Over the past decades sports have seen tremendous progress towards improved performances and the methodologies behind it. The concept of tracking an individual's movements during daily activities and their chosen sport has shifted the paradigm of performance enhancements into the hands – or as a matter of speaking – onto the wrists of the athlete, be it on the elite level or the recreational and the lifestyle athlete.

1.1 Winning with Numbers

In 2016 Elaine Thompson from Jamaica took Olympic gold in the women's 100m in 10.71 seconds (IAAF, 2016). One year later at the World Championships she won her semi-final in the event in 10.84 seconds and was expected to win the gold medal. However during the final race 2.5 hours later she faded into fifth position and finished in 10.98 seconds, falling out of the medal positions. Tori Bowie won the race in 10.85 seconds after running 10.91 seconds in her semi-final, 0.07 seconds slower than Elaine Thompson's semi-final time (IAAF, 2017b). Was there some indication in the running data of Elaine Thompson that night that could have estimated her performance in the final race? Could she have used this data to pace her semi-final and save energy for the final to improve her chances of winning, or at least medaling?

A new discipline is being developed that combine the knowledge of sport science, human physiology and kinetics, mathematics, big data, data science and statistics to answer questions such as this. This discipline is referred to as Sport Analytics. Large data sets are becoming available via technological advances in data capturing during sport participation or training (Passfield and Hopker, 2016). Wearable Microelectromechanical Systems (MEM) hold the possibility to improve training methods and aid in injury prevention actions (Wilkerson et al., 2016). A range of data can be captured using these wearable devices such as Global Positioning System (GPS) running watches, activity trackers, smart phones with monitoring applications and high speed cameras set up during races or games to name but a few. The mining and analysis of the datasets generated by these devices is changing sport science, as it provides the capability to establish links between on-field physiological and performance variables during training and later performance outcomes during competition. Innovative approaches to the descriptive and predictive analytical models that is built on the data may advance talent acquisition and development, training prescriptions, provide more accurate prediction of performance outcomes and even early injury detection (Passfield and Hopker, 2016). Passfield and Hopker (2016) commented that the study by Kosmidis and Passfield (2015) shifted the scientific paradigm of training prescription. By leveraging large data sets generated by wrist worn running watches they were able to determine which part of a runner's training methods in the field are most effective as opposed to testing re-active laboratory based prescribed training. Running watches have become increasingly available to athletes from different levels of competency. Both the elite and the recreational runner now have access to important metrics pertaining to their running form. Napier et al. (2017) reports on the advantages of using wearable tracking devices such as running watches in the rehabilitation of gait patterns on runners. Running gait or form assessments can now be moved from restricted clinical settings into the real world of slopes, streets, uneven surfaces and trails, different footwear and varying physiological states during activities.

1.2 Problem Background

Running as a sport is easily accessible and requires minimum equipment. The United States of America (USA) saw a sky-rocketing increase in road race finishers of 300% between 1990 and 2013. In 2015, 17.1 million finishers were reported. Although the USA running community saw somewhat of a decline in the number of finishers between 2015 and 2016, running as a sport and the business of running is still strong, with the road running industry being valued at \$1.4 *billion* (RunningUSA, 2016). Two iconic road races in South Africa have experienced tremendous growth since their inception. The Comrades Marathon entry cap for this year was increased by 1500 entries to satisfy the demand, after the original cap of 20 000 was reached (Comrades, 2018). Over the last decade, the Ultimate Human Race had a 37.9% increase in the number of finishers. That is 3808 more runners who took on the challenge and succeeded (Comrades, 2017). The Old Mutual Two Oceans Marathon grew from 26 runners in 1970 to nearly 26 000 entries in 2017 (OMTOM, 2018a). The world's most beautiful marathon generates approximately R675 million of income, raises R3.5 million for charity and creates thousands of jobs for the

City of Cape Town (OMTOM, 2018b). The market for wearables showed an increase of 7.3% from the third quarter in 2016 to Q3 in 2017. A total of 26.7 million devices were shipped globally during Q3 of 2017 (IDC, 2017). With such a large and growing running community and continued accessibility to the big data generated by wearable devices, the scene is set for data science and analytics to provide meaningful insight that will enhance the sport.

This Masters thesis initially set out to find new value in the big data generated by runners' GPS enabled running watches. This research project attempted to develop a minimum viable mathematical product to aid the monitoring, optimal training, development and career longevity of middle to long distance runners by mining the data generated by their GPS running watches during training and races. The focus was specifically on running form and their running environment. Running form refers to the cadence (or stride rate, i.e. strides per time unit) an athlete runs at and the pace they are able to achieve at various stride rates. The running environment is related to the surfaces on which the athletes ran as well as the inclination of the surface (i.e. uphill, downhill or level ground). The running activities are the proxy to the surface as well as the purpose of the training session and is subdivided into road running and racing (asphalt), track training (grass) and trail running and racing (gravel and single-tracks).

Gradually it became apparent that the data itself should be the focal point. The manufacturer of the devices clearly state that the data from these devices are meant to encourage a healthy, active lifestyle and not to prevent or cure any disease. On the other hand there is a call from medical and sport science research fields (Napier et al. (2017), Passfield and Hopker (2016)) to capitalise on the data generated by wearable devices such as running watches for the purposes of rehabilitation and athlete monitoring. However, scientific research relating to human performance and medical treatment requires high quality and reliable data sets that are as close representative of the truth as possible. In order to proceed with the use of the data from wearables, as suggested by the literature, the veracity of the data became a central figure.

1.3 Problem Statement

The techniques of data mining and in particular data visualisation on a large scale to identify patterns, outliers and associations have the potential to shift the boundaries of running form analyses and research across the athletic spectrum in the running community. The opportunity to use this data to develop insightful metrics and data presentation techniques pertaining to running form warrants further exploration and investigation. The center of interest for the research presented in this thesis is to open up the GPS container files and spread out the spatio-temporal data generated by the GPS running watches in order to analyse its patterns, its adherence to the results of published empirical field work and to explore visualistion of the data as an analytical approach to aid athletes and coaches in understanding their capabilities.

1.4 Research Design

The research deliverable is a comprehensive report on how and to what extent the data from the runners' watches can be used to analyse running form. The focus is on the data's fitness for the purpose to accurately report on running form on which statistical models can be constructed with the aim of explaining and/or predicting running form and physical performance during overground running. The project explored approaches in the analysis and visualisation of the data from runners' wrist-worn GPS watches. The study harnessed infinite time-series cadence, speed and altitude data and transformed it into concise descriptive and meaningful statistical models that represent all the data and are not limited to a time scale. These models were constructed using Generalised Linear Model (GLM) techniques as well as the regression analyses utilities of the GAM and SCM techniques illustrated by Otto (2012) and Kosmidis and Passfield (2015) respectively. Whereas Kosmidis and Passfield (2015) focused on the running speed of well-trained athletes during structured training sessions, this research project focused on the running form of lifestyle and recreational athletes in real world settings, which included structured track sessions, road running, road races, trail running, trail races and running on slopes.

Of key importance is that this research was focused on the individual athlete, it is not group research. This way the Flaw of Averages (FoA) was avoided and the question *"How does the individual athlete perform or respond?"* found an answer.

1.5 Research Method

Data generated from the wearable wrist devices that the athletes use during training and races were extracted from their on-line profiles after the data had been synced from the device to the online application. This data set include the GPS location (latitude and longitude), the altitude, a date-time stamp, distances covered, cadence and heart rate for each second over the course of the training session or race. Three variables were jointly and separately analysed, modeled and visually presented to describe the athlete's running form in terms of:

- Cadence (or stride turn-over);

- pace (time unit per distance);
- gradient (or grade of the slope).

Pace is a performance metric that is well understood and grasped by a runner. Minutes per kilometer is a more tangible metric to an athlete than meters per second or kilometers per hour. A smorgasboard of four analysis models are presented. Two interaction models were constructed in the study from the data on the lowest granular level:

1. The interaction between cadence and running activity and their effect on pace.
2. The interaction between cadence and grade and their effect on pace.

The lowest granular level accounts for every data point, that is every instance of the data. One data point represent one second. The interaction between cadence and grade was modeled using the Thin Plate Spline (TPS) approach to generate a three dimensional (3D) visual output. Two more models were constructed and are based on the work by Kosmidis and Passfield (2015). These models present the data on the highest aggregated scale. They represent the distribution curves from the histograms and density functions for cadence and pace. The aggregated models were also subset into the running activities.

The extracted data was organised, analysed and visualised in using a combination of the programming languages *R* and *SQL*. A database in *SQLite* was built in order to manage and store the data after extraction. The following packages in *R* were used in the study:

- *xml* to extract the data from GPS container files.
- *RSqlite* to communicate with the database.
- *ggplot* to visualise the time-series-, density and distribution curves as well as the output from the regression analyses.
- *mgcv* to construct the GAMs for the distribution curves and the joint analyses on pace, cadence and elevation.
- *scam* for the SCM for the distribution curves for pace and cadence.

The athletes continued with their normal training program, race schedule and tactics, i.e. there was no intervention from the researcher or tests for pre- and post-effects of any intervention. The research environment was therefore uncontrolled and representative of real world training and racing. Athletes were observed during track training from time to time in order to develop a mental image on their running style and bio-mechanics, which assisted the researcher during interpretation of the data models. Each athlete received a unique ID, in the form of Athlete 1,2,3 etc., which was stored with their data in the database. A local running club was approached where candidates for the study were recruited. All participants were older than 18 years of age were in good health. Because it is a series of case studies and the results were not generalised to the running population, the sample size was kept small. However, the data set per individual athlete was large with 12 weeks of data collected per athlete. Ethical approval for the study was obtained from both the Engineering and Health Sciences Ethics Committees at the University of Pretoria. Informed consent (Appendix B) was obtained from the four athletes that participated in the study.

1.6 Structure of the document

Chapter 2 is dedicated to the literature review. Chapter 3 describes the methodology (the process of data gathering) and the mathematical methods behind the fitted statistical models. The results and discussion of the models each have their own chapter and include suggestions for future work. These chapters are divided as follows:

- Chapter 4: The analysis of the interaction between cadence, the running activity and pace.
- Chapter 5: The analysis of the interaction between cadence, gradient of the running surface and pace.
- Chapter 6: The application of the GAMs and SCMs on the aggregated running form analysis.

The work concludes with final remarks in Chapter 7.

Chapter 2

Literature review [1]

This chapter contains the systematic literature review to showcase the work done in the fields of big data and sport, sport science and the chosen running metrics, the FoA and the regression methods utilised in the exploration of the data sets. A section is also dedicated to the ethical considerations surrounding big data and sport, which might not be apparent but have the potential to derail data analysis techniques and conclusions.

2.1 The conversion of terabytes of numbers into wisdom

Big data can be characterised in terms of the three V's (Kitchin, 2015):

- Volume: enormous data set sizes measured in terabytes or petabytes;
- Velocity: data is being created and transmitted in near-real time which results in an extremely fast arrival rate.
- Variety: the organisation of the data is diverse and presented as structured, semi-structured and unstructured.

Further the data is all-inclusive as it attempts to capture the whole population or system of interest. The resolution is granular (low-level of detail) and relational, meaning the data contains common fields that permits data sets to join. The data is both flexible in that it can easily add new fields and scalable with rapid expansion in size (Kitchin, 2015). The data is also highly variable in that it changes quickly.

The shear volume of fitness tracking data becoming available through wearable devices is a major challenge: what data to use and how to draw meaning from it. It reminds of a quote by John Naisbitt in 1982, which still holds value today:

---

[1] A modified portion of this chapter was accepted by the IEOM Conference 2018, South Africa.

“We are drowning in information but starved for knowledge” – *John Naisbitt*.

For instance, in one 5 *km* training run session of 30 minutes an athlete may produce more than 1800 unique lines of data across multiple variables. These data points start to add up as training and race frequency and the volume thereof increase. Kosmidis and Passfield (2015) collected 2.5 million time points in their analysis on running speed of well-trained athletes. Key to effective athlete performance and form monitoring is the selection of actual variables that are being monitored and the frequency of monitoring (Rein and Memmert, 2016). “To measure is to know” is the saying. However, measuring the wrong metric provides the wrong information and results in poor decision making. Adding to the volume of data problem, information overload is also not preferable as too much variable monitoring may become tedious, costly and lead to confusion.

Sands et al. (2017) emphasise that although something can be measured, it is not necessarily valuable. The premise is to do minimal measurement or testing but with maximum return of information. Information on measurements derived from the running tracking data provide runners with the general outcome of the training session or race, however it does not provide practical intelligence to the athlete pertaining to their holistic form or whether his or her training is effective. Averages of metrics only give the athlete a partial view of their performances, and may obscure important information necessary for decision making (Sands et al., 2017). An athlete may have been in good form for only a small percentage of the training session but was running out of form for the majority of the session. The average of a desirable metric may then be slightly more favourable due to the good form for a short period and not be representative of the whole training session.

Wisdom on the other hand is the implementation and use of the data to develop insight and practical intelligence. It is this practical intelligence that will enhance the athletes' and coaches' decision making capabilities. The aforementioned athlete must be know and understand how he or she performed throughout the whole training session or race in order to identify problematic areas and recognise improvements in running form, and not be mislead by a single number. This in turn will lend support to training methods, conditioning and race tactics in order to promote healthy running and avoid injuries. The data-to-wisdom maze cartoon by David Somerville shown in Figure 1 is an analog for the progression of raw data that has been mined from various sources to the beneficial use thereof. Raw data are just random dots on a page with no pattern, i.e. the GPS tuples generated by the device and stored on the server database. Information adds colour to the blank dots and starts to separate the random points. Information provides the athlete with the numbers and figures on how they performed during a run. It tells them how fast they were running, what their step frequency was, how far did they run and much

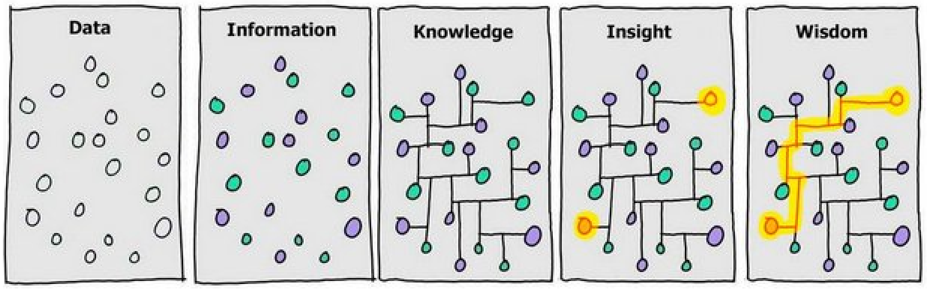


Figure 2.1: The data to wisdom maze by David Somerville (Kaushik, 2016)

more. However, these will remain just numbers if the runner does not understand them or knows how to interpret them. Knowledge shows the connection between the data points. To gain knowledge, the runner must learn and understand the context of the information and how the metrics are linked. Knowledge allows the runner to start thinking about the information they have at their disposal and how to actually use it to their benefit. Insight provides the starting and endpoints in the maze. When an athlete starts to impart their knowledge into practice, they develop insight into their capabilities and can use the insight to train better and relay their training into good racing results and general health. Wisdom shows the map to complete the journey from raw meaningless GPS tuples to practical intelligence that provides decision support to athletes and coaches regarding training regimens and race tactics.

Regardless of all its advancing possibilities, there is one characteristic that is not often mentioned in the literature that may make the data questionable: its veracity. Big data may display much commotion and noise with its accuracy sometimes drawn into question (Ahmed Memon et al., 2017) and subsequently the validity of the data becomes uncertain. These characteristics (now the four V's) are apparent in the running data captured by wearables. Cortes et al. (2014) reports on the number of workouts and resulting size data sets generated over five months by users of fitness applications, in particular runners. One month could reach as much as 37 558 648 workouts and a minimum of 16 510 934 workouts, which attest to the volume of the data sets. GPS data are generated as a runner moves along his or her route and is sent from the device to the online server as a tuple. The tuple contains, per time stamp, the latitude and longitude, the distance moved, the pace and the altitude. The maximum frequency (or put otherwise, the velocity) of GPS tuple data generation and transmission was roughly 25 000 tuples per second with a minimum of 10 000 tuples per second. The estimated number of tuples generated per month ranges between 2.8 and 6.3 billion (Cortes et al., 2014). Kosmidis and Passfield (2015) collected 2.5 million unique data points in their study on elite distance runners. The data were collected during the whole training and/or race session, with unique data points generated for nearly every second of the running activity. The data were diverse and had a very low level of detail (all-inclusive and presented with a variety in structure), ranging from GPS locations to physiological data such as heart rate. The data were generated in near-real time as the runner moved along their route and could be synced to an online platform during or right after the run. However, erroneous readings (poor veracity) which presented as outliers and missing data were reported and had to be cleaned or interpolated by the analysts before data interpretation. The outliers included extremely large values for distances covered in a short amount of time, resulting in humanly impossible speeds.

2.2 The importance of remaining injury-free

Injuries break the momentum of an athlete's career development or growth and can become costly to sponsors of elite athletes, the lifestyle athlete as well as recreational runners. The health benefits of running include weight control, physical functioning and cardio-respiratory function to name but a few. However, running increases risk of injury due to overuse and overloading of bio-mechanical structures on the runner's body, especially the spine and lower limbs (Hendricks and Philips, 2013). Hendricks and Philips (2013) performed a prospective study on a local running club in Cape Town, and found the prevalence rate of a running related injury to be 32% with an incidence rate of 0.67 per $1000km$ run. In their paper they cite another study which found an incidence rate of 30% per 1000 hours of running and a total of 163 new injuries from 629 runners over 8 weeks. Gijon-Nogueron and Fernandez-Villarejo (2015) reports on injury rates in their systematic review of the literature. One study reported that 63% of runners have had at some time in their lives experienced an injury to the lower limb. A total of 23% experienced symptoms that lasted longer than 6 months as localized pain. These runners had to change their training or lower the volume thereof, seek medical attention or used medication for the management of the injury. Between 30 to 70% of runners had to reduce training and up to 79% sought medical assistance.

Hendricks and Philips (2013) point out that common running related injuries may be mistreated because the underlying cause of the injury is overlooked. The wrongful diagnosis of an injury may be due to inadequate knowledge on the side of the practitioner of the pathophysiology and poor bio-mechanics related to running, erroneous patient history and incorrect physical examination. In defense of an ignorant wrongful diagnosis, running injuries are multifaceted and complex, with many intrinsic and extrinsic factors adding to or reducing risk factors. Risk factors range from footwear, training intensity and volume, bio-mechanical- and neuro-muscular responses to running and physiological statuses of a runner (Gijon-Nogueron and Fernandez-Villarejo, 2015). A single examination or observation does not provide the full picture of the athlete. However, with the wisdom that can be gained from data mining and analysis on a large scale the medical professional may be provided with another tool to evaluate patients with running related injuries. An image of the athlete's running form painted with numbers will provide the practitioner, coach and athlete alike with the opportunity to see a fuller picture and not just isolated evaluations in limited settings.

In team sports, an injury effectively results in downtime of the most important asset of a team – their athletes. Once injured, an athlete must withdraw from competition and go through a rehabilitation process in order to return to the sport. During this time, they are not a contributing member of the team and they are still dependent on the sponsor for their medical treatment and rehabilitation. Their career might also be in jeopardy, depending on the severity of the injury. Although running is mostly viewed as an individual sport, elite runners may form teams where runners support each other and make use of pacemakers to improve their final time. For instance, the women's only world record marathon time of 2:17:42 set by Paula Radcliffe in 2003 was broken by Mary Keitany from Kenya at the 2017 London Marathon (shown in Figure 2.2 on page 11), where she was paced by her training partner Caroline Kipkirui. She therefore did not have to set the pace herself, and could take advantage of the protection of pacer in front of her which saved her energy for the last portion of the race. They were running as a team. She eventually went out on her own finishing in a time of 2:17:01, taking 41 seconds of the world record (IAAF, 2017a). Injuries or poor performances in an elite team of runners such as this will impact on the final outcome of the race. Another example of team-based running is the relay. Athletes in relays are of different capabilities, however a team might fail in their outcome aspirations should the rest of the team not be able to pick up the slack left by an under-performing runner. Athletes across the spectrum and coaches who

Figure 2.2: Mary Keitany in the home straight (Monti, 2017)

are capable of preventing injuries and promoting performance by effective data-driven monitoring of running form will have a competitive advantage. For this to materialise, they need their data to be communicated clearly, understandable and on-time.

2.3 Metrics for running form

Three metrics are used to describe running form: cadence, surface grade (i.e. gradient of slopes) and pace.

2.3.1 Cadence and pace

Common metrics used in running form is cadence (stride turn-over in steps per minute) and stride length in meters (Heiderscheit et al., 2011). These two metrics govern speed as shown in Equation 2.1:

$$Speed = Cadence \times Stride\ Length \tag{2.1}$$

Speed may therefore be increased by either increasing the cadence, the stride length or both. However, the former is strongly supported in the systematic review of the literature of Schubart et al. (2013). Increasing the stride length may pre-dispose the athlete to over-striding, whereby the the body's centre of mass is behind the base of support when the foot strikes the ground. They concluded that reducing the magnitude of the load on the ankle, knee and hip joints and the lower spine during running by a higher step rate reduces the risk for injury. Napier et al. (2017) reports that the increase in cadence, lowering of vertical load in the lower limbs and foot strike adjustments have the most consistent effect on injury incidence. Heiderscheit et al. (2011) studied the changes in joint mechanics during altered levels of runners' preferred cadence. Runners' preferred step rates were varied between $-10\%, -5\%$, $+5\%$ and $+10\%$. Results showed that less mechanical energy was absorbed at the knee joint during intervals with 5% and 10% increases in step frequency while maintaining a constant speed. The decrease in mechanical energy absorption subsequently lowers the load on the lower limbs and joints during running, whereas the increase in energy absorption increases the load. More energy was absorbed when the step rate was decreased at a constant speed. To achieve constant speed to adhere to Equation 2.1, stride length had to be either decreased in the case of a higher cadence whereas it was increased for the lower cadence.

However, increasing cadence comes at a metabolic cost to the runner. More muscle activation (and thus energy production) of the lower limbs is required to increase the turn-over of steps. Re-training the neuro-muscular system of the athlete may therefore lead to an initial decrease in performance due to the body that must adapt to the increases in muscle activation (Heiderscheit et al., 2011). Nonetheless, following good conditioning programs will improve muscle strength and re-train the neuro-muscular system to adapt to the higher cadence or alteration in running form (Brukner and Khan, 2006). By increasing cadence and keeping stride length within a safe range, combined with a good conditioning regimen, a runner may favourably influence his/her performance and subsequently run at higher speeds, whilst at the same time protect themselves from injury risk factors.

2.3.2 Graded running [2]

Running on slopes (i.e. graded running) impacts the biomechanics and the musculoskeletal system of the human body during motion. The understanding of how graded running affects a runner's body have significant implications in injury prevention, performance improvements and training. Chen et al. (2008) studied the effects of 30-min level running (LR) for six consecutive days after a 30-min downhill running (DR) session. They found significant changes in muscle damage markers as well as running economy for seven days after the DR. Downhill running is known to cause delayed onset muscle soreness. The LR sessions did not seem to have positive neither negative effects on the recovery of the runners. Studies on graded running is less abundant than research on LR. The systematic literature review by Vernillo et al. (2016) delivered only nine significant studies where the effect of graded running on spatio-temporal variables were assessed. Results from these studies do not always agree on significant changes in cadence, contact time and aerial time in uphill running (UR) and DR when compared with LR. Five studies reported a significant higher stride frequency in UR when compared to LR while the rest did not report any significant changes in stride frequency. Snyder and Farley (2011) compared LR, UR and DR to find new information on how nine runners manage their stride frequency. Although the mean stride frequency did increase with 0.04 $strides\ s^{-1}$ across increases in the slope (from -3 to 0 to +3 degrees), these changes were not significant. In contrast to Snyder and Farley (2011), Padulo et al. (2013) reported an increase of 4% in stride frequency for an increase in slope from 0% to 7%. These studies were limited by time, space and subsequently data set size. The harnessing of big data generated by a running watch can extend the monitoring (or observation) time and experimentation environment and provide a comprehensive description on running form during LR, UR and DR.

---

[2] A modified version of this section was communicated to the IEEE Transactions on Big Data.

2.4 Applications of data mining in sport

This section systematically summarises studies and research completed regarding data mining and its application in the sports field, or how data generated by fitness trackers were used for other purposes with physical activity still in mind. The tables list the contributing authors, the method and/or approach followed and the conclusions and/or comments from each research contribution.

Adamakis and Zounhia (2016) tested and validated (to some extent) the use of fitness trackers (and specifically the Garmin Vivofit, which is the same manufacturer as the devices used by the athletes who participated in this research project) to count the steps of adolescents during both walking and running. Adams et al. (2016) concluded from their study that the running watch is a reliable tool to detect changes in cadence, vertical oscillation and ground contact time. They found high inter-class correlation coefficients between a commercial running watch (Garmin fenix 2), a chest strap (HRM-Run, Garmin) and the gold standard instrument for gait dynamics (Vicon; OMG plc, Oxford, UK) for cadence (0.931), vertical oscillation (0.963) and ground contact time (0.749). They extracted the same data from the Garmin Connect software as this research study and based their work on the cadence, the vertical oscillation and the ground contact time as provided by the software. On the 95% confidence interval, the minimal detectable change in cadence by the watch was 2.53 steps per minute, compared to the 1.29 steps per minute for the motion capture system. This is a relatively large difference and must be kept in mind when analysing differences in cadence over time when data from the watch are used. Therefore, although the inter-class coefficient for cadence is high, the cadence as obtained from the watch is still not as accurate as those obtained with the gold standard. However, using a watch extends the sample time and allows the athlete to run in the real-world where data can be collected over uneven surfaces and varying conditions. Although this experiment was carried out under laboratory conditions and on a treadmill on which the speed was controlled, the high correlation between the running dynamic metrics is a good indication that the watch may be used outside of the laboratory for running gait analysis despite the differences in the minimal detectable changes.

The definition of training load monitoring is categorised by Bourdon et al. (2017) as internal (relative physiological parameters such as heart rate) and external (parameters such as cadence, distance, speed, location analysis and more). Hardware (both stationary and mobile) exist to monitor both these categories, with GPS enabled devices being used more and more often in training load monitoring to aggregate external training loads. The output data from these devices are being validated and the reliability tested by various studies. Nonetheless, practitioners are cautioned that the algorithms used for the derivation of the output from these devices differ between manufacturers and software companies, and subsequently their analysis methods must be approached with this in mind (Bourdon et al., 2017). Wilkerson et al. (2016) comments that wearable devices have the potential to advance the methods and initiatives behind athlete monitoring and injury prevention, but the evidence from research is still lacking to support the data that these devices provide for decision making regarding training, performance optimisation and the reduction of injury risks.

Table 2.1 contains the summarised information from research studies for mining big data in sport (from websites) or from activity trackers, wearable devices and running watches.

Table 2.1: Data mining in sport

Metadata mining in sport (running and cycling)			
Author(s)	**Title**	**Approach or Technique**	**Conclusions**
Hochmair et al. (2017)	Estimating bicycle trip volume for Miame-Dade county from Strava tracking data	Extracted metedata on cycling behaviour. Used the data with other relevant spatio-temporal data in linear regression to quantify bicycle ridership.	Assisted in differentiating between cycling types (work commute or leisure) to assist road network planners.
Balaban and Tuncer (2017)	Visualizing and analyzing urban leisure runs by using sports tracking data.	Collaborated the GPS spatio-temporal data from Singapore runners' personal fitness trackers (data were mined from Endomondo and Strava) with climate, socio-demographic and topology data in joint analysis to develop data visualisations of the city and the runners' behaviour.	Dynamic visualisations helped city planners understand runner behaviour across time frames and their choice of routes. The attractiveness of a area for runners could be quantified that will help city planners to design the infrastructure that will accommodate a healthy and active lifestyle by its inhabitants.
Passfield and Hopker (2016)	A mine of information: can sport analytics provide wisdom from your data?	Retrospective analysis on 67 503 race results from 5561 riders who rode in 25 major junior and elite races. They extracted the data from websites using web-crawlers.	Discovered the relative age effect in world-class cycling and provide a description of the progression from elite junior to elite senior level.

Training load and performance monitoring			
Author(s)	**Title**	**Approach or Technique**	**Conclusions**
Sands et al. (2017)	Modern techniques and technologies applied to training and performance monitoring.	Review of monitoring techniques and the role of data mining technologies.	Propose a framework of athlete monitoring: trend analyses, rules-based analysis, and statistical process control.
Kosmidis and Passfield (2015)	Linking the performance of endurance runners to training and physiological effects via multi-resolution elastic net.	Developed the training distribution profile that visualised the runners' over-all performances by expressing the amount of time that they spent at each achievable speed.	Identified the effective training speeds achieved in the field that made significant contributions to performance improvements. Encouraged further development of the training distribution profile to other metrics.
Wisbey et al. (2009)	Quantifying movement demands of Australian Football League (AFL) football using GPS.	Used GPS tracking devices attached to the vest of players to describe the movement patterns of nomadic players, forwards and defenders. The researchers extracted physical load metrics such as speed, accelerations, total distance etc. and analysed the tracking data over four seasons to determine whether the physical demands have changed.	The physical demand on players have increased over four seasons. This has fitness and game regulation implications for athletes, coaches and game authorities.

Individual athlete injury detection

Author(s)	**Title**	**Approach or Technique**	**Conclusions**
Ahmad et al. (2018)	Monitoring and prediction of exhaustion threshold during aerobic exercise based on physiological system using artificial neural network.	A wearable device monitored physiological metrics of fatigue during physical exercise and a supervised machine learning (artificial neural network) algorithm was developed to predict exhaustion level (ranging from very light to maximum).	The algorithm can be incorporated into smart phone applications to predict an exhaustion threshold of its user and warn him/her of possible injury due to exertion.
Sands et al. (2017)	Modern techniques and technologies applied to training and performance monitoring.	Individual monitoring of an athlete's vital signs throughout training and during competition (time-series data).	The resting heart rate data warned that the athlete was not in good physical condition, but were unfortunately ignored and the athlete suffered a career-ending injury.
Wilkerson et al. (2016)	Utilisation of pratice session average inertial load to quantify college football injury risk.	Retrospective analysis of archived inertial loads derived from Inertial Measurement Units (IMU) of college football players to identify associations of inertial load metrics and the occurrence of injuries. A cohort of 45 athletes wore the IMU's during training for a period of 15 consecutive weeks.	Inertial loads (accumulated and variability thereof) are associated with increased risk for injury. The use of IMU is remarkably useful in guiding of individualised performance management and enhancements.

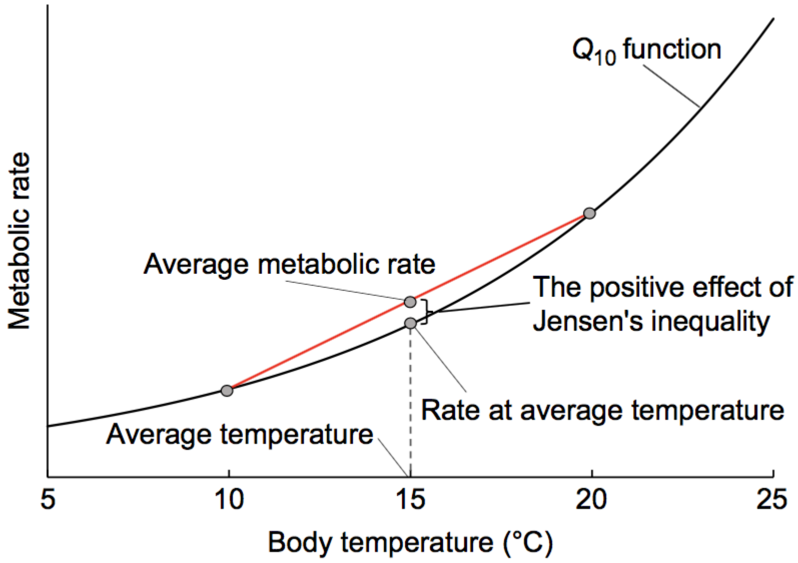


Figure 2.3: Metabolic rate as a function of body temperature (Denny, 2017)

2.5 Flaw of Averages

The FoA as explained by Kruger and Yadavalli (2016) is basically the concept that *average inputs do not necessarily provide average outputs*. Making major decisions based on single, presumably constant figures on the average number ignores the interaction between variables and may lead to subsequent under- or overvaluation of outcomes. The authors refer to Jensen's inequality as the mathematical basis for the FoA. The basic premise is that the function of the mean is not equal to the mean of the function as stated in Equation 2.2.

$$f(\overline{x}) \neq \overline{f(x)} \tag{2.2}$$

The only prerequisite of the FoA is for the input-output function to be non-linear and that at least one input variable is subject to uncertainty (Kruger and Yadavalli, 2016). Figure 2.3 provides a basic visual description on the FoA with a non-linear input-output transformation, borrowed from Denny (2017). At the average body temperature of 15^oC the actual metabolic rate is lower than the average metabolic rate at average body temperature. Denny (2017) continues to explain the FoA in statistics within the context of the variability in natural occurrences. The physiological capacity varies among individuals and ecological interactions shift from place to place and over a time period. Subsequently, the wrong assumptions and conclusions are drawn for individuals who form part of the population of interest whose value of the decision variable is based on the average of the population. This is illustrated in the comic of the drowning statistician, who made the wrong assumption about the depth of the pond being 3 feet throughout the width of the pond, as depicted in Figure 2.4 on page 20. The FoA in sport science is due to the great

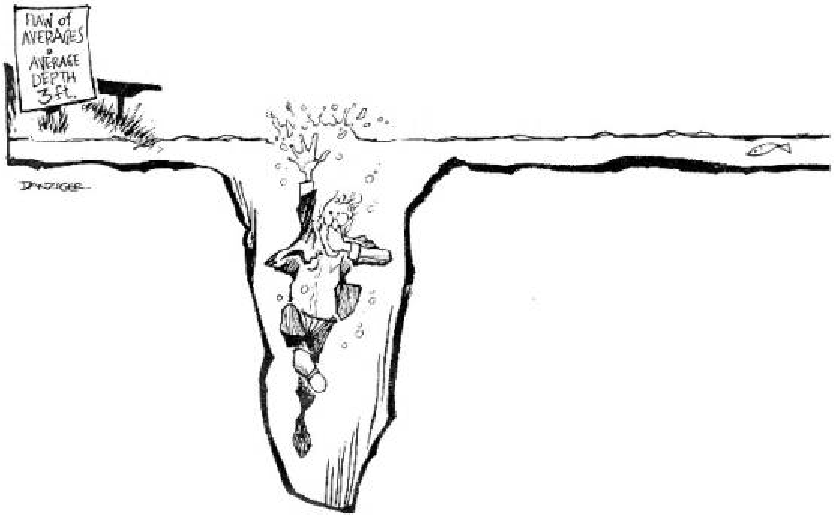


Figure 2.4: The drowning statistician (Savage, 2000)

variability in human nature and that no single athlete will respond the same to input variables. Refer to Table 2.2 for examples on focused research on the FoA.

Table 2.2: Flaw of averages

Author(s)	**Title**	**Approach or Technique**	**Conclusions**
Denny (2017)	The fallacy of the average: on the ubiquity, utility and continuing novelty of Jensen's inequality.	Mathematical and graphical illustration of the Jensen's inequality in biology.	Encourage researchers in the natural sciences to consider the FoA in response functions as to provide accurate results from their research.
Sands et al. (2017)	Modern techniques and technologies applied to training and performance monitoring.	Athletes' responses to training and competition stress are idiosyncratic. Measuring an athlete's stress response to an average of metrics obtained from group research may lead to failure in recognising physiological warning signs that an athlete is under duress and need intervention.	Proposed rule-based assessment to identify the important individual metrics or variables by which an athlete must be monitored.
Kruger and Yadavalli (2016)	Probability management and the flaw of averages.	Monte Carlo simulation on the news vendor problem to illustrate the FoA under various stochastic conditions and probability distributions.	The FoA will be present and severe when the input/output transformation function is non-linear and at least one of the input variables is stochastic.

2.6 Regression methods

The definitions and backgrounds for the regression methods used in this project are defined and outlined in detail in Chapter 3. The literature review covers only the application of these methods in real-world problems.

Interaction models are an extension of GLMs and are valuable to analyse the re-enforcing or interfering effects that independent variables have on each other and their joint effect on the response variable. The effect of multicollinearity must be considered in multiple regression analysis with interaction effects. The presence or absence of independent variables in the model may result in extensive variation of the estimated regression coefficients of the separate effects. The separate effect of the independent variable on the response remains uncertain as no absolute sense of any effect can be established . Fortunately the inferences made on the response variable remain mostly unaffected by multicollinearity, given that the inference is made within the limits of the observed independent variables (Neter et al., 1988). The interaction models in this study are only concerned with the size and direction of the interaction terms within the range of the input variables. The separate effects that the independent variables have on the response is not of current concern.

The transformed interaction models for cadence and running activity and for cadence and grades are described in Chapter 3 (mathematical background), 4 and 5. Table 2.3 contains three research projects where the value added by an Interaction Model (ICM) is illustrated. Two studies' focus areas are in sport, while the third showcases how an interaction multivariate regression model can explain a complex ecological system which contains both categorical and continuous variables.

Table 2.3: The application of interaction models

Author(s)	**Title**	**Approach or Technique**	**Conclusions**
Mahmoudi et al. (2014)	Environmental variables and their interaction effects on chlorophyll-a in coastal waters of the southern Caspian Sea: Assessment by multiple regression grey models.	Included both categorical and continuous variables in two-way interactions to model the complicated behaviour of the ecosystem throughout the four seasons of the year.	Although the R_a^2 values decreased from spring to autumn, both tools were satisfactorily able to find the most important main and interactions effects and enhance the understanding of how they influenced the response throughout the four seasons.
Ortega et al. (2010)	Cardiovascular fitness modifies the associations between physical activity and abdominal adiposity in children and adolescents: the European Youth Heart Study.	Used a multiple regression analyses to determine the associations between the input variables (physical activity and cardiovascular fitness) and waist circumference (response).	Found an unexpected and paradoxal relationship between the the level of fitness, vigorous physical activity and waist circumference, which opens the discussion for further research in the field.
Ullrich-French and Smith (2009)	Social and motivational predictors of continued youth sport participation.	Used logistic regression analyses in a hierarchical method to find the odds ratios of soccer continuation with the presence and interaction of social and relational variables.	The final model with two-way and three-way interactions between the social and relational predictors had the best capability to predict team participation continuation. The McFadden R_L^2 increased for each hierarchical addition of predictor variables and interactions.

Hofner et al. (2014) explained that the GAM is used in statistics to allow for flexible, data-driven estimation of the influence of covariates on outcome variables, where the assumption on linearity of the regression model can (and perhaps should) be relaxed. A simple search with the keywords “generalised additive models” on the “Publications” tab on the Research Gate platform delivered a trove of articles, ongoing projects and studies where the functionality of the GAM is illustrated. The application fields range from the effect of air pollution on morbidity and mortality, the lameness of dairy cows, bird species density, spatial distribution of marine life, river nutrient concentrations in water mechanics models, dynamic analysis in linguistics and much more. The GAM has shown its value to analyse data and explain relationships where the system dynamics are complex and no mechanistic models exist to fit the data. Dominici et al. (2002) reports on the use the GAM in the non-linear regression analysis in various medically related fields. In time-series studies on air pollution and mortality the GAM was the most widely used, because it can account for seasonality, trends, and weather variables. Table 2.4 contains some application studies in real-world settings. These studies display how a GAM may be used instead of or alongside the GLM, when non-linearity is problematic and the modeler requires more freedom of the functional form than what is allowed by a GLM.

Table 2.4: Generalised Additive Modeling

Author(s)	**Title**	**Approach or Technique**	**Conclusions**
Hastie and Tibshirani (1986)	Generalized Additive Models.	Compared GLM techniques and non-parametric smooths on binary response and survival data.	The non-parametric smooths were able to identify non-linear co-variate effects.
Barrio et al. (2013)	Use of generalised additive models to categorise continuous variables in clinical prediction.	P-spline smoothers were used to find the association between the continuous predictors and the response. The continuous predictors were then categorised and regressed using the GAM to find cut-off points for the categorisation of the predictor variables.	The GAM's performance with the categorised variables were as successful as the original continuous predictors. This is useful for clinicians as now have categorised reference points to make clinical decisions and prediction rules.
Rodríguez-Álvarez et al. (2012)	Analysing visual receptive fields through generalised additive models with interactions.	Fitted a Poisson GAM with interactions to smooth the receptive field maps of a single neuron by including spatial effects. The fitted GAM also estimated the temporal evolution of the receptive field maps (they vary under different experimental conditions).	The GAM is a flexible statistical tool to map receptive fields and further work is suggested to develop specialised GAMs for the mapping of receptive fields in the visual system.

Rodríguez-Álvarez et al. (2012) followed the GAM approach suggested by Wood (2006), as a flexible method to model the temporal development of visual receptive fields, across varying conditions. Although their work has got nothing to do with sport science and is unrelated to the type of data analysed in this project, their work serves as an illustration of the effectiveness of a GAM in an unexpected application field.

Hofner et al. (2014) points out that in some situations the researcher might have priori knowledge of the effects of a input variable, such a monotonicity, cyclic characteristics or it's adherence to boundary conditions. Chen and Samworth (2016) explains that shape restricted curves can be monotonic (i.e. either increasing or decreasing, never both), concave or convex as illustrated in Figure 2.5 on page 26. Hofner et al. (2014) suggested a framework on how to deal with these shape constraints of variables in regression analysis and offered some case studies on how to apply constrained model packages developed in R. Two packages in R were used to develop the effects estimate for time on mortality due to pollution in a city. This is essentially the combination of seasonal weather patterns and the long-term trend between pollution levels and mortality. In a study such as this, the researcher has to deal with a great amount of variability in nature, both environmental (weather) and human physiological responses to air pollution and linearity should not be assumed in order to develop an accurate model. The researchers used their priori knowledge on cyclical seasonal changes in the weather during the year as well as a reasonable assumption of increasing monotonicity between air pollution levels and mortality to develop a shape constrained model. The *mboost* and *scam* packages in R displayed similar results in the effects estimate for time.

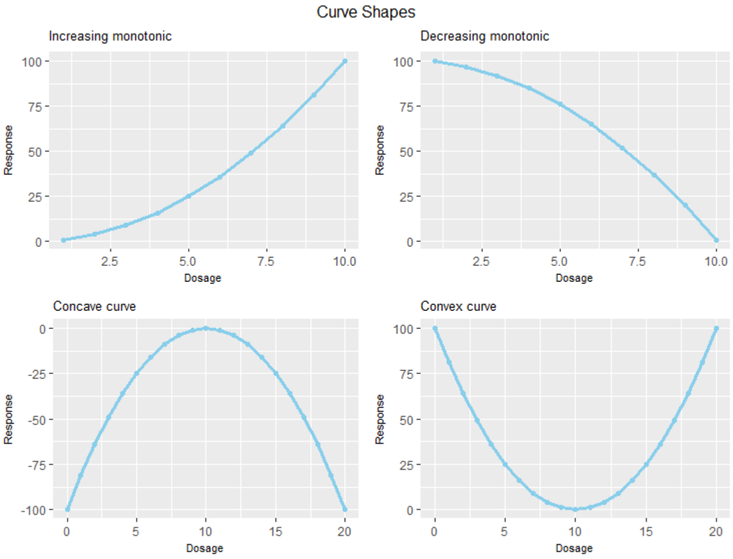


Figure 2.5: Illustrated shape constrained curves

Table 2.5 contains three studies where the SCM was implemented to model monotonic curves. Only one study is directly linked to sport analytics, but the other two studies showcase the flexibility of SCMs in medical or physiological research fields.

Table 2.5: The SCM in application

Author(s)	**Title**	**Approach or Technique**	**Conclusions**
Mašić et al. (2016)	On the use of shape-constrained splines for biokinetic process modeling.	Used Shape Constrained Splines (SCS) to describe substrate affinity with parametric flexibility in bacterial growth rates (monotonic and concave).	The SCS model is capable to fit the investigated growth rates when presented with an unconventional bacterial growth rate.
Kosmidis and Passfield (2015)	Linking the performance of endurance runners to training and physiological effects via multi-resolution elastic net.	Fitted a SCM using Poisson response variables for the amount of time spent at speed thresholds. Used the *scam* package in *R*.	The fitted line respected the positivity and monotonicity of the distribution profile. The outcome of the training distribution profile from the SCM can be used as a response or a covariate in functional regression analysis. The average times of the most effective speeds in the fields were included in the expression that predicts performance.
Pya and Wood (2015)	Shape constrained additive models.	Fitted a SCM and an unconstrained GAM on expected monotonically decreasing data relationship (decreased rate of cancer for increased distance from a municipal incinerator)	The Akaike Information Criteria (AIC) score for the SCM is lowest and the GAM resulted in a non-monotone smooth. The SCM is the preferred model.

2.7 Data visualisation – painting by numbers [3]

The concept of information visualisation for the masses was proposed by Kerren et al. (2011) as a new research direction pertaining to how information visualisation can be used to address real-world problems and opportunities. The importance of making data visualisations understandable to patients in a health care setting is underscored by Peterson (2016). When patients are provided with a visualisation that they understand they are empowered to make informed treatment decisions and are better able to monitor their condition. The same idea applies to athletes who make use of the visualisations of their tracking data. The advent of big data has brought along opportunities for data visualisations on a unprecedented scale. The visualisation of big data is able to add value and deliver new insight to otherwise inaccessible and complicated data due to its scale and velocity.

Lessons learned from the business intelligence fields have shown just how valuable a meaningful and customised visualisation of data can be to decision makers. It improves the quality of information as well as the quantity that can be absorbed and interpreted in a short time span, leading to better decisions being made more quickly. Meaningful visualistaion of training load data as generated by wearables or other monitoring devices will assist coaches and athlete in making informed decision regarding training and performance management (Bourdon et al., 2017).

The online fitness and activity tracking applications such as Strava and Garmin Connect have developed many visualisations for runners and cyclists to analyse their training and performance. These visualisations range from overlayed time-series analysis to aggregated categorised information such as the amount of time spent in heart rate zones. The effective use of visualisation in sport analytics is illustrated by Pileggi et al. (2012) who developed heat maps to indicate shot density in ice hockey. These heat maps equipped the analysts with means to quickly evaluate shooting patterns among teams and ice rinks. The shooting patterns include shot lengths, location of the shooting player and conversion rates of shots into goals. The ability to analyse teams' shooting behaviour from different positions on the ice will impact team strategies and tactics on the field. It also provided analysts with opportunities and avenues to change original theories and become more creative in their approach to game tactic ideas.

Table 2.6 contains the summaries of four studies focused on data visualisation of complicated data sets. Three of these studies mined their data from wearable activity trackers.

---

[3] A modified version of this section was communicated to the IEEE Transactions on Big Data.

Table 2.6: Data visualisation in medicine and sport

Author(s)	**Title**	**Approach or Technique**	**Conclusions**
Balaban and Tuncer (2017)	Visualizing and analyzing urban leisure runs by using sports tracking data.	Developed data visualisations of the routes and streets used by runners in Singapore from the GPS locations from their personal fitness trackers (data mined from Endomondo and Strava).	Output: 1) Dynamic geographical heatmaps of frequented routes and their temporal boundaries (time of the day as well as seasons). 2) Static heatmap with runs per kilometer per street that can be filtered on time frames and climate conditions.
Balaban and Tuncer (2016)	Visualizing Urban Sports Movement.	Collected publicly available personal fitness tracking data from users on Endomondo.	Display the fitness activities on a map and specify time, location, activity, gender, age groups etc. This information indicates usage patterns of certain spaces. This study was the fore-runner for visualisation tools for city planners (Balaban and Tuncer, 2017).
Chen et al. (2016)	Toward pervasive gait analysis with wearable sensors: a systematic review.	Used the data from an inertial sensor in a wearable device to generate geometric representations of the dynamical system of walking (also called phase portraits).	Created a visualisation of the phase portraits to analyse angular velocities, mechanical energy, gait regularity, gait stability and complexity in a clinical setting.

Author(s)	**Title**	**Approach or Technique**	**Conclusions**
Pileggi et al. (2012)	SnapShot: Visualization to Propel Ice Hockey Analytics	Developed an interactive visualisation system of ice-hockey shots based on analyst requirements, real-world contextualisation and data categorisation.	Delivered a high-resolution interactive visualisation using radial heatmaps with the ice-rink as backdrop. Analysts can filter the data, compare different scenarios, present to stakeholders and collaborate with each other.

The visual presentation of data of the runners' behaviour to designers and city planners is a convincing method to explain how the phenomenon works. However, the visualisation does not isolate a parameter's influence on the overall behaviour. Therefore analytics must be included to determine the effect of the visualised parameter on the phenomenon (Balaban and Tuncer, 2017). The tracking data from the fitness applications helped the researchers to isolate the spatio-temporal behaviour of the runners and to identify the areas to include in the regression model that followed.

2.8 Ethics in big data and sport

There are some concerns relating to the ethical considerations of big data in sport. Three of them will be discussed here: data validity, data security and athlete autonomy.

2.8.1 Data validity

Is the data accurate and does it correctly represent what it claims to represent? The reliability of sport tracking data is important as performance decisions are based on them. Incorrect readings will lead to over- or under- determination of performance capabilities and subsequently harmful decisions may be made. An athlete might push themselves physically to far or falsely assume fatigue due to some performance detriment (Karkazis et al., 2017). The veracity of the data is influenced by the commotion, or noise, that may accompany tracking data due to a faulty device, signal interruptions between the device and its interacting environment or the body of the user. Although a wearable presents the athlete with the opportunity to assess their physical performance, the user must still have the knowledge to distinguish between logical results and noise. Some commonly used fitness tracking devices were tested on their accuracy. For step count, the best performance of a device had an average 1.05% error rate with the poorest performing device having an average error rate of 27.28%. On measuring the distance the lowest error rate was 3.72% and the highest error rate 11.17% over 400 meters (on a track) (Fangfang et al., 2013). These statistics underscore the problem of veracity in the data. Error rates such as these will lead to inaccurate calculations on cadence (stride frequency) and running speed or pace. An athlete may now believe they are under or over-performing and subsequent change training or race tactics.

Although algorithms designed to clean and present data should be objective, they still suffer from bias due to the perspectives and assumptions of the developer. The analysis of sport biometric data presents a unique challenge to algorithms: there is an overload of data that requires interpretation but an undersupply of historical, validated data to develop a valid algorithm. Somewhere data will have to be validated for algorithms to become reliable and true representatives of the actual data that is generated by a runner. The signal-to-noise ratio in biometric data remains low when compared to data collected in a structured experiment (Karkazis et al., 2017). Interpretation of the data must always be accompanied by some domain knowledge and not just taken as the full truth at face value. On the aggregate scale, such as the data used in the urban planning studies, data suffers from bias as it is limited to the population characteristics of those who are actually using the device. In the urban planning study for cycling the data from Strava is skewed towards male users at 87%. Nonetheless the extensively large scale on which the fitness application data could be collected outweighed the bias disadvantage (Hochmair et al., 2016).

2.8.2 Data security and protection

The recent Facebook-Cambridge Analytica debacle testifies to the ease at which private user data can be extracted and the consequences thereof. Users of platforms react by reluctance to share their data (which will negatively impact research in the market area) or loose trust in the brand completely (Gupta and Schneider, 2018). Personal data generated by fitness trackers are not fully and effectively protected, with as much half of data that requires protection are actually protected (Cortes et al., 2014). Hackers may gain access to sensitive personal information when a runner records their run session. The starting and/or end location may indicate their home address. A runner's routine (i.e. when they are away from home, where they run, how long they are away etc.) may be revealed after some remote cyber surveillance. Their health data such as heart rate and other biomechanics may also be disclosed by malware or cyber-attacks. Gupta and Schneider (2018) emphasize that protecting users data requires more than just anonymising their identities. The anonymised large datasets generated from wearables and extracted for aggregation purposes such as what was done in the urban planning study are not immune against data privacy risks. It was found that 87% of the United States population may be identified by their zip code, gender and birthday (Sweeney, 2000). From GPS data it is easy to identify a runner's physical location while their gender and birthday is captured on the on-line fitness profile.

2.8.3 Athlete autonomy

In the midst of the hype surrounding the advantages that big data and sport analytics can provide, a lingering question is surfacing: where is the line between data working for the athlete and data working against the athlete? Data analytics should remain an adjunct tool in the athlete's quest for sporting excellence and career longevity, it should not become the driver of performance. Athletes risk loosing their autonomy and intuition when they completely rely on their biometric data to govern their performance efforts and race or game tactics. When an athlete is reduced to an entity basically consisting of only numbers and visualisations the enjoyment of the sport will be diminished, which is contradictory to the desired effect of sport participation in the first place for both the recreational and elite athlete. The literature has shown that it is possible for sports disciplines to capitalise on new found knowledge hidden in the extensively large data sets generated by wearables. When used correctly this knowledge lends decision support to athletes, coaches, management teams and even urban planners that was not previously possible. Caution should be exercised in the midst of the hype that might be created by big data: the wisdom gained from the data mining from activity trackers provides the athlete with a tool to make an informed decision, it should not govern the athlete.

Chapter 3

Methodology: From time-series analysis to interaction and distribution modeling

This chapter deals with the approach that was followed for the completion of this project. It describes the pathway to gather and organise the raw data and the calculation of variables. The regression analysis methods that were used are analysed and the application thereof illustrated. Finally model adequacy techniques used for model selection are discussed with the application thereof on the constructed models.

3.1 Data gathering

Ethical approval for the study was obtained from both the Engineering and Health Sciences Ethics Committees at the University of Pretoria following the required procedures set out by each committee. The following considerations were in place to ensure a sound ethical study and protection of the subjects:

- Only participants who have granted informed consent were used in this study.
- There was no incentive for participants.
- There was no sponsor with technological or financial interest involved.
- The study forms part of the researcher's fulfillment duties for MEng: Industrial and is funded by the researcher herself.
- Participants were informed that they may withdrew from the study at any point of time.

- Participants' anonymity were ensured and all their data are still safeguarded.
- Data was stored on a *SQLite* database on a password protected laptop. The database does not allow for any remote queries, so outsiders will not be able to access the data. The data that are kept on the database do not include any identifiable information. The laptop is well protected against malware. Back-ups of the data will be kept in an encrypted file in the cloud. The names of the athlete to which the athlete ID is linked will be stored in an encrypted file and known only to the researcher.
- Participants have access to all research and information relating to them personally at any stage during or after completion of the study, including the final findings and conclusion of the study.
- There were no participants under the age of 18 years.
- All data and results will be and remain the property of the University of Pretoria.
- Result may be released as part of the dissertation for attainment of the Masters Degree in Industrial Engineering and published in a selected scientific journal (yet to be determined).
- No video or other image recordings were made during the training observations.
- There was no personal conflict of interest for the researcher or other parties that promote training methods or race tactics because the athletes will carry-on with their own training program and race tactics as they see fit.
- The athlete was also under no obligation to train or participate in races.
- The raw data did not include any personal information that may identify the athlete.
- The researcher did not measure any of the data herself using instruments or apparatuses, but only used the data as generated by the wearable device.
- It was made clear to the athlete that this model will not provide the exact outcome of the race or injury status. The model cannot account for factors that are outside of the model's control, such as the weather, injury or illness developed on race day or mental state. This measure was to ensure that the athlete is not misled by the model.
- The manufacturer of the wearable devices have declared that the devices are not medical devices and that the data are not intended to be used to diagnose or cure

any ailment or disease. The data is to be used to encourage a healthy lifestyle (Garmin, 2018). This research study is also not intended to diagnose nor cure injuries, diseases or ailments.

- The data analysis technique demonstrated in this study will have to undergo further clinical evaluation to be confirmed as a usable approach in injury risk detection and running form monitoring.

A local running club was approached to recruit athletes. In the end four male runners participated. The inclusion criteria are as follows:

- Athletes had to be middle- to long distance runners.
- All participants had to be older than 18 years of age and be of good health.
- All participants had to be familiar with a running watch and how to use it. No distinction was made between which watch had to be used. The runners used their own watch of preference. It so happened that all the runners who participated had Garmin watches.

Informed written consent (Appendix B) was obtained for each athlete who participated in the study. Because it is single-subject research and the results were not generalised to the running population, the sample size was kept small. However, the data set per individual athlete was large. Roughly 12 weeks of data were extracted per athlete. Athletes continued to run, train and race with no intervention from the researcher. After a run, runners synced their data from the device to the fitness application (Garmin Connect) as they normally do. The researcher gained access to their online profile (with written permission from the athlete) and extracted the GPS container files with the required data from the running sessions. This file is accessed via the *Option* icon on the specific activity session and selected as *Download tcx format.* Each run session was visually inspected on Google Earth to classify the running activity based on the location. The on-line application, Garmin Connect, has this functionality built-in so it was easy to inspect the route and classify it as road, track or trail. The container files were downloaded and stored on a laptop. This data file include the GPS location (latitude and longitude), the altitude, a date-time stamp, distances covered, speed, cadence, and heart rate for nearly each second over the duration of the run activity. The GPS container files were named per the following convention: *athlete ID - run type - date of activity - activity number.* The activity number is auto-generated from the on-line application. The athlete ID is the unique identification number given to each athlete to ensure anonymity of the data. The type of running activity can be:

- rr for road running;
- tt for track training;
- rrace for a road race;
- tr for trail running;
- trace for trail race.

The date of the activity was captured in the format yyyy-mm-dd. For example, the file

7rr2018041124578398

is the GPS container file for athlete number 7 on 11 April 2018 for a road run session. The last eight digits constitute the file name as the activity number originally generated by the application. This naming convention was used in order to manage the data per athlete, run type and dates. This constructed file name was also imported together with the data set for the sake of trace-ability and data management control.

3.2 Data organisation

A database in *SQLite* was built in order to manage and store the data after extraction. The statistical programming language *R* was the software tool of choice to perform the analysis and construct the models. The *xml* package in *R* was used to extract the data from downloaded GPS container files. Data from the container files was sorted and converted to *R* data frames, which were then written to data tables in the database for storage using the package *RSqlite*. First a temporary staging table is created in the data base that contains the newly extracted data from the container file. The data from the staging table is inserted into the main table in the data base. The staging table is then dropped from the data base, in order to generate the new staging table with new data from the next container file.

Each athlete received a unique identification key (an integer between 1 and 5) as the variable *athlete id*. The date-time stamp was converted to a UNIX numerical time stamp with origin date as 31 December 1989 23:59:59. This numerical time stamp represents the number of seconds passed since the date of origin. All date-time stamps are extracted on GMT zone from the application and were converted to the South African time zone in *R* using the *POSIXct* function. Each line of data received a unique primary key, which is a concatenation of the *athlete id* and the numerical time stamp. Each imported run activity received a unique identification key as the concatenation from the athlete id, the first numerical time stamp and a portion of the first latitude co-ordinate where the athlete started the run. This is referred to as the *session id* and is a unique combination of the athlete and when and where the session started. An athlete can be at the same location on different times, but not at different places at the same time. A categorical variable was added to classify the run session as road running, trail running, track training or a race with the same abbreviations used for the file name.

At each import the raw data from the *xml* file had to be checked for inconsistencies in headings or column variables. A standard was set to import only relevant data from each GPS container file. The raw data from the GPS container files were cleaned from missing values (which are casted in *R* as *NA*) before the data was written into the database. A view was constructed in the database to organise and package only the required data to be read back into *R* for the purposes of analysis and statistical modeling. This view grouped the data to the unique time stamp level to eliminate any duplication of rows or entries. Data in this view was ordered by the athlete ID and date.

3.3 Variable calculations

The organised view in the database was accessed via the *RSqlite* package and the data was imported back into *R*. Although speed forms part of the extracted container file, it was decided to calculate speed from the changes in distances and time in the container file. There are two reasons for this, namely:

- Abnormal high speeds were observed in the raw data, in excess of 10 m/s. This occurs when the device's signal jumps between satellites when the signal is poor or momentarily lost.
- The speed algorithm on the device does not respond instantaneously when the runner becomes stationary. Instead it seems that a rolling average is calculated and it slowly decreases to near 0 m/s.

It would therefore be an inaccurate estimate to use this speed in the container file to determine the athlete's moving status. The following variables were calculated for each line per unique run activity, i.e. per unique GPS container file. The subscript j represents one instance, or one data point in time.

- Cumulative time passed in seconds.

$$T_{cum_{j+1}} = T_{cum_j} + (time_{j+1} - time_j) \tag{3.1}$$

- Speed (m/s)

$$v = \frac{\delta\ distance}{\delta\ time} \tag{3.2}$$

- Pace (min/km)
$$p = \frac{1}{v} \times \frac{1000}{60} \tag{3.3}$$
Because pace is the inverse of speed, the value decreases as the runner is running faster.

- Elevation (meters) per instance.
$$\delta \ Elevation_j = altitude_j - altitude_{j-1} \tag{3.4}$$

- Net elevation (meters).
$$Net \ Elevation = altitude_{start} - altitude_{end} \tag{3.5}$$

- Gradient of the slope (in percentage) at the j^{th} instance:
$$g_j = 100 \times \frac{altitude_j - altitude_{j-1}}{distance_j - distance_{j-1}} \tag{3.6}$$

- Move type (i.e. walking or running), which is a categorical variable. A theoretical threshold was set to 2 m/s as the transition speed from walking to running (Neptune and Sasaki, 2005). Any speed below this threshold was considered walking, and any speed greater than or equal to this threshold was considered running.
$$\text{move type} = \begin{cases} \text{running} & : speed \geq 2 \\ \text{walking} & : speed < 2 \end{cases} \tag{3.7}$$

These variables were appended as columns to the original imported data and saved to R-object files for retrieval during data analyses and modeling. The pseudocode for the calculations code is shown in Algorithm 1 and uses athlete 3 as an example.

3.4 Descriptive analysis

Data from the R-object files was extracted and subset to include only running data, i.e. all walking instances and too low cadence values were excluded from further analysis. Cadence must be greater than 75 *cycles/minute*. This was an arbitrary value chosen by the researcher and loosely based on the application's criteria for running cadence. Any entries where the speed was above 7.9 m/s were also excluded. This is based on the exlsusion criteria of Kosmidis and Passfield (2015). This is world-record pace for the

Algorithm 1: Variable Calculations

```
Connect to the database
Extract only the relevant data per athlete from the view per athlete ID and rename as data frame rd31
Set up the unique id to subset the data frame into separate running sessions
Set an empty data frame to populate later as rd30
Loop through all the entries in the extracted data frame and perform the calculation for the variables
for each activity i in extracted data frame rd31 do
    Subset a new data frame per run activity as rd32
    Create empty vectors for the following variables:
    speed
    elevation
    time difference
    cumulative time in seconds
    for each instance j in data frame rd32 do
        speed = $(distance_{j+1} - distance_j)$ / $(timestamp_{j+1} - timestampj)$
        elevation = $altitude_{j+1} - altitude_j$
        time difference in seconds = $time\ stamp_{j+1} - time\ stamp_j$
        cumulative time passed in seconds = cumulative time up to instance $j$ + $(time\ stamp_{j+1} - time\ stamp_j)$
    end
    Test for any NA values in the vectors and remove them
    Append the calculated variables to the subset data frame rd32
    Find pace by taking the inverse of speed across the data frame's speed column
    Create empty vector as move type
    for each second j in the data frame rd32 do
        Test running speed against transition threshold
        if $speed \leq threshold$ then
            Assign variable as walking
        else
            Assign variable as running
        end
        Append the assigned variable to the vector move type
    end
    Append the vector move type to the data frame rd32
    Populate the empty data frame rd30 each time by binding the rows from the newly subset data frame rd32
end
Disconnect from the database
Save the completed data frame rd30 as an R-object file
```

men's 800m and is reasonably considered as being outside of the participating athletes' capabilities. The original time-series, histograms and density plots were constructed from the subset data for the following key variables:

- Cadence
- Pace

The time-series, concentration (or density) and distribution curves were constructed for a single session per athlete in order to illustrate the data transformation from a time-series into a smoothed cumulative distribution curve. The histograms and density plots were constructed from all the data and then subset into the running activity, i.e. road running, road racing, track training, trail running and trail racing. Descriptive statistics for the graphed variables were calculated. These are the means (μ), variances (σ^2), coefficient of variance (CV), ranges, minimum, maximum and the 95% confidence intervals (CI). These values are not necessarily shown in the results, but used to explain observations where necessary.

The analyses on the spread of the data around the mean is a continuation of the descriptive statistics. The third moment around the mean is used to calculate the skewness of the data, or put otherwise, how symmetrical the data is. A positive value implies the data is skewed to the right (it has a long right tail) and a negative value implies skewness to the left (a long left tail). A distribution is considered perfectly symmetrical when the value is equal to 0. The third moment, or the skewness measure, is calculated as follows:

$$\begin{aligned} m_3 &= \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^3}{n-1} \\ m_3' &= \frac{m_3}{s^3} \end{aligned} \tag{3.8}$$

where $\bar{x}$ is the sample mean of the data set $x_1, x_2, ..., x_n$ and m_3' is the standardised skewness measure.

Skewness was specifically used to describe the symmetry of pace and grade per cadence level. This formed part of the holistic behavioural analyses of the data and was the motivation to initiate secondary clean-ups of the data in order to correct the skewness. A boxplot is an effective visualisation technique to evaluate general symmetry and skewness of a data set and is illustrated in Figure 3.1. The Inter-quartile Range (IQR) is captured by the white rectangle with the lower limit (or hinge) representing the 25^{th} percentile and the upper limit (hinge) representing the 75^{th} percentile. The median of the data is presented by the line inside the box. The upper whisker of the boxplot extend up tot the highest value that is smaller than or equal to 1.5 IQR from the upper hinge. The

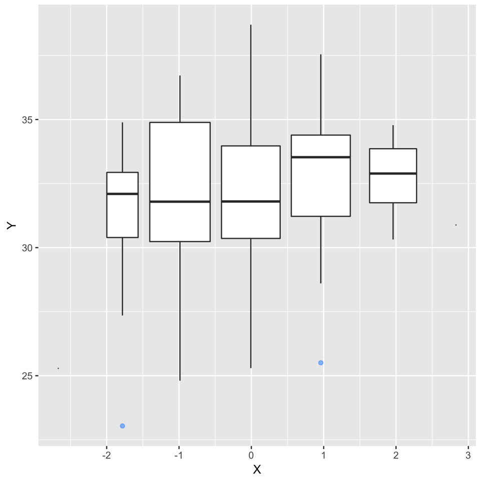


Figure 3.1: Simple boxplot of randomly generated variables

lower whisker of the boxplot extend up to the lowest value that is smaller than or equal to 1.5 IQR from the lower hinge. Any data outliers beyond the end points of the whiskers are plotted as blue points.

3.5 Data clean-up

Secondary clean-ups followed after the analyses on the patterns of the data's skewness. The first clean-up operation involved removal of outliers of pace per cadence level. For each cadence level, an upper and a lower limit were calculated and any pace and gradient values below or above those values for each cadence level were removed from the data set.

$$\begin{aligned} Q_{up} &= Q_{p75} + (1.5 \times IQR_p) \\ Q_{lp} &= Q_{p25} - (1.5 \times IQR_p) \end{aligned} \tag{3.9}$$

where IQR_p is the inter-quartile range of the pace per cadence level.

$$\begin{aligned} Q_{ug} &= Q_{g75} + (1.5 \times IQR_g) \\ Q_{lg} &= Q_{g25} - (1.5 \times IQR_g) \end{aligned} \tag{3.10}$$

where IQR_g is the inter-quartile range of the grade per cadence level. The pseudo-code for the removal of the outliers are shown in Algorithms 2 and 3.

Algorithm 2: Removal of outliers of pace per cadence level

```
Set up the vector of cadence levels (each value in the cadence range becomes a level)
Set-up a clean data frame to populate with the outcomes for each level i in the cadence level vector
do
    subset the data frame to the cadence level i
    set up the empty vectors to to test for the upper and lower limits of pace
    for each instance j in data set do
        test the pace against the $Q_{up}$ for cadence level i
        if $pace_j > Q_{up}$ then
            assign a value of 0
        else
            assign a value of 1
        end
        test the pace against the $Q_{lp}$ for cadence level i
        if $pace_j < Q_{lp}$ then
            assign a value of 0
        else
            assign a value of 1
        end
    end
    Append the outcome per instance j to the data set
    Append each completed subset of data set the final data frame
end
```

Algorithm 3: Removal of outliers of grade per cadence level

```
Set up the vector of cadence levels (each value in the cadence range becomes a level)
Set-up a clean data frame to populate with the outcomes for each level i in the cadence level vector
do
    subset the data frame to the cadence level i
    set up the empty vectors to to test for the upper and lower limits of grade
    for each instance j in data set do
        test the grade against the $Q_{ug}$ for cadence level i
        if $grade_j > Q_{up}$ then
            assign a value of 0
        else
            assign a value of 1
        end
        test the grade against the $Q_{lg}$ for cadence level i
        if $grade_j < Q_{lp}$ then
            assign a value of 0
        else
            assign a value of 1
        end
    end
    Append the outcome per instance j to the data set
    Append each completed subset of data set the final data frame
end
```

3.6 The transformation: time-series to distribution curve

An illustration of how the data was aggregated from a simple time-series to the final distribution is shown in Figure 3.2. Cadence for each captured instance in the GPS container file is plotted against the time duration of the session. A histogram shows the distribution of the same time-series plot. It is clear that the distribution is right-skewed. The outliers to the right correspond to the scattered instances plotted in the time series plot where the points deviate far from the mean line (roughly before the 25^{th} minute and between minutes 32.5 and 35). There is one instance after the 55^{th} minute that is also considered as an outlier.. The distribution plot now shows the accumulated time (in percentages) that the athlete spends at or above each cadence level. The lowest recorded cadence is 79 and in the distribution curve the athlete has spent all of his time (100%) at or above 79 cycles/minute. The line starts dipping between 84 and 85 cycles/minute, which is the point in the histogram where the height of the bars start to increase. This implies that the athlete has spent just above 95% of his time at a cadence of 85 or higher. The athlete has spent roughly 16% at a cadence of 90 or higher. The line between the cadence levels changes it slope throughout the range. The changes is slope is an important concept, as it may be indicative of how difficult it might be for an athlete to run at a higher cadence. A more gradual slope may imply the athlete is comfortable to move to the next cadence level, where a sharp decline may mean the athlete prefers the previous level and therefor a great distance in percentages is observed between the lower and higher cadence level. Moving from a preferred level to the next may take much more effort when the slope is steep than when the slope is gradual. At the opposite end of the cadence range the slopes of the lines change more abruptly than at the higher cadence levels. The athlete has spent only a very small percentage at a cadence of 91 or higher. The highest recorded cadence is 104, which is virtually at 0%.

3.7 Regression modeling

Regression analysis is the collection of statistical tools used to explore and describe non-deterministic relationships between variables. The basic premise of regression analysis is to explain how much of variation of the response variable Y, also referred to as the dependent variable can be explained by the predictor variable/s x_i (also called the co-variates, the independent variable, the regressor or the estimator). It produces an empirical function which mathematically describes the relationship between the dependent variable Y and the independent variables x_i to some degree of certainty (Montgomery et al., 2006).

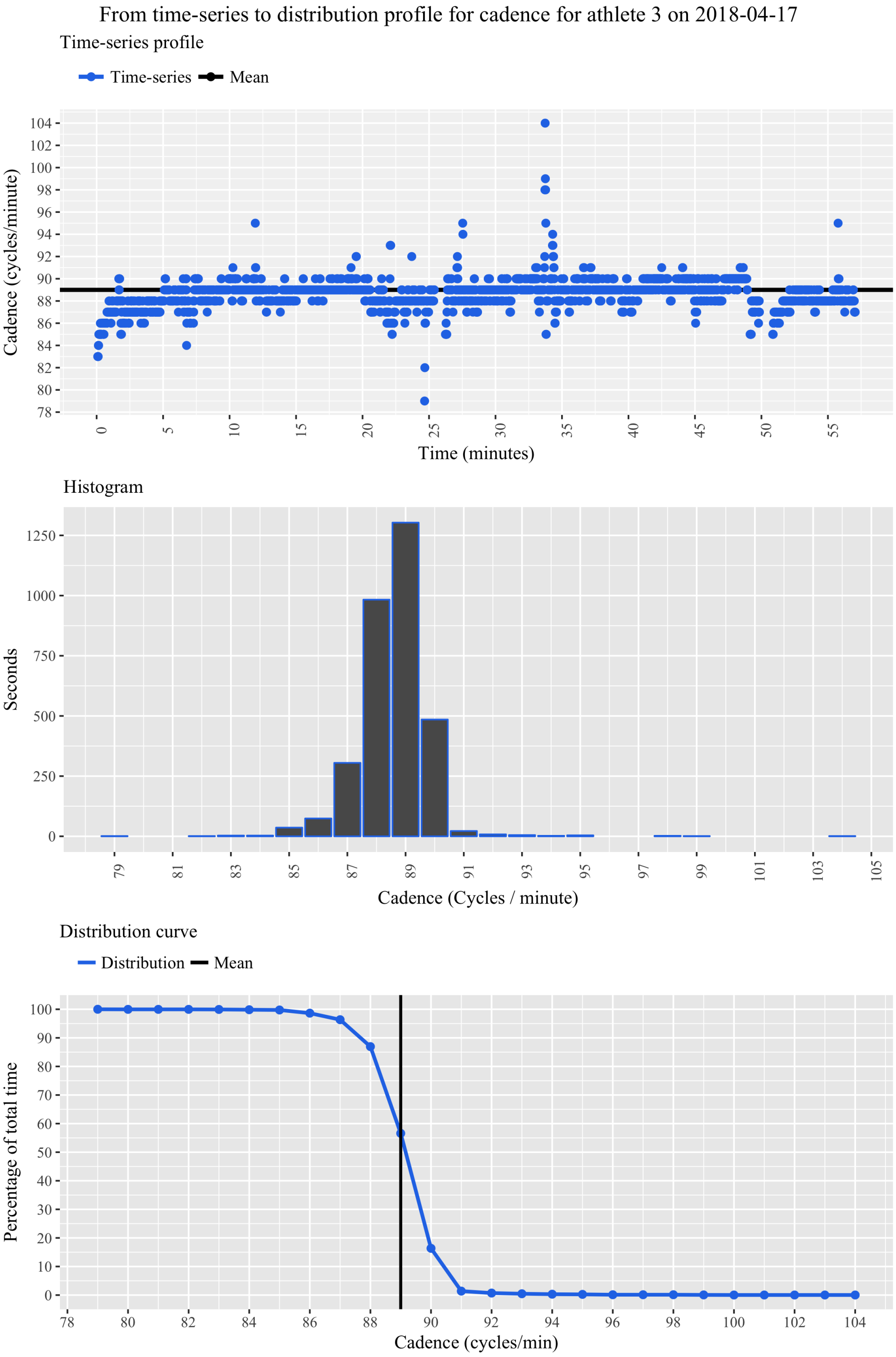


Figure 3.2: The transformation of cadence data from a time-series to the distribution curve for a single session.

3.7.1 Regression mathematics: parametric models

A regression analysis is a method to determine the expected value of a response variable y as a function of a set of predictor variables $x_1, x_2, x_3, ..., x_n$ with parameters $\beta_1, \beta_2, \beta_3, ..., \beta_n$ (Montgomery and Runger, 2011).

General Linear Modeling

In the standard linear form the regression model is presented in Equation 3.11:

$$E(y \mid x_1, x_2, x_3, ..., x_n) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{3.11}$$

The Simple Linear Regression (SLR) will be used to first provide the basics behind parametric regression modeling. In SLR the relationship is assumed to be linear, i.e. that the observations in a scatter plot more or less fall around a straight line and there is only a single regressor (Montgomery et al., 2006). The basic model function for the SLR is shown in Equation 3.12 where Y is the response variable, β_0 is the intercept, β_1 is the slope and x is the regressor.

$$Y = \beta_0 + \beta x \tag{3.12}$$

However, the data points are random and do not all fall exactly on the straight line. The model in 3.12 must be adjusted to include an error term that accounts for the model's inability to fit all the data points exactly. The error term, ε, is added to the equation in 3.12 as a statistical, random variable to make up for the difference between the Y-value on the fitted line and the actual observed value at the data point i. Equation 3.12 is adjusted in 3.13 to include the error term ε:

$$Y = \beta_0 + \beta x + \varepsilon \tag{3.13}$$

By fixing the value of x the properties of Y are determined by the random component ε. Let $\varepsilon \sim N(0, \sigma^2)$, then it follows that

$$E(y \mid x) = \mu_{y|x} = E(\beta_0 + \beta x + \varepsilon) = \beta_0 + \beta x \tag{3.14}$$

and

$$Var(y \mid x) = \sigma^2_{y|x} = Var(\beta_0 + \beta x + \varepsilon) = \sigma^2 \tag{3.15}$$

where the variability in Y at a certain x is determined by the variance of the error term ε. The regressor x is controlled by the analyst under the assumption that the measurement has negligible error, while the response variable Y is a random variable. Thus there exists a probability distribution for Y at each value of x, with the mean of the distribution as

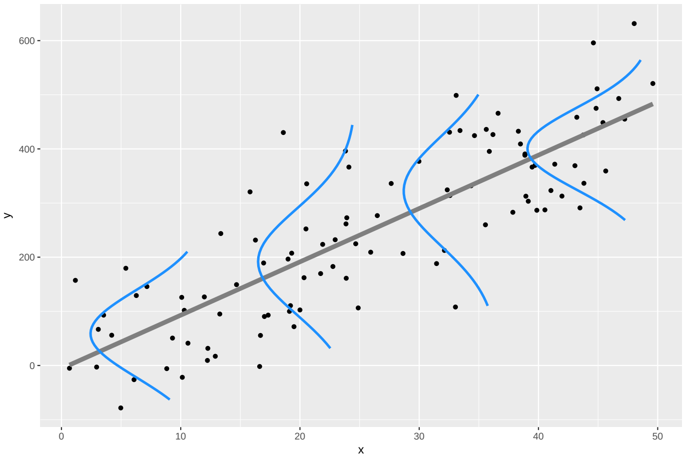


Figure 3.3: The straight linear regression line with the distribution of y_i

3.14 and the variance as 3.15. The variance remains constant for each fitted y across all x. This is the homogeneity of the data (Otto, 2012). An illustration is adapted from Montgomery and Runger (2011) in Figure 3.3. The straight line passes through the data with the Gaussian density curve plotted over the line for various values of x.

The coefficients β_0 and β must be estimated. This is done using the method of *least squares*. The coefficients are chosen such that the sum of the squares of the differences between the fitted y and the observed value y_i is minimised (the right hand side of 3.16).

$$\begin{aligned} S(\beta_0, \beta_1) &= \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i)^2 \\ &= \sum_{i=1}^{n} (y_i - y)^2 \end{aligned} \tag{3.16}$$

Let the estimators for the coefficients be defined as $\beta_0 = \hat{\beta}_0$ and $\beta_1 = \hat{\beta}_1$. To minimise 3.16, the partial derivatives of $S(\beta_0, \beta_1)$ for a given $\hat{\beta}_0$, $\hat{\beta}_1$ must be equal to 0:

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial S}{\partial \beta_1} &= -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{aligned} \tag{3.17}$$

Equation 3.17 can be re-written as:

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2 &= \sum_{i=1}^{n} y_i x_i \end{aligned} \tag{3.18}$$

Solving the equations in 3.18 simultaneously yields the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, expressed in terms of $\overline{x}, \overline{y}$ and n (the number of data points) in 3.19:

$$\begin{aligned} \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \\ \hat{\beta}_1 &= \frac{\sum\limits_{i=1}^{n} y_i x_i - \dfrac{\left(\sum\limits_{i=1}^{n} y_i\right)\left(\sum\limits_{i=1}^{n} x_i\right)}{n}}{\left(\sum\limits_{1=1}^{n} x_i^2 - \dfrac{\left(\sum\limits_{i=1}^{n} x_i\right)^2}{n}\right)} \end{aligned} \tag{3.19}$$

where $\overline{y} = ({}^1\!/\!{}_n) \sum\limits_{i=1}^{n} y_i$ and $\overline{x} = ({}^1\!/\!{}_n) \sum\limits_{i=1}^{n} x_i$. $\hat{\beta}_0$ and $\hat{\beta}_1$ are the *least squares estimators* of the intercept and slope respectively (Montgomery et al., 2006). Equation 3.14 can now be re-written as the fitted SLR model, where $\hat{y}$ is the point estimate for a mean value of y for a given x, i.e. $\mu_{y|x}$:

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ &\text{and for a particular value at point } i : \\ \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \end{aligned} \tag{3.20}$$

When there are more than one predictor variable, the model is a Multiple Linear Regression (MLR) and takes on the form of Equation 3.11. The coefficients are also estimated using the same basic principles as for the SLR.

The difference between the actual observations y_i and their point estimates $\hat{y}_i$ is called the residual and is expressed in 3.21:

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i) \end{aligned} \tag{3.21}$$

The residuals are important when evaluating model adequacy, or its ability to represent the data with the least amount of uncertainty. Model adequacy is discussed further in a subsequent section in this chapter. The behaviour of the residuals provide information on whether the underlying assumptions for a linear model are met. If the assumptions are not met based on the residual analysis, the decision to model the data under a linear relationship needs to be re-visited.

Generalised Linear Modeling

A basic assumption is the homogeneity of the model, i.e. that the variance for all fitted y-values across all x remains constant. However in some cases this assumption is violated and the variance is functionally related to the mean y. In such cases the response variable is transformed in order to stabilise the variance. Depending on the functional relationship between y and the variance, transformations include taking the square root, arc-sin, logarithm, recirpocal square root or just the reciprocate of y as y' in 3.22 (Montgomery et al., 2006):

$$\begin{aligned} y' &= \sqrt{y} \\ y' &= \sin^{-1}(\sqrt{y}) \\ y' &= \log(y) \\ y' &= y^{-1/2} \\ y' &= y^{-1} \end{aligned} \tag{3.22}$$

This leads to the development of the GLM. In a GLM the transformation may may be decided upon by specifying a function to fit the data if there is a known theoretical relationship, or the transformation may be selected based on the trend observed in a scatter plot of the data or the residual plots. Non-linear relationships that can be successfully transformed are referred to as being intrinsically linear or transformably linear (Montgomery and Runger, 2011). Examples of functions which can be transformed are: The exponential function

$$y = \beta_0 e^{\beta_1 x} \varepsilon \tag{3.23}$$

can be transformed by taking the natural logarithm of y:

$$ln\ y = ln\beta_0 + \beta_1 x + ln\varepsilon \tag{3.24}$$

which yields the transformed linear model

$$y' = \beta_0' + \beta_1 x + \varepsilon' \tag{3.25}$$

Similarly a reciprocal transformation may be applied to the function

$$y = \beta_0 + \beta_1(\frac{1}{x}) + \varepsilon \tag{3.26}$$

where $x' = \frac{1}{x}$ which yields

$$y = \beta_0 + \beta_1 x' + \varepsilon \tag{3.27}$$

In each case the residuals must be analysed after the transformation to assess whether the transformation is a good fit to the data. It is important to remember that the predicted values, the coefficients and the residuals are now in the transformed scale and the transformed function will have to be converted back to the original scale. The residual analysis is discussed further under the section *Model adequacy.*

Otto (2012) provides a basic note on how to recognise an intrinsically linear function: *it is the linear combination of a series of parameters where no parameter appears as an exponent or is multiplied by another parameter.* Basic forms of non-linear functions that cannot be linearalised are shown in 3.28:

$$\begin{aligned} y &= {x_1}^{x_2} \\ y &= {\beta_1}^{x_1 x_2} \\ y &= x_1 x_2 \end{aligned} \tag{3.28}$$

Transformations have proved to be useful to present non-linear data in a linear form for the purposes of simplification and explanation. It is also useful when an asymptote must be enforced on the prediction model for practicality reasons.

Polynomial regression models

The polynomial regression models are included in parametric linear regression modeling (Montgomery et al., 2006). They are useful when the data is curvilinear in shape and does not fit a priori mechanistic model. For example the second degree polynomial in one variable,

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{3.29}$$

describes a quadratic relationship between x and y, where β_1 is the linear effect parameter and β_2 is the quadratic effect parameter. β_0 is the mean of y where $x = 0$ if 0 is included in the range of x, otherwise it holds no meaning. The general expression for the k^{th} polynome in one variable is given as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \varepsilon \tag{3.30}$$

When $x_j = x^j,\ j \in \{1, 2, ..., k\}$, Equation 3.30 becomes a MLR in k regressors for $x_1, x_2, ..., x_k$.

The R_a^2-values are used to compare polynomial models, as it punishes the model when introducing more higher order terms for a better fit. However, there are cases when a low order polynomial is a poor fit to the data and the introduction of higher order terms does not improve the fit. This situation is apparent when the residual sum of squares fail to stabilise or the residual plots persist in showing unexplained structures. The underlying problem to this is that the function behaves differently in different ranges of x. Non-paramteric models are fitted to solve this problem (Montgomery et al., 2006) and is discussed in section *Regression mathematics: non-parametric models.*

3.7.2 Interaction models

An ICM is an extension of the GLM. Interaction effects consists of interference and reinforcement effects. The predictor variables in an interaction model are not additive but may either reinforce each others effects on the response variable or interfere with one another. Their interactions are modeled in a regression model by adding a cross-product or interaction terms in the linear model (Neter et al., 1988). Let x_ik be the i^{th} observation for regressor k and let β_k be the coefficient for regressor k. If there are two regressors (then $k = 2$), the linear model becomes:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \epsilon_i \tag{3.31}$$

where the third term $\beta_{12} x_{i1} x_{i2}$ is the interaction term for the two regressors. The effect of x_1 depends on the effect of x_2 and vice-verse and therefore the effects of x_1 and x_2 are no longer additive. The interaction terms can be a combination of categorical and continuous variables. Let x_1 be a continuous variable. When x_2 is a categorical variable (such as pre-determine levels), then x_2 becomes an indicator variable. It is easier to think of x_2 as an integer with binary values $0, 1$ where:

$$x_2 = \begin{cases} 1 & \text{the categorical variable is present} \\ 0 & \text{the categorical variable is absent} \end{cases} \tag{3.32}$$

When x_2 is absent, the terms involving x_2 in Equation 3.31 will become 0 and subsequently the model is reduced to include only x_1. In the case where x_2 is present and is assigned the value of 1, the effect of x_1 on y increases (or decreases) by the value of β_{12}, or put otherwise, the slope β_1 increases (or decreases) by the value of β_{12}. Equation 3.31 becomes:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \epsilon_i \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) x_{i1} + \epsilon_i \end{aligned} \tag{3.33}$$

The response function becomes (when the categorical level evaluates to true or is present, i.e. $x_2 = 1$):

$$\hat{Y} = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) x_1 \tag{3.34}$$

In the case where x_2 is a continuous variable, the coefficient for the interaction term is explained as: the effect of x_1 on y increases by β_{12} for a one unit change in x_2. Equation 3.31 remains as is and the parameters are substituted into the model. The response function becomes:

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \tag{3.35}$$

In the case of a log transform on response Y, the expected or predicted values $\hat{Y^*}$ must be transformed back to the response scale. The coefficients provided in the output from the linear modeling function, *lm*, in *R* are given in the linear form. The exponent is taken on either side of the response function to return the fitted values to the response scale:

$$\begin{aligned} \hat{Y^*} &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \text{ (log-scale)} \\ e^{\hat{Y^*}} = \hat{Y} &= e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2} \end{aligned} \tag{3.36}$$

3.7.3 Regression mathematics: non-parametric models

Sometimes the relationship between y and x is non-linear, there exists no preceding knowledge of the relationships and there is no mechanistic mathematical model to explain y as a function of x. The analyst may also not necessarily be interested in the function itself, but want to study a relationship between the data. Empirical, non-parametric models that make use of piece-wise functions address this problem.

Regression splines

One solution is to segment the range of x and approximate a different polynomial curve in each segment. This practice is called piece-wise polynomial fitting, with the piece-wise polynomials referred to as splines of order k. Burden et al. (2016) provides a simplified real-world analog of the spline. The word "spline" originates from same word as splint, which can be though of as a flexible piece of wood used to join two boards. The word was later used to refer to a long flexible metal strip that could be used to draw a continuous

smoothed curve passing through all the data points and trace the curve. A more formal definition of a spline is as follows:
A spline is defined by polynomials of order k on sub-intervals i of the data range x, segmented by a series of h knots denoted as t_i where $t_1 < t_2 < ... < t_h$ such that the pieces are joined smoothly at the knots (Montgomery et al., 2006). A spline $S(x)$ of the order k is presented as a power series:

$$S(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \sum_{i=1}^{h} \beta_i (x - t_i)^k{}_+$$

where

$$(x - t_i)_+ = \begin{cases} (x - t_i) & \text{if } x - t_i > 0 \\ 0 & \text{if } x - t_i \leq 0 \end{cases} \tag{3.37}$$

The most used piece-wise polynomial approximation involves cubic polynomials between pairs of nodes or knots and is referred to as cubic spline interpolation. The cubic spline is analysed here first for n points on $n - 1$ intervals , i.e. the spline must go through all the data points (Burden et al., 2016). Let f be a function defined on interval $[a, b]$ and a set of points $a = x_0 < x_1 < ... < x_n = b$. S is a cubic spline interpolant for f with a set of conditions:

1. $S(x)$ is a cubic polynomial denoted $S_i(x)$ for $x \in [x_i, x_{i+1}]$ for each $i = 0, 1, ..., n-1$
2. The cubic spline must go through all the data points:

$$S_i(x_i) = f(x_i) \text{ and } S_i(x_{i+1}) = f(x_{i+1}) \text{ for each } i = 0,\ 1, ...\ n - 1 \tag{3.38}$$

3. There are matching conditions for each $i = 0, 1, ..., n - 2$. This ensures that the y-values at x_i are the same for the two joining functions.

$$S_{i+1}(x_{i+1}) = S_i(x_{i+1}) \tag{3.39}$$

4. Smoothing conditions for each $i = 0, 1, ...i = n - 2$ ensures continuity of the curve at the joining points:

$$\begin{aligned} S'_{i+1}(x_{i+1}) &= S'_i(x_{i+1}) \\ S''_{i+1}(x_{i+1}) &= S''_i x(_{i+1}) \end{aligned} \tag{3.40}$$

5. Boundary conditions at x_0 and x_n:

$$\begin{aligned} S''(x_0) = S''(x_n) = 0 &\text{ for a natural cubic or boundary free} \\ S'(x_0) = f'(x_0) \text{ and } S'(x_n) = f'(x_n) &\text{ for a clamped boundary} \end{aligned} \tag{3.41}$$

The basic cubic expansion for each interval's S_i is:

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \tag{3.42}$$

From 3.42, there are four unknown coefficients (a_i, b_i, c_i, d_i) to solve per interval, that is a total of $4 \times (n - 1)$ unknowns. The cubic spline must go through all the data points as in 3.38, enforcing n constraints. There are no matching points at the start and at the end of the curve, therefore the matching and smoothing constraints only apply on $n - 2$ points. The constraint in 3.39 provides $(n - 2)$ constraints across the interval $[a, b]$. The constraints in 3.40 provide $2(n - 2)$ constraints across the interval $[a, b]$. Constraint 3.41 provide another two constraints at the boundaries, depending on which condition the modeler chooses. The total number of constraints are $n + (n - 2) + 2(n - 2) + 2 = 4(n - 1)$. The number of unknowns and the number of constraints are the same. The coefficients a_i, b_i, c_i, d_i are solved for each i^{th} point on $[a, b]$ using a system of equations and enforcing the matching, smoothing and boundary conditions. Two adjacent pairs of coordinate points, namely $\{(x_i, f(x_i)), (x_{i+1}, f(x_{i+1}))\}$ are needed to set up the system of linear equations to solve for the coefficients for the interval $[x_i, x_{i+1}]$:

$$\begin{aligned} S_i(x) &= a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \\ S_{i+1}(x) &= a_{i+1} + b_{i+1}(x - x_{i+1}) + c_{i+1}(x - x_{i+1})^2 + d_{i+1}(x - x_{i+1})^3 \end{aligned} \tag{3.43}$$

A cubic spline $S_i(x)$ is then defined for each coordinate pair of $(x_i, f(x_i))$ within the range $[x_i, x_{i+1}]$. This method is repeated until each coefficient a_i, b_i, c_i, d_i have been solved.

When the natural boundary conditions are used, the spline adopts the shape of a long flexible rod when forced to go through all the data points $\{(x_0, f(x_0)), (x_1, f(x_1)), ..., (x_n, f(x_n))\}$. The spline extends linearly in the regions $x \leq x_0$ and $x \geq x_n$. In the case of a clamped spline, the end points of the flexible spline is fixed. It forces the spline to follow a specified direction at each endpoint. This becomes important when two spline functions must match at the endpoints. The fixing of the endpoints is accomplished mathematically by assigning values to the derivative of the curve at the endpoint (Burden et al., 2016).

Returning to piece-wise splines, the range of x is subdivided into segments or intervals that contain more than one data point (Montgomery et al., 2006). The same constraints as explained above are applicable to the spline at the knots t:

1. The function values at the knots are the same, as per 3.39.
2. The first $k-1$ derivatives agree at the knot, as per 3.40. For a piece-wise cubic spline, this implies that both $S'(x=t)$ and $S''(x=t)$ have the same values for the joining functions S from either side of the knot t.

The cubic spline with h knots at data points $t_1 < t_2 < ... < t_h$ with continuous first and second order derivatives can be defined as the smoothed function $S(x)$ over the range of x:

$$E(y) = S(x) = \sum_{k=0}^{3} \beta_{0k} x^k + \sum_{i=1}^{h} \beta_i (x - t_i)^3{}_+ \\ \text{where} \\ (x - t_i)_+ = \begin{cases} (x - t_i) & \text{if } x - t_i > 0 \\ 0 & \text{if } x - t_i \leq 0 \end{cases} \tag{3.44}$$

The constraints for a continuous spline are:

1. Matching condition: The values of the spline from the two joining sub-intervals $i, i+1$ must be the same at the knot t_i.

$$S_i(t_i) = S_{i+1}(t_i) \tag{3.45}$$

2. Smoothing conditions. The first and second order derivatives from the joining sub-intervals $i, i+1$ must be the same value at the knot t_i.

$$S_i'(t_i) = S_{i+1}'(t_i) \\ S_i''(t_i) = S_{i+1}''(t_i) \tag{3.46}$$

From 3.44 the cubic expansion for a single knot t_i on sub-interval i is given as:

$$S_i(x) = \beta_{00} + \beta_{01}x + \beta_{02}x^2 + \beta_{03}x^3 + \beta_i(x - t_i)^3 \\ \text{where } \beta_i(x - t_i)^3 \text{ is expanded as} \\ a_i + b_i x + c_i(x - t_i)^2 + d_i(x - t_i)^3 \tag{3.47}$$

When the location of the knots are known, the parameters are found using *least squares* as per normal linear regression. The number of knots and their position as well as the order of the polynomial are important when considering the flexibility of piece-wise polynomial fitting. The extent of flexibility offered by splines may lead to the over-fitting of data, which is not the primary goal of regression analysis. Wold (1974) (in (Montgomery et al., 2006)) suggests the following to avoid over-fitting of the data:

- That there should be as few knots as possible .
- Each segment must contain at least four data points.
- A maximum of one extreme point per segment, preferably in the center of the segment.
- One point of inflection per segment, preferably close to the knots.

A parameterised piece-wise polynomial may be used to predict future outcomes, however the cubic spline's pieces are only valid for the data range or interval it were specified for.

In simple terms, non-parametric models is made up from estimated piece wise functions from a scatter plot with input variables on the x axis and outcome variable y on the y axis that gets smoothed with scatter plot smoothers at the naught points (i.e. the knots) where the function changes. A simple segmented piece wise function representative of a scatter plot is illustrated in Figure 3.4. Upon visual inspection it may be considered that the data is following some cyclical pattern or have boundary conditions it must adhere to. A single linear regression line representing all the data points would not have been a good fit to the scatter plot at all. From the example in Figure 3.4 and after some more iterations and smoothing at the naught points a non-parametric model may be developed. The piece-wise function will be more representative of the data's shape as Y continues to change for increasing values of x.

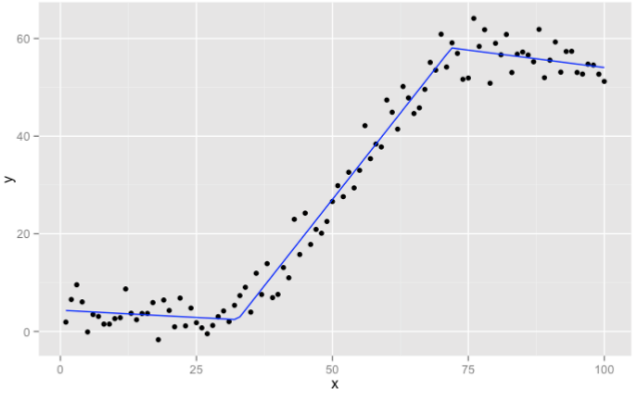


Figure 3.4: Fitted segmented function from scatter plot data (Vaidyanathan, 2012)

Non-parametric regression is closely connected to piece-wise polynomial regression, with the difference that is a model-free basis for estimating a response y over the range of data x. It generates an empirical (or heuristic) model that is good for interpolation but not recommended for extrapolation as the model is only applicable on the range of data on which it was built. Instead the parameters β are replaced by a smoothing matrix S (Montgomery et al., 2006). The predictors are divided into neighbourhoods of x_i. A neighbourhood N_i may be defined as a group of points whose x values are close tot x_i. The size of the neighboorhood is the span or window size, w. The window is defined as the proportion of the total data points within each N. A larger span includes more data points per x_i and the resulting smoother is applied to a greater proportion of the data. The smooth is then less sensitive to variance in the data patterns and may result in a poorer fit. On the other hand a small window contains less data points in the neighbourhood and a smoother is applied to smaller proportions of the data (Montgomery et al., 2006). The analyst must exercise caution when using a small window size as the risk for over-fitting increases. The premise of a smooth is to identify a pattern in the data and not explain each data point (Otto, 2012). The width of the span is thus an important decision when smoothing curves.

Locally Weighted Regression Modeling (LOESS)

The Locally Weighted Regression Modeling (LOESS) is a non-parametric regression method that uses the data points contained in neighboorhood N_i to produce a Weighted Least-Squares (WLS) estimate of the response for the data point x_i. The WLS procedure uses a low order polynomial (usually simple linear regression) or quadratic regression within the neighboorhood. The weights for the WLS section of estimation are based on the distance between the data points in the neighbourhood and the focal point, x_i (Montgomery et al., 2006). A commonly used weighting procedure is the tri-cube weighting function. Let x_0 be the point of interest and $\triangle(x_0)$ be the distance to the furthest point in the neighboorhood. The tri-cube weighting if defined in 3.49 as

$$W\left[\frac{\mid x_0 - x_j \mid}{\triangle(x_0)}\right] \tag{3.48}$$

where

$$W(t) = \begin{cases} (1-t^3)^3 & \text{for } 0 \leq t < 1 \\ 0 \text{ elsewhere} & \end{cases} \tag{3.49}$$

Finally, the LOESS model is presented as $\boldsymbol{y} = \boldsymbol{S}\boldsymbol{x}$, where $\boldsymbol{S}$ is the smoothing matrix generated by the locally weighted regression. The LOESS method was used to graphically model the behaviour of the residuals in the residual plots of the GAM and SCM on the distribution curves for cadence and pace.

Generalised Additive Modeling

The GAM as developed by Hastie and Tibshirani (1986) extends both the linear and the GLM to non-parametric smoothed functions. GAM's are data-driven instead of modeldriven, i.e. the relationship is built on the data. The resulting fitted values for y do not originate from a priori model (Otto, 2012). Whereas the normal linear regression models present the functional relationship between y and $x_1, x_2, ..., x_n$ in some parametric form, the GAM stems from the trend where the dependence of y on $x_1, x_2, ..., x_n$ is modeled in a non-parametric manner (Hastie and Tibshirani, 1986). The basic definition by Hastie and Tibshirani (1986) is that the general linear form $\sum \beta_i X_i$ with parameters β is replaced by the sum of smoothed functions in Equation 3.50, hence *generalised additive models*

$$E(y \mid x_1, x_2, ..., x_n) = s_0 + \sum_{i=1}^{n} s_i x_i \tag{3.50}$$

where $s_i(.)$'s are the smooth functions estimated from a scatter plot smoother. Since each variable is represented as a separate function in 3.50, all predictors but one must be fixed. It is therefore assumed that $E_{s_i}(x_i) = 0$ for all i except the modeled co-variate.

Wood (2006) extends the basic definition in 3.50 as *a linear model with a linear predictor involving a sum of smooth functions of covariates.* The term *linear* refers to the addition of terms whereby $s_i(.)$ becomes a linear model and not the shape of the function. The general structure of the model is broken down as:

$$g(u_i) = \boldsymbol{X_i}^{*}\boldsymbol{\theta} + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}, x_{4i}) + ... \tag{3.51}$$

where

$$\begin{aligned} \mu_i &\equiv E(y_i) \\ y_i &\sim \text{ some exponential distribution} \\ \mathbf{X_i}^{*} &: \text{ a row in the model matrix with paramatric components} \\ \boldsymbol{\theta} &: \text{ corresponding parameter vector} \\ s_i &: \text{ smooth functions} \\ x_k &: \text{ covariates} \end{aligned} \tag{3.52}$$

The specification of the response variable's dependence on x in therefore flexible. The model is specified in terms of the smooth functions s_i instead of detailed parametric relationships between β and x. In this way the sometimes cumbersome and complicated models generated by GLMs and MLR are avoided. The non-parametric model for a single co-variate is presented as $y = s(x) +$ *error* where $s(x)$ is the smoothed function (Hastie and Tibshirani, 1986). A scatter plot smoother is used to estimate $s(x)$. Such smoothers include the following:

- A running mean

- A running median
- A running least squares line, also referred to as the LOESS
- A kernel estimate
- A spline (piece-wise polynomials)

The smoothed functions are estimated one at a time, moving forward in a series of distinct stages using the *local scoring* procedure. The data is divided into windows and scatter plot smoothers are fitted inside the window. The resulting smoothed functions can be employed for data description and estimation or suggest transformations within the range of x values provided. A smooth estimate can be produced for all the co-variates or a linear fit may be enforced for some portion of the data (Hastie and Tibshirani, 1986), as presented in the extended model in 3.51.

For $s(x)$ to become a linear model a base must be chosen (Wood, 2006). The base represents the space of functions of which s is an element. Choosing a base implies the selection of basis functions. Let $b_j(x)$ be the j^{th} basis function for $s(x)$ where $s(x)$ is presented as:

$$s(x) = \sum_{j=1}^{q} b_j(x)\beta_j \tag{3.53}$$

where β is an unknown coefficient. Substituting 3.53 into 3.51 yields the linear model for response variable y.

Two spline functions from the *gam* function in the *mgcv* package in *R* were used in the developed models and will be discussed further, i.e. the cubic regression and thin plate regression splines. A method to define a basis for the cubic spline is to parameterise the spline in terms of its values at the knots. The *gam* function from the *mgcv* package in *R* uses the cardinal basis function, which will be described here. Let $f(x)$ be a cubic spline with k knots at $x_1, ..., x_k$ over the range of x. Let $\beta_j = f(x_j)$ and $\delta_j = f''(x_j)$. For $x_j < x < x_{j+1}$ the spline is written as:

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1} \tag{3.54}$$

where the basis functions $a_j^-, a_j^+, c_j^-, c_j^+$ are defined as:

$$\begin{aligned}
a_j^-(x) &= \frac{x_{j+1} - x}{h_j} \\
a_j^+(x) &= \frac{(x - x_j)}{h_j} \\
c_j^-(x) &= \frac{\frac{(x_{j+1} - x)^3}{h_j} - h_j(x_{j+1} - x)}{6} \\
c_j^+(x) &= \frac{\frac{(x - x_j)^3}{h_j} - h_j(x - x_j)}{6}
\end{aligned} \tag{3.55}$$

where $h_j = x_{j+1} - x_j$. The condition for continuity up to the second derivative at each knot and that the second derivative at the end knots at x_1 and x_k should evaluate to 0 implies:

$$\boldsymbol{B\delta^-} = \boldsymbol{D\beta} \tag{3.56}$$

where $\boldsymbol{\delta^{-1}} = (\delta_2, ..., \delta_{k-1})^T$ (recall that $\delta_1 = \delta_k = 0$) and $\boldsymbol{B, D}$ are non-zero matrix elements defined in 3.57 and 3.58. For $i = 1, ..., k-2$:

$$\begin{aligned}
D_{i,i} &= \frac{1}{h_i} \\
D_{i,i+1} &= -\frac{1}{h_i} - \frac{1}{h_{i+1}} \\
D_{i,i+2} &= \frac{1}{h_{i+1}} \\
B_{i,i} &= \frac{h_i + h_{i+1}}{3}
\end{aligned} \tag{3.57}$$

For $i = 1, ..., k-3$

$$\begin{aligned}
B_{i,i+1} &= \frac{h_{i=1}}{6} \\
B_{i+1,i} &= \frac{hi+1}{6}
\end{aligned} \tag{3.58}$$

Further let $\boldsymbol{F}^- = \boldsymbol{B}^-\boldsymbol{D}$ and

$$\boldsymbol{F} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{F}^- \\ \mathbf{0} \end{bmatrix} \tag{3.59}$$

where $\mathbf{0}$ is a row of zero's, then $\boldsymbol{\delta} = \boldsymbol{F\beta}$. The spline from 3.54 can now be re-written in

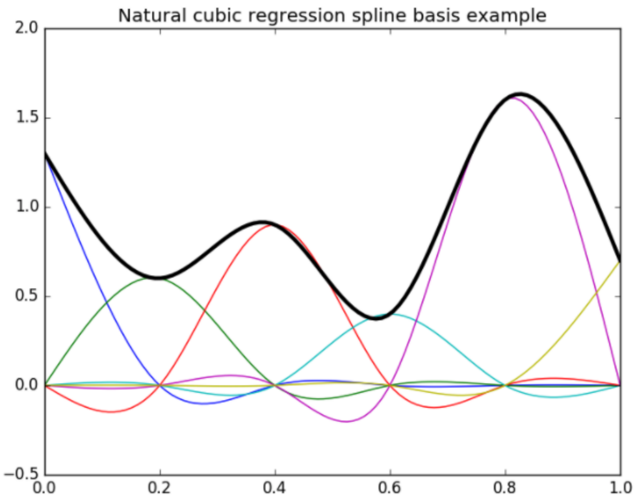


Figure 3.5: The basis functions for a cubic regression spline (Smith, 2015)

terms of β:

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\boldsymbol{F_j\beta} + c_j^+(x)\boldsymbol{F_{j+1}\beta} \\ \text{on } x_j < x < x_{j+1} \tag{3.60}$$

The spline in 3.60 can be re-written as

$$f(x) = \sum_{i=1}^{k} b_i(x)\beta_i \tag{3.61}$$

with $b_i(x)$ the basis functions. Thereby with a given set of x-values at which the spline can be evaluated, β can be mapped to the evaluated spline (Wood, 2006).

Figure 3.5 is borrowed from sample code by Smith (2015) to lend some visual support for the natural cubic regression spline, using the cardinal base functions. Let x_h represent the location of the knots. There are four interior knots, located at $x_h \in (0.2, 0.4, 0.6, 0.8)$ and two knots at the endpoints, namely x_1 and x_6. Six plotted basis functions, $b_i(x)$ shown in colour are constructed throughout the range of $x \in [0, 1]$. Each basis is then multiplied by its corresponding coefficient, β. The combination of the basis functions and their coefficients are summed up to yield the fitted spline in black, which is the GAM, $f(x)$ from Equation 3.61. The blue line to the left peaks at knot x_1, the light green line peaks at x_6 with the other curves each peaking at an internal knot. The coloured lines are all zero at the other knots.

One of the theoretical challenges of the GAM is to decide on the degree of smoothness of the fitted spline. One solution is to select the knots h as well as their spacing, i.e.

changing the dimensions of the basis function. However this may become problematic with uneven knot spacing leading to poor fits. An alternative to altering the base functions' dimensions is to keep the base dimensions fixed and to add a penalty term to the *least squares* objective that will control the "wiggliness" of the model. Resulting splines are referred to as penalised regression splines (Wood, 2006). Rather than minimising only *least squares* as in linear modeling,

$$||\boldsymbol{y} - \boldsymbol{X\beta}||^2 \tag{3.62}$$

a penalty term is added to the *least squares* matrix and the objective is to minimise:

$$||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \int [f''(x)]^2 dx \tag{3.63}$$

where the integrated square of the second derivative of the smooth function f penalises the model when it is too "wiggly". The smoothing parameter λ controls a trade-off between model fit and smoothness. Larger values produce smoother curves and lower values produce more wiggly curves. If $\lambda \rightarrow \infty$ the penalty term dominates and forces $f''(x) = 0$ everywhere in the domain of x. The result is a straight line estimate for f. When $\lambda \rightarrow 0$ the penalty term becomes negligible and the smooth becomes an unpenalised regression spline (Wood, 2006).

The natural cubic spline is regarded by Wood (2006) as being the smoothest interpolator for a set of points $x_i, y_i : i = 1, ..., n$ where $x_i < x_{i+1}$ as well as the spline with optimal interpolation. Wood (2006) provide proof from Green and Silverman (1994) that the cubic spline is the smoothest interpolator. Let $g(x)$ be a natural cubic spline and $f(x)$ be all the functions that are continuous on $x \in [a, b]$, have continuous first derivatives and can interpolate $\{x_i, y_i\}$. Then $g(x)$ is the smoothest function which mimimises the second order derivative of f on $[a, b]$:

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_{a}^{b} f''(x)^2 dx \tag{3.64}$$

Smoothing for a cubic spline g is done by introducing the penalty term to the *leasts squares* and minimising:

$$\sum_{i=1}^{n} \{y_i - g(x_i)\}^2 + \lambda \int g''(x)^2 dx \tag{3.65}$$

The λ in 3.65 is the tuneable parameter to control for the conflicting objectives of the model to match the data and produce a smooth spline. The resulting $g(x)$ is a smoothed cubic spline.

The cubic spline bases do have some disadvantages as listed here (Wood, 2006):

- Knot locations must be chosen, which may introduce some researcher subjectivity.
- The model allows for only one predictor variable.
- The differences between bases as to which is best is not clear.

An approach has been developed to produce knot-free bases that can smooth over multiple predictors, which leads to the discussion on thin plate splines.

The Thin Plate Spline [1]

The TPS is similar to the cubic spline, however the smooth that is now estimated involves more than one co-variate. It generates an estimated smooth function of multiple co-variates, from the noisy observations of the function at certain values of the co-variates (Wood, 2006). It is a powerful tool to examine the combined effect of multiple co-variates on the outcome. A TPS may be compared to the physical bending of a thin metal sheet. In the physical setting the plate is deflected in the z direction, orthogonal to the x, y plane. The plate is “lifted” by a displacement of the x and y coordinates within the plane. Two thin plate splines are required to specify the two-dimensional coordinate transformation (Belongie, 2018). The TPS is the two-dimensional analog of a one-dimensional cubic spline. Rather than a single flat curve, it is presented as a bendable surface with a multi-dimensional appearance and has the functional form

$$U(r) = r^2 \ln r \tag{3.66}$$

Belongie (2018) provides a simplified explanation of the TPS: *With a given set of data points a weighted combination of thin plate splines centered about each data point gives the interpolation function that passes through the points exactly while minimising the so-called “bending energy.”* The bending energy is the penalty imposed to measure the “wiggliness” of the fitted curve and follows the same logic as the penalised cubic spline in one co-variate. For two co-variates it is defined as the double integral over $\mathbb{R}^2$ of the squares of the second derivatives of $f(x, y)$:

$$I\left[f(x,y)\right] = \iint_{0}^{\mathbb{R}^2} (f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2)\, dx\, dy \tag{3.67}$$

---

[1] A modified version of this section was communicated to the IEEE Transactions on Big Data.

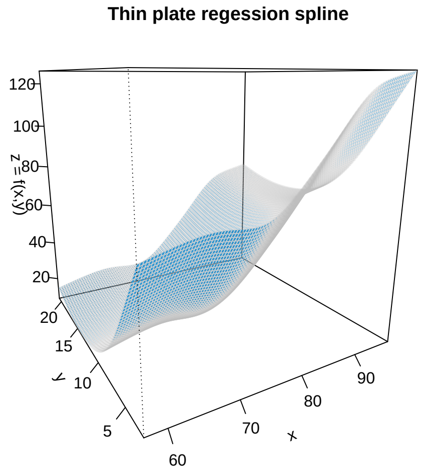


Figure 3.6: The thin plate regression spline for two variables, x, y on z

Let z be the response variable and $g(x, y)$ be a thin plate smooth in the terms of the two co-variates x, y from n observations. For the i^{th} combination of the observed z, x, y the model is expressed as:

$$z_i = g(x_i, y_i) + \varepsilon_i \tag{3.68}$$

Thin plate spline smoothing estimates the function g by finding a function $\hat{f}$ that minimises

$$||\boldsymbol{z} - \boldsymbol{f}||^2 + \lambda I\left[f(x, y)\right] \tag{3.69}$$

where $\boldsymbol{z}$ is the vector of z_i data and $\boldsymbol{f} = \left[f(x), f(y)\right]^T$. Equation 3.69 is the three dimensional analog for 3.65, where $I\left[f(x, y)\right]$ is the double integral taken over the second derivatives of the smoothing function $f(x, y)$ and λ is the smoothing parameter to control the "wiggliness" of the TPS surface. Figure 3.6 provides an illustration of a thin plate spline as a three-dimensional image. Two co-variates are each plotted on their individual axis creating a bi-variate $xy-$ plane. The outcome or response variable is plotted on the $z-$axis continuously across the bi-variate plane and creates a three-dimensional image of the unique observational data z_i, x_i, y_i. The bending of the "plate" is apparent for the different combinations of z_i, x_i, y_i.

One disadvantage of a TPS is the computational cost if constructing a full TPS, as these smoothers have as many unknowns as data (the number of unique predictor combinations). To counter this, the thin plate regression spline truncates the space of the wiggly components of the TPS, thereby simplifying the model and lowering the computational costs.

Shape Constrained Additive Models

The SCM proposed by Pya and Wood (2015) employ shape constrained P-splines as basis functions. P-splines are essentially penalised B-splines. B-splines are commonly used as basis functions for smooths because of their flexibility, smooth interpolation property, and local support. B-splines use several Bézier curves that are joined end to end. A degree n Bézier curve is a smoothed combination of piece-wise polynomial functions and is defined by $n+1$ control points P_i:

$$\begin{aligned} C(t) &= \sum_{i=0}^{n} b_{i,n}(t) P_i \\ b_{i,n}(t) &= \binom{n}{i} t^i (1-t)^{n-i} \end{aligned} \tag{3.70}$$

with $\binom{n}{i} = {}^nC_i$. As an example, the cubic Bézier is expanded as:

$$C(t) = (1-t)^3 P_0 + 3t(1-t)^2 P_1 + 3t^2(1-t) P_2 + t^3 P_3 \tag{3.71}$$

A k degree B-spline defined by $n+1$ control points will consist of $n-k+1$ Bézier curves (Shiach, 2015).

A B-spline of degree k is defined in Patrikalakis et al. (2009) as:

$$C(t) = \sum_{i=0}^{n} N_{i,k}(t) P_i \tag{3.72}$$

where $(P_0, P_1, ...P_n)$ are the control points of the spline and $N_{i,k}(t)$ are the basis functions. The knot vector $\boldsymbol{T}$ is defined as $\boldsymbol{T} = (t_0, t_1, ..., t_m)$ where the points t are non-descending. For a knot vector $\boldsymbol{T}$ the basis functions are determined by the Cox-de Boor recursion formulas. For $k = 1$, $i \in [0, n]$:

$$N_{,i,1}(t) = \begin{cases} 1 & \text{for } t_i < t < t_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

and for $k > 1$, $i \in [0, n]$:

$$N_{i,k}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} N_{i,k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} N_{i+1,k-1}(t) \tag{3.73}$$

The resulting curve from the of the linear combination of the basis functions $N_{i,k}(t)$ within the ranges of the control points $P_0, P_1, ..., P_n$ is called a B-spline. As for the cubic regression splines, B-splines have continuity constraints to ensure smoothness at the knots.

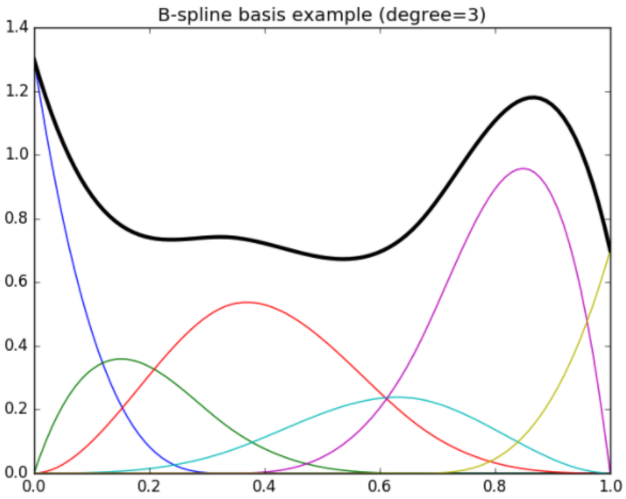


Figure 3.7: The basis functions for a B-spline regression (Smith, 2015)

The function $N_{i,k}(t)$ has C^{k-2} continuity at each knot (Shiach, 2015). These continuities are:

- C^0 continuity: The last point on the first Bézier curve and the first point on the second (adjoining) Bézier curve have the same co-ordinates. That means the curves join in the same spot.
- C^1 continuity: The first derivatives of the curves at the joining points (i.e. the last point on the first curve and the first point on the adjoining curve) are equal.
- C^2 continuity: The second derivatives of the curves at the joining points (i.e. the last point on the first curve and the first point on the adjoining curve) are equal.

Six B-spline basis functions, using cubic splines, are fitted to data in Figure 3.7, borrowed from Smith (2015). Each line is fitted across the range of $t \in [0, 1]$ and represents a basis function N_i. The basis functions are multiplied by their control points to yield one smooth, $C(t)_i$ These smooths are then summed as in 3.72 to produce the black curve $C(t)$.

In a P-spline the coefficients of the B-spline is partially determined by the actual data, with the added influence of a discrete penalty function that forces smoothness but still avoids overfitting. Pya and Wood (2015) refer to Eilers and Marx (1996) for objective function that must be minimised to obtain the P-spline. The penalty is based on higher-order finite differences of the coefficients of adjoining B-splines:

$$S = \sum_{i=1}^{m} \left[y_i - \sum_{j=1}^{n} a_j B_j(x_j) \right]^2 + \lambda \sum_{j=k+1}^{n} (\triangle^k a_j)^2 \tag{3.74}$$

where n is the number of B-splines, m is the number of observations and the term $\triangle^k a_j$ is the difference between adjacent coefficients of the B-spline. The term λ is the tunable parameter to control the trade-off between goodness of fit and overfitting of the data. Pya and Wood (2015) suggested a new non-linear extension of the P-splines from Eilers and Marx (1996) with novel discrete penalties for the SCM, referring to them as the Shape Constrained P-splines (SCOP).

The SCM is presented as the model:

$$g(\mu_i) = \boldsymbol{A\theta} + \sum_j f_j(z_{ij}) + \sum_k m_k(x_{ik}) \tag{3.75}$$

for $y_i \sim exponential family$ with parameters (u_i, ϕ) where:

y_i is a univariate repsonse variable with mean μ_i

g is a known smooth monotonic link function

$\boldsymbol{A}$ is a model matrix

$\boldsymbol{\theta}$ is a vector of unknown parameters

f_j is an unknown smooth function for predictor variable z_j

m_k is an unknown shape constrained smooth function for predictor variable x_k

The shape constraints imposed by the m_k differentiates the SCM from the GAM. In the one-dimensional case let:

$$m(x) = \sum_{j=1}^{q} \gamma_j B_j(x) \tag{3.76}$$

where

q is the number of basis function

B_j are B-spline basis functions of at least second order

γ_j are spline coefficients

(3.77)

For $m'(x) \geq 0$ over an interval $[a, b]$ it is required that $\gamma_j \geq \gamma_{j-1} \, \forall j$. This condition could be imposed by re-parameterising β:

$$\boldsymbol{\gamma} = \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}} \tag{3.78}$$

where

$$\begin{aligned}\boldsymbol{\beta} &= [\beta_1, \beta_2, ..., \beta_q]^T \\ \tilde{\boldsymbol{\beta}} &= [\beta_1, exp(\beta_2), ..., exp(\beta_q)]^T\end{aligned} \tag{3.79}$$

and

$$\Sigma_{ij} = \begin{cases} 0 & : i < j \\ 1 & : i \geq j \end{cases} \tag{3.80}$$

Let $\boldsymbol{m} = [m(x_1), m(x_2), ..., m(x_n)]^T$ be the vector of m values at observed data points x_i and $\boldsymbol{X}$ be the matrix so that $X_{ij} = B_j(x_i)$. Then

$$\boldsymbol{m} = \boldsymbol{X}\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}} \tag{3.81}$$

where $\boldsymbol{X}$ is a $n \times q$ matrix. The monotonically increasing smooth can be extended to other monotonic functions, such as monotonic decreasing, concave, convex or a combination of these. The form of $\boldsymbol{\Sigma}$ and $\boldsymbol{D}$ provides the differences in the shape constraint. For monotonic increasing:

$$\Sigma_{ij} = \begin{cases} 0 & \text{if } i < j \\ 1 & \text{if } i \geq j \end{cases}$$

$$\begin{aligned} D_{i,i+1} &= -D_{i,i+2} = 1 \text{ for } i \in [1, q-2] \\ D_{ij} &= 0 \text{ otherwise} \end{aligned} \tag{3.82}$$

For monotonic decreasing:

$$\Sigma_{ij} = \begin{cases} 0 & \text{if } i < j \\ 1 & \text{if } j = 1 \, i \geq 1 \\ -1 & \text{if } j \geq 2, \, i \geq j \end{cases}$$

$$\begin{aligned} D_{i,i+1} &= -D_{i,i+2} = 1 \text{ for } i \in [1, q-2] \\ D_{ij} &= 0 \text{ otherwise} \end{aligned} \tag{3.83}$$

The penalty term for $m(x)$ becomes:

$$\lambda ||\boldsymbol{D}\boldsymbol{\beta}||^2 \tag{3.84}$$

where $\boldsymbol{D}$ is a $(q-2) \times q$ matrix with all zeros except for $D_{i,i+1} = -D_{i,i+2} = 1$ for $i \in [1, q-2]$ and λ is the smoothing parameter to control the trade-off between goodness of fit and smoothness. The value of $\boldsymbol{\beta}$ is now chosen such that

$$||\boldsymbol{y} - \boldsymbol{X\Sigma}\tilde{\boldsymbol{\beta}}||^2 + \lambda||\boldsymbol{D\beta}||^2 \tag{3.85}$$

is minimised.

Equation 3.75 is now replaced by matrices for the sake of computation. $\sum_j f_j(z_{ji})$ is replaced by $\boldsymbol{F}_i\boldsymbol{\gamma}$ where $\boldsymbol{F}$ is the model matrix constructed from basis functions and their constraints and $\boldsymbol{\gamma}$ is the vector of coefficients. The shape constrained term m_k is now presented by a matrix $\boldsymbol{X\Sigma}$ and the corresponding vector of coefficients, with $\boldsymbol{X}$ being an $n \times m$ matrix and $\boldsymbol{\Sigma}$ the variable covariance matrix. All m_k terms are combined in a model matrix $\boldsymbol{M}$ with $\tilde{\boldsymbol{\beta}}$ a vector that contains the model coefficients β_i and the exponentiated coefficients $(exp(\beta_i))$:

$$\sum_k m_k(x_{ki}) = \boldsymbol{M}_i\tilde{\boldsymbol{\beta}} \tag{3.86}$$

Equation 3.75 becomes:

$$g(\mu_i) = \boldsymbol{A_i\theta} + \boldsymbol{F_i\gamma} + \boldsymbol{M_i}\tilde{\boldsymbol{\beta}} \tag{3.87}$$

Simplifying, the model matrices are combined into one model matrix $\boldsymbol{X}$:

$$g(\mu_i) = \boldsymbol{X_i}\tilde{\boldsymbol{\beta}} \tag{3.88}$$

where $\tilde{\boldsymbol{\beta}}$ has absorbed $\boldsymbol{\theta,\gamma}$ but still include the original $\tilde{\boldsymbol{\beta}}$. In the same way, $\boldsymbol{\beta}$ has been expanded to include the original $\boldsymbol{\beta}$ as well as $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. The penalty has the general form $\boldsymbol{\beta^T S_\lambda \beta}$, where

$$\boldsymbol{S_\lambda} = \sum_k \lambda_k \boldsymbol{S_k} \tag{3.89}$$

The original penalty matrices $\boldsymbol{S_k}$ are expanded with zeros everywhere except for the elements that correspond with the k^{th} smooth's coefficients. The objective function that must be minimised to estimate the coefficients can now be written as:

$$||g(\boldsymbol{y}) - \boldsymbol{X}\tilde{\boldsymbol{\beta}}||^2 + \tilde{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{S_\lambda}\tilde{\boldsymbol{\beta}} \tag{3.90}$$

How much smoothing: The generalised cross validation score

The General Cross Validation (GCV) score is used to determine the amount of smoothing that is required in the approximating model g (Wood, 2006). It is used to estimate the smoothing parameter λ from the penalty term $\lambda \int g''(x)^2 dx$ in the penalised cubic function

3.65. The general form of the ordinary cross-validation error, ν is given as:

$$\nu = \frac{\sum_{i=1}^{n}(y_i - \hat{g}^{-i}(x_i))^2}{n} \tag{3.91}$$

where $\hat{g}^{-i}(x_i)$ is the predicted value for $\hat{y}_i$ when the model g was fitted to all the data excluding y_i.

For the domain of λ, there exists a optimal λ_0 that produces the lowest cross validation error. Let $\hat{g}_\lambda^{-i}$ be a generated smoothed function from the actual data, where data point i have been left out of the data set. Then λ_0 will be chosen where

$$\nu_\lambda = \frac{\sum_{i=1}^{n}(y_i - \hat{g}_\lambda^{-i}(x_i))^2}{n} \tag{3.92}$$

is minimised. However the ordinary cross-validation present with some challenges when fitting additive models. Minimising multiple smoothing parameters becomes computationally expensive and it has some worrisome lack of invariance. To solve these problems, a generalised version of the ordinary cross-validation was developed by Craven and Wahba (1979) in Wood (2006), given as:

$$\nu_g = \frac{n||\boldsymbol{y} - \hat{\boldsymbol{\mu}}||^2}{[n - tr(\boldsymbol{A})]^2} \tag{3.93}$$

which is known as the *Generalised Cross-Validation* score. Hasti and Tibshirani (1990) suggested a globally applicable GCV score for the generalised additive model in terms of the deviance D of the estimated parameters $\hat{\boldsymbol{\beta}}$ and the trace of the influence matrix $\mathbf{A}$ (Wood, 2006):

$$\nu_g = \frac{nD(\hat{\boldsymbol{\beta}})}{(n - tr(\boldsymbol{A}))^2} \tag{3.94}$$

where $\hat{\boldsymbol{y}} = \boldsymbol{Ay}$.

3.7.4 Distributions considered for regression models

The Gaussian normal distribution, as well as its log-transform of the response variable were used in the statistical models.

The Gaussian distribution

The Gaussian distribution forms part of the continuous probability distributions that are described by the Gaussian equation. This equation represents an exponential decaying function which is centered around the mean μ and is scaled by the standard deviation,

σ. The mean is the center of the observations and the standard deviation is a measure of dispersion of the data. The Gaussian distribution has a PDF:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{3.95}$$

for $-\infty < x < +\infty$ with parameters mean, μ and variance σ^2 where $-\infty < \mu < +\infty$ and $\sigma^2 > 0$. Also

$$\begin{aligned} E(x) &= \mu \\ Median &= \mu \\ Mode &= \mu \\ V(x) &= \sigma^2 \end{aligned} \tag{3.96}$$

with notation $N(\mu, \sigma^2)$.

The function $f(x)$ represents the height of the curve at point x. The standard deviation determines both the width, or the stretch, of the curve as well as the height. A large standard deviation results in a wider curve with a lower maximal point, whereas a smaller standard deviation results in a narrower curve with a higher maximal point. From 3.95 for $\sigma >>$:

$$\begin{gathered} \lim_{\sigma \to \infty} {}^{1}/_{\sqrt{2\pi\sigma}} = 0 \\ \lim_{\sigma \to \infty} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = 1 \end{gathered} \tag{3.97}$$

it follows that $\lim_{\sigma \to \infty} f(x) = 0$. By implication, the spread of the data becomes wider and the height of the curve becomes lower and lower for ever increasing values of σ. On the other hand, for $\sigma <<$ the height of the curve increases and the dispersion becomes narrower.

Figure 3.8 illustrates the Gaussian normal distribution, where $\mu = 0$ for various values for σ. The larger σ results in a flatter and more dispersed curve with lower height $f(x)$.

The properties of the Gaussian distribution are:

- The distribution is symmetric around μ. This implies that 50% of the observations fall below the mean and 50% will fall above the mean.

- The total area under the curve is equal to 1.

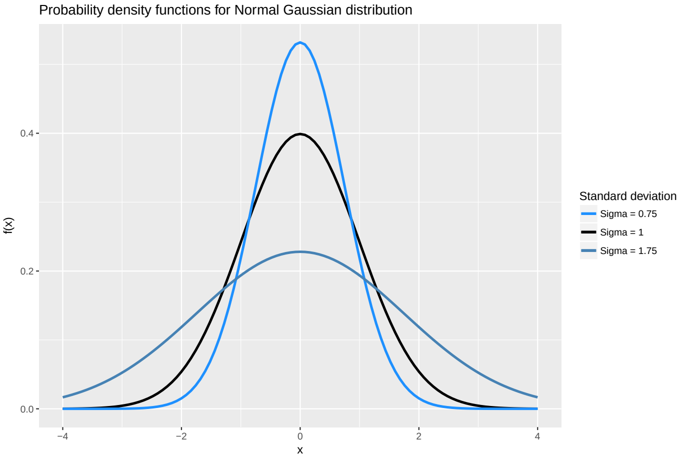


Figure 3.8: The PDF curve for a Gaussian normal distribution

- It is unimodal, that is:

$$\begin{aligned} f'(x) &> 0 \text{ for } x < \mu \\ f'(x) &< 0 \text{ for } x > \mu \\ f'(x) &= 0 \text{ for } x = \mu \end{aligned} \tag{3.98}$$

- The function tapers. Most observations occur close to the mean with less and less observations or events for values of x further away from the mean. This is a logical deduction from a exponential decaying function.

The empirical 68-95-99.7 rule states that the probability of values drawn from a normal distribution are within one, two and three standard deviations from the mean are as follows:

$$\begin{aligned} p(\mu - \sigma < x < \mu + \sigma) &\approx 0.682 \\ p(\mu - 2\sigma < x < \mu + 2\sigma) &\approx 0.954 \\ p(\mu - 3\sigma < x < \mu + 3\sigma) &\approx 0.997 \end{aligned} \tag{3.99}$$

The probabilities in 3.99 implies that 68% of the values from a normal distribution lies within one standard deviation from the mean, 95% are within two standard deviations and 99.7% fall within three standard deviations.

To calculate the probability of an occurrence, the Cumulative Distribution Function

(CDF) of the Gaussian distribution is the integral of the PDF over $x \in \mathbb{R}$:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = 1 \tag{3.100}$$

Because it is a continuous distribution function, the $p(x = b) = 0$. The probabilities are therefore always calculated within an interval $[a, b]$. Thereby, for the probability that $x < a$ the CDF becomes:

$$P(x < a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{3.101}$$

There is no closed form solution for 3.100 and the probability must be calculated numerically.

The log-transform of the Gaussian variables

For the distribution curves, the response variables, i.e. the fractional cumulative values for both the distribution curves for cadence (F_a) and pace (F_p), were transformed using the natural logarithm. The natural logarithm function for the transformed response function of Y has an asymptote at $y = 0$:

$$\lim_{y \to 0} \ln(y) = -\infty \tag{3.102}$$

This property of the natural logarithm is fundamental to the modeling of the smoothed distribution curves data, as it forces the fitted response variable to tend to zero but never become negative. The response in terms of the independent variables x_i is transformed back to the original scale:

$$\begin{aligned} e^{ln(y)} &= e^{-\infty x} \text{ the expected value of } Y \text{ becomes :} \\ \hat{Y} &= e^{-x} \end{aligned} \tag{3.103}$$

This is a distinguishing property for model selection of the distribution curves, as negative values hold no biological value and is in fact impossible. For the interaction models (between cadence and running activity and between cadence and grade) the response variable, pace, was also log-transformed to force an asymptote at $Y = 0$.

3.7.5 Associations between cadence, grade and pace

Correlation coefficients between pace and cadence and between pace and grade were calculated in order to establish whether there is a linear relationship between the variables (Montgomery and Runger, 2011). The correlation coefficient is defined as:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.104}$$

where σ_{xy} is the co-variance between variables x and y.

$$\sigma_{xy} = E[(x - \mu_x)(y - \mu_y)] = E(xy) - \mu_x \mu_y \tag{3.105}$$

The standard deviations of x and y will always be greater than zero, that is $\sigma_x > 0$ and $\sigma_y > 0$. The co-variance may take on a negative, positive or zero value. The range for the correlation is therefore $-1 \leq \rho \leq 1$. A ρ closer to -1 implies the relationship's slope to be negative and a value closer to $+1$ suggests a positive slope for the linear relationship. A value close to zero indicates there is no relationship between x and y.

3.7.6 Application: the interaction between cadence and running activity

The intercepts from the interaction models are considered as irrelevant to the analyses, as it represents a cadence of 0 which is not possible. What is of importance is the effect on cadence during the various running activities. The more negative the interaction term's value in the exponent is, the more steep the gradient of the slope (rate of change in the pace) in the linear combination of the function. The pace is becoming faster when:

$$\beta_{12} < \beta_1 < 0 \tag{3.106}$$

$$\beta_1 < \beta_{12} < 0 \tag{3.107}$$

The slope is negative for both 3.106 and 3.107 because $(\beta_{12} + \beta_1) < 0$, although 3.107 has a lower reinforcement effect on cadence than 3.106. The rate of change in pace is slower when $\beta_{12} > 0$ and interferes with cadence, but $(\beta_{12} + \beta_1) < 0$ still holds. When $(\beta_{12} + \beta_1) > 0$ the pace is decreasing for every unit increase in cadence (the runner runs slower for every unit increase in cadence). The adjusted coefficient of determination, R_a^2 was used to evaluate the interaction models. The model is coded in R as follows where the * indicates the interaction requirement:

$$lm(log(Pace) \sim cadence * running\ activity) \tag{3.108}$$

3.7.7 Application: the interaction between cadence and grade

The interaction model to analyse the reinforcing and interference effects on cadence from grade was fitted using the log-transform on pace. The interaction with grade was modeled as categorical variable (elevation_type) for DR, LR and UR in 3.109.

$$lm(log(Pace) \sim cadence * elevation\_type) \tag{3.109}$$

The same logic for interference and reinforcement as explained in 3.106 and 3.107 applies to the effect of the type of the slope.

A parametric second degree polynomial interaction model and two non-parametric methods were used to generate the 3D output for the visual analysis on graded running. The polynomial model was fitted in *R* using the *lm* function with the interaction variables specified as grade and cadence and the output as pace in 3.110:

$$lm(Pace \sim poly(cadence,\ 2,\ raw\ =\ TRUE)*(poly(grade,\ 2,\ raw\ =\ TRUE) \tag{3.110}$$

The LOESS was applied using the quadratic polynomial with the span set to 0.75 as in 3.111:

$$\begin{aligned} model = loess(Pace \sim cadence + elevation, data = data\ set, span = 0.75,\\ family = "gaussian", method = c("loess", "model.frame")) \end{aligned} \tag{3.111}$$

Although the model presented with good adequacy and was able to capture the pattern of the data, its computation is slow when applied to large data sets in excess of 1000 points. The LOESS was therefore considered as an impractical solution.

A GAM using the functionality from the *mgcv* package in *R* was applied next. Due to the size of the data set, the *bam* function was used in stead of the *gam* function. The base functions for both smooths (s) were set to “tp” for thin plate regression. The *family* setting was set to *gaussian*. All other settings were left as default. The basic formula in the *bam* function is given as follows:

$$\begin{aligned} bam(Pace \sim\ & s(cadence, bs = \text{“}tp\text{”}) +\ s(grade, bs = \text{“}tp\text{”}),\\ & data = data\ set,\ famliy = \text{“}gaussian()\text{”}) \end{aligned} \tag{3.112}$$

The *bam* creates an object of the model containing all the relevant output from the regression as well as the original input. The following data were extracted from the *bam* model objects and saved to a data frame named *extracts*:

- Original predictors (or estimators, in this case the cadence and grades)

- Original response variables
- Fitted values
- Linear predictors
- Partial residuals

The error terms e_j (or residuals) were calculated as simply the difference between the fitted value and the original response value:

$$e_j = \hat{Y_j} - Y_j \text{ for data point } j \tag{3.113}$$

The diagnostic plots (fitted values versus residuals and the qq-plot) for both variables were graphed. A 3D presentation of the TPS was constructed to show the combined effect of cadence and grade on pace with the *visreg2d* function in the *visreg* package.

The AIC scores and the R^2-values were extracted both the TPS and the second degree polynomial models that were fitted on both the original and clean-up data sets. This data frame was saved as a R-object file.

3.7.8 Application: smoothing of the distribution curves

The GAM and SCM regression techniques smooth the distribution curves for pace and cadence. For both distribution plots of cadence and pace it was decided to regress the continuous form of the counting variables (as a fraction of the total) instead of the discrete counts themselves. It is a more practical interpretation of data for a runner to read from an axis or a summary page the percentage time they have spent at a pace or cadence level than the physical count in seconds. This arrangement also simplified the models to two distribution families, i.e. *gaussian* and the *gauss-log* transform. A GAM was applied to generate a 3D TPS of the interaction between cadence, grade and pace.

The values for the distribution curves for both pace and cadence had to be calculated using the pseudocode in Algorithms 4 and 5 respectively. A bin sequence for pace was created and is referred to as the pace levels (PL), starting at the runner's minimum pace and ending at the runner's maximum pace in increments of the 0.05 min/km (this is equal to 3 seconds/km). The pace values were subset to separate vectors for the various running activities. Each instance of pace in the data set and subset vectors was evaluated against the pace value in the bin. A value of less than or equal to the pace level received a 1, otherwise 0. This code loops through all the entry lines and a 0 or a 1 is appended to the count vector (P) as in 3.114. Finally the sum is taken over the count vector in Equation 3.115. This sum is thus the total instances where the pace was less than or equal to the reference pace level. This sum is the the total number of instances that the runner was running at the pace level or slower.

$$P_j = \begin{cases} 1 & : P_j \leq PL \\ 0 & : P_j > PL \end{cases} \tag{3.114}$$

$$P = \{P_j, P_{j+1}...P_n\} \text{ where } \ j \text{ is a data point}$$

$$T_p = \sum P \text{ for pace level } p \tag{3.115}$$

These steps are repeated for each bin in the pace level sequence until all the bins have been filled.

The summed total counts (T_p) for pace were transformed into continuous variables by expressing the total count in each bin as a percentage (or fraction) of the total instances (T_t).

$$F_p = 100 \times \frac{T_p}{T_t} \tag{3.116}$$

The F_p is now the response variable in the regression analyses.

The bins for cadence were created as the sequence starting with the runner's minimum cadence up to their maximum cadence in increments of one cycle. As for pace, the data were subset into running activity and each were given its own vector. Each cadence instance in the data set was evaluated against the cadence level (CL) in the sequence. The code looped through all the data entries with each cadence value being evaluated. If the evaluated cadence was more than or equal to the cadence level in the bin, a value of 1 was assigned, otherwise 0. The assigned value was appended to the counting vector, A.

$$A_j = \begin{cases} 1 & : A_j \geq CL \\ 0 & : A_j < CL \end{cases} \tag{3.117}$$

$$A = \{A_j, A_{j+1}, ..., A_n\} \text{ where } \ j \text{ is a data point}$$

$$T_a = \sum A \text{ for cadence level } a \tag{3.118}$$

The same logic applies to cadence as for pace: T_a represents the total instances that the runner has spent running at a cadence level or higher. The summed total counts T_a for cadence in each bin were expressed as the percentage (or fraction) of the total instances.

$$F_a = 100 \times \frac{T_a}{T_t} \tag{3.119}$$

The F_a is now the response variable in the regression analysis.

Once the values for the distribution curves were calculated, the fractions F_p and F_a

Algorithm 4: Distribution curves for pace

```
Extract the data from the saved R-object file rd32
Subset the extracted data frame rd32 to include only running data and rename to rd34
Construct the distribution curve for pace: Set up bins for pace using the sequence function and name as pace levels
Set up the empty vectors for the summation vectors for the overall count and the running activities
Subset the data frame into vectors for pace from the different running activities
for each bin i in the pace levels do
    Set up empty counting vectors for each the overall count and per run activity
    for each pace entry j in rd34 do
        if $pace \leq pace\ level$ then
            assign the value 1 to the counting vector for all
        else
            assign the value 0 to the counting vector for all
        end
    end
    for each pace entry j in subset vector on each running activity do
        if $pace \leq pace\ level$ then
            Assign the value 1 to the counting vector for running activity
        else
            Assign the value 0 to the counting vector for running activity
        end
    end
end
Sum over the counting vectors and append the answer from the bin to the summation vector overall and per run running activity
Append the pace levels bin sequence and the summation vectors as columns to a data frame for the distribution curve on pace
Calculate the fraction of the total counted frequency as per Equation 3.116 for each pace level bin
```

Algorithm 5: Distribution curves for cadence

```
Extract the data from the saved data frame rd32
Subset the extracted data frame rd32 to include only running data and rename to rd34
Construct the distribution curve for cadence: Set up bins for cadence using the sequence
  function and name as cadence levels
Set up the empty vectors for the summation vectors for the overall count and per
  running activity
Subset vectors for the cadence of the different running activities
for each bin i in the cadence levels do
    Set up empty counting vectors for the overall count and each running activity
    for each cadence entry j in rd34 do
        if cadence $\geq$ cadence level then
            assign the value 1 to the counting vector for all
        else
            assign the value 0 to the counting vector for all
        end
    end
    for each cadence entry j in subset vector on running activity do
        if cadence $\geq$ cadence level then
            Assign the value 1 to the counting vector for running activity
        else
            Assign the value 0 to the counting vector for running activity
        end
    end
    Sum over the counting vectors and append the answer from the bin to the summation
      vector per running activity
end
Append the cadence levels bin sequence and the summation vectors as columns to a data
  frame for the distribution curve on cadence
Calculate the fraction of the total counted frequency as per Equation 3.119 for each bin
  in cadence level
```

were regressed against the pace and cadence levels respectively. The scatter plot for the fractional pace against the pace levels was smoothed using first the GAM and then the SCM. The base function, i.e. the spline to use to do the fit, was selected as cubic regression (cr). The *gam* function in R self-selects the number knots and their location. Function 3.120 represents the GAM:

$$\begin{aligned} &gam(Fraction\ Pace \sim s(pace\ levels, bs = \text{“}cr\text{”}),\\ &\qquad data = data\ set,\ family = \text{“}gaussian()\text{”}) \end{aligned} \tag{3.120}$$

The first iteration set the family to *gaussian* with the second iteration using the *gaussian log-link* transform. The function in 3.121 looks different for the SCM, where the base function is replaced by *mpi*, which enforces the extra penalty from Equation 3.85 to construct a strictly monotonic increasing spline.

$$\begin{aligned} &scam(Fraction\ Pace \sim s(pace\ levels, bs = \text{“}mpi\text{”}),\\ &\qquad data = data\ set,\ family = \text{“}gaussian()\text{”}) \end{aligned} \tag{3.121}$$

The scatter plot for the fractional cadence against the cadence levels was also smoothed using first the GAM and then the SCM. The first iteration was done using the *gaussian* family with the second iteration set to the *gauss-log* transform for both models. The GAM for cadence is shown in Function 3.122.

$$\begin{aligned} &gam(Fraction\ Cadence \sim s(cadence\ levels, bs = \text{“}cr\text{”}),\\ &\qquad data = data\ set,\ family = \text{“}gaussian()\text{”}) \end{aligned} \tag{3.122}$$

For the SCM, because the cadence distribution profile is monotonic decreasing, the base function is set to *mpd* which enforces the extra penalty from Equation 3.85 to construct a strictly monotonic decreasing spline in 3.123:

$$\begin{aligned} &scam(Fraction\ Cadence \sim s(pace\ levels, bs = \text{“}mpd\text{”}),\\ &\qquad data = data\ set,\ family = \text{“}gaussian()\text{”}) \end{aligned} \tag{3.123}$$

3.8 Model selection

“Truth is elusive; model selection tells us what inferences the data support, not what full reality might be” from Burnham and Anderson (2002).

The criteria for the best fit are not only based on the values for the adequacy tests, i.e.

the lowest values for the AIC and the highest adjusted coefficient of determination, but also on some self-defined criteria by the researcher for practical implications of the model. These criteria include:

- Fitted negative values.
- The fitted line's behaviour near the 0-mark on the Y -axis.
- The monotonicity of the fitted line.
- The ability of the fitted line to pick up the change in the slope of the calculated distribution curve.

The discussion starts on the model adequacy tests, followed by the application thereof as well as the self-defined criteria.

3.8.1 Model adequacy

Three adequacy measures were used to determine the best model fit to the raw data, namely:

1. The AIC score.
2. Visual results from the residual analysis.
3. The coefficient of determination, R^2 and/or the adjusted R^2, denoted as R_a^2.

The Akaike's Information Criterion

The AIC was developed by a Japanese statistician, Hirotugu Akaike in the early 1970's. The AIC estimates the relative quality of a statistical model for a certain set of data. It estimates the quality of each model relative to the other models in a group of models that are representing the same data set. It deals with the tension between the goodness of fit and the parsimony of a model by estimating how much information is lost when a certain model is used to present the data. The model with the lowest AIC score is therefore the closest to the unknown reality generated by the data set. Important to note is that it does not provide the absolute quality of any single model, only the quality relative to others. The AIC enables the analyst to choose the best model from the possible options, but does not provide information on how well the model actually fits the data. Therefore, if all models are poor the AIC will not provide any indication thereof However, Akaike's work has provided for great practical and theoretical progress in model selection for complex data (Burnham and Anderson, 2002).

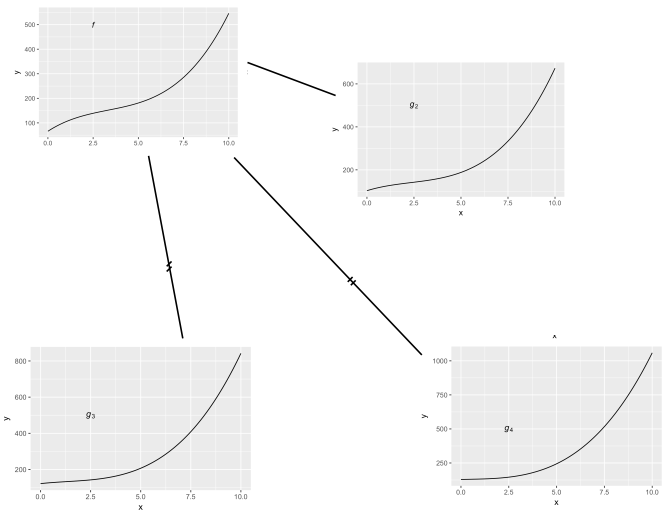


Figure 3.9: The K-L distance from candidate models g_i to reality function f

The AIC is derived from the K-L distance and the Maximum Likelihood Estimate (MLE) method from parametised models. In science and information theory the K-L distance between models is a fundamental quantity and forms the logical basis for model selection together with likelihood inference. The K-L distance can be thought of as the directed, orientated distance between two models f and g. The measure from f to g is not the same as the measure from g to f. It is therefore a measurement of discrepancy between the real function, f and the collection of functions g_i used to approximate that reality. Synonyms for the distance are divergence, information and number (Burnham and Anderson, 2002). The K-L distance for continuous distributions is defined as follows:

$$I(f,g) = \int f(x)\, ln\left(\frac{f(x)}{g(x|\theta)}\right)\, dx \tag{3.124}$$

where $I(f,g)$ is the information lost when $g(x)$ given parameters θ is used to approximate the real function f. Empirically , $I(f,g)$ is the distance from g to f. Figure 3.9 lends visual support to the definition of the K-L distance. The models g_3 and g_4 are not representative of the real function f and are placed far away from reality f in the figure. Model g_2 more closely resemble f and is therefore closer to f in the figure. The K-L distance allows the analyst to know which of the approximating models g_i are closest to f.

In order to compute the K-L distance, both f and g and their parameters must be known. Since f presents the truth or full reality it is in essence not a model but a reflection of complex processes that generated the observed data. Conceptually f might be regarded as having infinite number of parameters that gave rise to the data range x. If the unknown

f is treated as a constant C of unknown value, then the relative distance between f and g_i becomes the measurement of importance. Equation 3.124 can be re-written as:

$$I(f,g) = \int f(x)\ln(f(x))\,dx - \int f(x)\ln(g(x|\theta))\,dx \tag{3.125}$$

Both terms on the right hand side are statistical expectations with regards to $f(x)$. It follows that the K-L distance can be expressed as the difference between two statistical expectations:

$$I(f,g) = E_f\left[\ln(f(x))\right] - E_f\left[\ln(g(x|\theta))\right] \tag{3.126}$$

Let $E_f\left[\ln(f(x))\right]$ be a constant C that depends only on the unknown reality of f, then 3.126 becomes:

$$\begin{aligned} I(f,g) &= C - E_f\left[\ln(g(x|\theta))\right] \\ I(f,g) - C &= -E_f\left[\ln(g(x|\theta))\right] \end{aligned} \tag{3.127}$$

The measure $E_f\left[\ln(g(x|\theta))\right]$ is now a relative directed distance and the quantity of interest in model evaluation and selection. With f presented as a constant term that remains the same across all the candidate models and no assumptions are required about f, it becomes irrelevant for comparison.

In data analysis, the true value of the parameters θ are unknown and are estimated from empirical data. Therefore, $g_i(x|\theta)$ is denoted as $g_i(x|\hat{\theta})$. The relative distance between the candidate models g_i and the truth f now becomes the estimated relative distance. The model that gets selected has the smallest estimated distance to the reality f. Put otherwise, the model that loses the least amount of information is selected.

The K-L distance forms the logical basis for the development of the AIC. Both the parameters for f and g must be known to calculate the discrepancy. Akaike's work produced a rigorous method to approximate this distance based on the functions' empirical log-likelihoods at its maximum point. There exists a parameter θ that minimises $I(f,g)$, i.e. this value is the true value underlying the MLE of θ. The unknown value depends on:

- Truth f,
- the models g_i going through the structure of f,
- the parameters' space,
- the sample space (the structure and nature of the data that makes up f.

Let θ_0 be the true value of θ than minimises $I(f,g)$. Then θ_0 is the absolute best value for θ to be assumed in the model g. If it was known that model g was the closest to

f with the lowest K-L distance, then MLE $\hat{\theta}$ is an estimate for θ_0. The property of the model $g(x|\theta_0)$ that minimises $I(f, g)$ for $\theta \in \Theta$ is important in the derivation of the AIC. When a model is based on estimated parameters $\hat{\theta}$ for θ_0 and not on the true θ, it changes the model selection criteria from minimising the actual K-L distance to minimising the expected K-L distance.

Akaike proved the most important problem for the generation of an applied K-L model was to estimate the double statistical expectation with regards to truth, f:

$$E_y E_x \left[\ln(g(x|\hat{\theta}(y))) \right] \tag{3.128}$$

Equation 3.128 is the selection target of model selection approaches, based on the K-L discrepancy. Akaike went further to prove that estimation of 3.128 using the maximised $\ln(\mathscr{L}(\hat{\theta}|data))$ for each g_i is biased upwards of the selection target. This bias is almost equal to the number of estimable parameters in the model. A bias-correction term, K, the number of estimable parameters is introduced to counter this effect:

$$\ln(\mathscr{L}(\hat{\theta}|data)) - K \tag{3.129}$$

of which the result is the same as

$$\ln(\mathscr{L}(\hat{\theta}|data)) - K = C - \hat{E}_{\hat{\theta}} \left[I(f, \hat{g}) \right] \tag{3.130}$$

where $\hat{g} = g(\cdot|\hat{\theta})$. For “historical reasons”, Akaike multiplied 3.129 with -2 to formulate the Akaike’s information criterion:

$$AIC = 2K - 2\ln(\mathscr{L}(\hat{\theta}|y)) \tag{3.131}$$

where $\ln(\mathscr{L}(\hat{\theta}|y))$ is the log-likelihood of the maximum point that the model estimated for parameters θ. The number of estimable parameters K must include the intercept and σ^2

When all the models assume normally distributed errors with constant variance then AIC can be calculated from the *least squares*:

$$AIC = n\ln(\hat{\sigma^2}) + 2K \tag{3.132}$$

where:

$$\begin{aligned} n &= \text{sample size} \\ \sigma^2 &= \frac{\sum \hat{\varepsilon_i^2}}{n} \text{ (the MLE of } \sigma^2\text{)} \\ K &= \text{ number of parameters, inluded the error term} \end{aligned} \tag{3.133}$$

It is possible to decrease $I(f, g)$ by adding more parameters to g_i. However unknown parameters that must be estimated adds more uncertainty to the expected relative K-L distance. Equation3.130 can be re-arranged as:

$$\hat{E}_{\hat{\theta}}\left[I(f, \hat{g}\right] = C + K - \ln(\mathscr{L}(\hat{\theta}|data)) \tag{3.134}$$

At some point, as K becomes larger, the relative K-L distance, $\hat{E}_{\hat{\theta}}\left[I(f, \hat{g}\right]$, will start to increase due to the uncertainty, i.e. noise, added by unknown estimated parameters. With respect to 3.131, $2K$ will increase with more parameters and $2\ln(\mathscr{L}(\hat{\theta}|y))$ tend to decrease with more parameters. The AIC will therefore eventually increase with the addition of too many unknown parameters. This property of the AIC provides the trade-off between goodness of fit and information loss, i.e. between over- and underfitting of data. The over- or underfitting of data is fundamental to parsinomy or simplicity of data models, where a model is able to closely present the true behaviour of the data but is not trying to explain each data point.

As a conclusion on the AIC, Burnham and Anderson (2002) states: *"Rather than having a simple measure of the directed distance between two models (i.e. the K-L distance), one has instead an estimate of the expected, relative distance between the fitted model and the unknown true mechanism (perhaps of infinite dimension) that actually generated the data."*

Residual analysis

The analysis of the residuals, or the error terms, e, was briefly touched on in the section on general linear modeling. Residual analysis is usually the first diagnostic tool to see whether the model that was generated adheres to the underlying assumptions about the actual data. For a strictly linear model, these assumptions are:

- The error terms are uncorrelated random variables.
- They are normally distributed with mean of zero and a constant variance, i.e. $e \sim N(0, \sigma^2)$. The variance remains constant for all fitted $\hat{y}_i$, that is they are homogeneous.

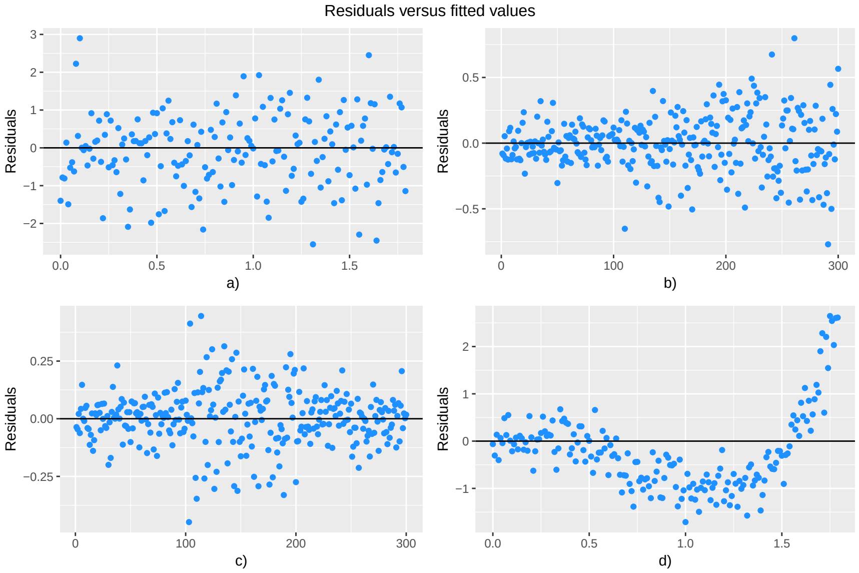


Figure 3.10: The possible patterns exhibited in a residual plot

- The order of the model is correct, that is if a linear model is fitted the assumption is that the underlying true data has a linear relationship.

Two diagnostic plots, namely the residuals plot and a qq-plot will be discussed here. The residual plot is simply a scatter plot of the error terms versus the fitted values. The patterns observed in the residual plots provide almost instant information about the normality and homogeneity of the residuals. A plot without any pattern where the error terms are scattered randomly around the zero line confirms the underlying assumptions about the residuals and that the model choice is good. A residual plot that exhibits some pattern, such as a funnel, a bow tie, curvature (non-linear) indicates that the underlying assumptions have been violated and the model choice should be reconsidered.

Figure 3.10 from Montgomery and Runger (2011) provides patterns for residual plots where $\mu_{error} = 0$. Plot *(a)* shows the random distribution of the error terms around zero and is regarded as satisfactory. Plot *(b)* has a funnel shape, *(c)* is a bow-tie shape and *(d)* shows curvature. Plots *(b)* to *(d)* violates the assumptions for linearity, therefore the model choice should be reconsidered. In both the bow-tie and funnel shaped anomalies the variance is changing for increasing value of y and a transformation must be considered to stabilise the variance. Variables may be transformed as generalised linear models, more terms may be added or other non-parametric data modeling must be considered.

In a normal probability qq-plot, the residuals are standardised as $d_i = {}^{e_i}/\sqrt{\sigma^2}$. If the error terms are indeed normally distributed, 95% of the values should be within the range $(-2, 2)$. When $e_i < -2$ or $e_i > 2$ they indicate an outlier or an observation that does not share the same characteristics of the rest of the data. Further investigation into the

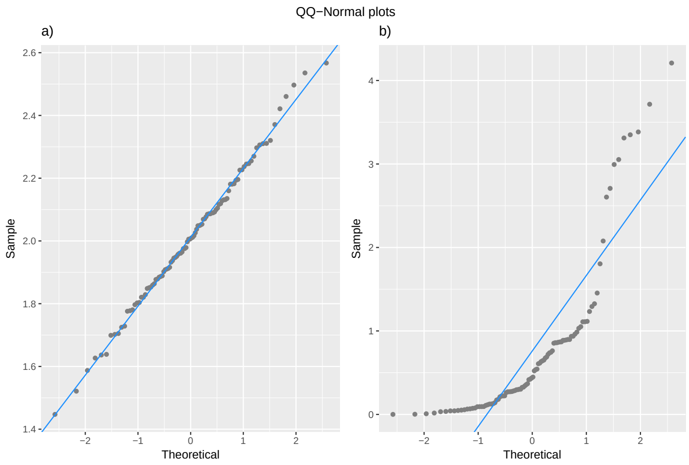


Figure 3.11: The scatter of error terms around the normalised straight line

outliers should be conducted as to ascertain their value or otherwise insightful meaning into the data. The actual residuals e_i are on the $y-$axis and the standardised or theoretical residuals, d_i are on the $x-$axis. Departure from the straight line indicates non-normality of the residuals when the assumptions about the underlying data are wrong (Montgomery and Runger, 2011). Figure 3.11 shows a normal probability qq-plot for some residuals. The residuals in plot *(a)* is scattered closely around the straight line with most of the points falling between -2 and $+2$. The residuals in plot *(b)* clearly departs from the normal line. The underlying data that generated the residuals in plot *(b)* is not normal and may have arose from an exponential distribution. Transformation on the response variable may solve this problem and stabilise the variance. The qq-plot generates a straight line around which the error terms must fall. The intercept of the straight line is equal to the mean of the error and the slope is equal to the standard deviation. For the errors to be normally distributed, the errors must have a mean of 0 and a standard deviation of 1. The straight line that results from a normally distributed line is there the line $y = x$ and runs at a 45° angle. However, when the mean of the errors are not equal to 0 and the standard deviation is not equal to 1, the line does not fall on $y = x$ and the errors therefore do not comply with the normal distribution. This quick evaluation of the intercept and slope of the straight line is also an indication of the expected behaviour of the data.

The coefficient of determination, R^2

The coefficient of determination can be generally defined as the amount of the variability in the data explained by the regression model. It is a ratio of the sum of squares:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \tag{3.135}$$

whereby $0 \leq R^2 \leq 1$ and where

$$\begin{aligned}
SS_E &= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \text{ i.e. error sum of squares} \\
SS_R &= \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \text{ i.e. the regression sum of squares} \\
SS_T &= \sum_{i=1}^{n} (y_i - \overline{y})^2 \text{ i.e. total corrected sum of squares} \qquad (3.136)
\end{aligned}$$

SS_E is the sum of the squared errors between the actual observed value y_i and the fitted mean value for that point, $\hat{y}_i$. SS_R is the sum of the squared errors between the mean fitted $\hat{y}_i$ and the overall mean for y, i.e. $\overline{y}$. SS_T is the sum of the squared errors between the actual observed y_i and the overall mean $\overline{y}$ and is also referred to as the *analysis of variance identity.* A R^2 value closer to 1 implies a good model fit, while a value nearing 0 is indicative of a poor fit to the actual data. Whereas the AIC provides the analyst with the model that is closest to the truth amongst a group of models, the R^2 tells the analyst how well the generated model fits the data.

Although relatively simple to calculate from the output of a regression model, the R^2 number should be used with caution. The value can easily be improved by simply adding another term to the model, which may lead to the overfitting of the data (Montgomery and Runger, 2011). For example, a perfect fit may be achieved by a polynomial of degree $n-1$ for n data points, that is the line goes though every data point. Looking at SS_E, this value will decrease as $\hat{y}_i$ comes closer to y_i due to additional parameters. The term SS_T however remains constant. A decreasing term, ${}^{SS_E}/_{SS_T}$ is subtracted from 1, resulting in a R^2 closer to 1.

The adjusted R^2, denoted as R_a^2 is preferred above the normal R^2 with multiple linear regression (as done in the interaction and polynomial models), as it takes into account the addition of more independent variables into the model (Neter et al., 1988). It is expressed as:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\left(\frac{SS_E}{SS_T}\right) \qquad (3.137)$$

where p increases by 1 when another independent variable is added to the model. R_a^2 will become smaller if it is not counteracted by a sufficient decrease in SS_E. The additional independent variable must therefore add true value to the model, otherwise R_a^2 will become smaller and the regression model becomes inadequate.

3.8.2 Application: fitted negative values

For both the pace versus cadence and the distribution lines negative values hold no biological meaning and is impossible to achieve. A scoring variable, called the Zero Score (ZS) was set up find the total number of instances where the fitted line gave a negative value. A score greater than 0 means the line dips below $Y = 0$. A counting vector (N) was used to identify instances when a negative value was fitted as is defined as per Equation 3.138. The algorithm in 6 codes the logic behind the extra model evaluation criteria for the ZS calculation.

$$N_j = \begin{cases} 1 & : \hat{Y_j} < 0 \\ 0 & : \hat{Y_j} \geq 0 \end{cases} \tag{3.138}$$

The ZS is the sum over the counting vector N:

$$ZS = \sum N \tag{3.139}$$

Algorithm 6: Extra Model Evaluation Criteria for Negative Values

```
for each fitted value j in extracts do
    if Fitted value $\hat{Y_j} \leq 0$ then
        Assign the value 1
    else
        Assign the value 0
    end

end
Take the sum of the assigned values as the variable $ZS$
```

3.8.3 Behaviour near $y = 0$

The same logic applies as for the fitted negative values, however in the case of the distribution curves, a smaller distance to 0 with a small error term is an indication of a better fit. The fitted line must not pass the 0-line on the y -axis. It is important to consider the distribution models' behaviour at the maximal end for cadence and the minimal point of pace, as these are the points where the fitted lines near 0. A simple variable was created to find the value near 0. The first fitted Y-value for pace and the last fitted value for cadence was used to evaluate the proximity to 0. This value is the PZ, or its *proxy zero*. The error terms for these fitted values were also extracted as *error zero*.

3.8.4 Monotonicity of the distribution curves

Both the distribution curves are by their nature and construction monotonic, thus the fitted line must display similar behaviour. A monotonic curve in a singular direction is defined as:

$$\begin{aligned} &f(x) \text{ is increasing for every } x \text{ when } f(x_{i+1}) > f(x_i) \\ &f(x) \text{ is decreasing for every } x \text{ when } f(x_{x+1}) < f(x_i) \end{aligned} \tag{3.140}$$

The monotonicity score (MS) counts the number of instances where the sign of the instantaneous slope between two adjacent points changes. The score should be equal to 0 for a monotonic curve. A score greater than 0 means that the fitted line is not monotonic. The counting vector (C) was defined differently for cadence (monotonic decreasing, C_d) and pace (monotonic increasing, C_i). Monotonic decreasing for cadence is defined in Equation 3.141 with the pseudocode for its calculation in Algorithm 7.

$$C_d = \begin{cases} 1 & : \hat{Y}_{j+1} - \hat{Y}_j > 0 \\ 0 & : \hat{Y}_{j+1} - \hat{Y}_j < 0 \end{cases} \tag{3.141}$$

Algorithm 7: Monotonicity Scores for Cadence

```
for each fitted value j in extracts do
    if difference between $\hat{Y}_{j+1}$ and $\hat{Y}_j > 0$ then
        Assign the value 1
    else
        Assign the value 0
    end
end
Take the sum of the assigned values as the monotonicity score $MS$.
```

Monotonic increasing for pace is defined in Equation 3.142 and Algorithm 8.

$$C_i = \begin{cases} 0 & : \hat{Y}_{j+1} - \hat{Y}_j > 0 \\ 1 & : \hat{Y}_{j+1} - \hat{Y}_j < 0 \end{cases} \tag{3.142}$$

The MS is the sum over the counting vector:

$$MS = \sum C \tag{3.143}$$

The best fit for each model per athlete were chosen based on a combination of the

Algorithm 8: Monotonicity Scores for Pace

```
for each fitted value $j$ in extracts do
    if difference between $\hat{Y}_{j+1}$ and $\hat{Y}_j < 0$ then
        Assign the value 1
    else
        Assign the value 0
    end
end
Take the sum of the assigned values as the monotonicity score $MS$.
```

model adequacy measures as obtained form the model objects and the summary, the line's ability to asymptotically convert to $Y = 0$, i.e. it's PZ and the error term near 0, the ZS and in the case of the distribution curves the MS. A lower AIC-score and GCV with a higher R^2 from the adequacy checks are preferred when models are compared to each other. Furthermore, the ZS and MS must be equal to 0. Should either of these scoring values be greater than 0, the model is rejected irrespective of its favourable adequacy checks. A lower PZ is preferred, i.e. the first (pace) or last (cadence) fitted value comes closer to 0.

Chapter 4

The interaction between cadence and running activity as explanatory variables for running pace

The raw data (containing only running instances) are subset by the type of running activity for the analysis. Cadence as provided by the device is the number of complete stride cycles. A complete stride cycle is from toe-off to the next footfall of the same foot. The measure thus accounts for two full steps. It is sometimes referred to as cadence levels but it is treated as an integer variable. Running activity serves as a categorical variable with five levels. Running activity can be a proxy for the running surface, as most track training takes place on grass and trail running is on gravel roads. Road running and road racing occur on asphalt road. The premise was that there will be significant dissimilarities in running form while running on the various surfaces and under differing conditions. An athlete has a different mind set during racing than training and the body responds differently to various running surfaces. An Analysis of Variance (ANOVA) was performed to test the differences in pace on the different running surfaces (or during the different running activities). The ANOVA tests showed significant changes in running pace between the running activities for all the athletes ($p - value < 0.01$). Data had to be cleaned from the significant outliers per cadence level based on Equation 3.9 and using Algorithm 2. Another clean-up based on visual inspection followed so that the essence of the pattern of the pace on each cadence level could become more clear. Working with the cleaner data set concentrated the data to include only instances within expected reasonable limits. Two interaction models were fitted to the data, both before and after the clean-ups:

1. A log-transform on pace as the response variable with cadence and running activity as explanatory variables. The logarithm of pace was used in order to avoid negative

fitting of pace.

2. A cubic polynomial model with pace as response and cadence and running activity as explanatory variables was fitted to capture expected curvature in the response.

The objective of these models is to explain the expected interaction between cadence and the running activity and find the re-enforcement or interference effects. These models are therefore aimed at explaining running form on different surfaces, not to predict pace.

4.1 Case study A: the semi-professional all-rounder

Athlete 3 is a highly capable runner able to compete both on road and trail. He partook in all five running activities during the time of data collection: road running, road racing, trail running, trail racing and track training. A total of 399 365 data points covering 1256 km of running were extracted for this athlete. The mean average time between data points is 1 second. He has the largest data set of the four participating athletes. Figure 4.1 shows the raw data subset by running activity with pace on the y-axis and cadence on the x-axis. It is clear that the combinations of pace and cadence differ for the type of running activity. The facet plot enhances the original spread of the data in the overall scatter plot. Each activity seems to form a cluster with the rest of the data spreading out from the cluster. Both road running and track training appear to have two clusters. Track training shows the largest spread of data across the cadence levels.

Important to notice is the great variability in the pace for each cadence level irrespective of the running activity. In Figure 4.1 it is seen that pace for a running cadence of 90 ranges between the minimum of 2.5 min/km up to the slowest values at roughly 8.3 min/km. It becomes difficult to associate cadence levels with clear ranges of pace. Furthermore it is visually challenging to find a definite pattern in the scatter plot alone, except for the data becoming scarce at roughly a cadence of 102 and above 5.5 min/km, as well as below 4.5 min/km with cadences between 76 and 82 cycles/min. The analysis on the spread of the pace values per cadence level revealed some interesting information, more specifically the behaviour of the data around its median for each cadence level. The third moment around the mean of the data is used to quantify its skewness. The boxplots in Figure 4.2 a) illustrate the skewness in the data (outliers are coloured light blue). The line inside the boxplot represent the median and the dot inside the box is the mean for the pace on that specific cadence level. The spread per cadence level is skew to the left up until a cadence of 84 is reached, where after the distributions become skew to the right. The skewness in the data gets amplified for the higher cadences: at a cadence of 102 the mean coincides with the 75th percentile of the data and is highly skewed to the right with

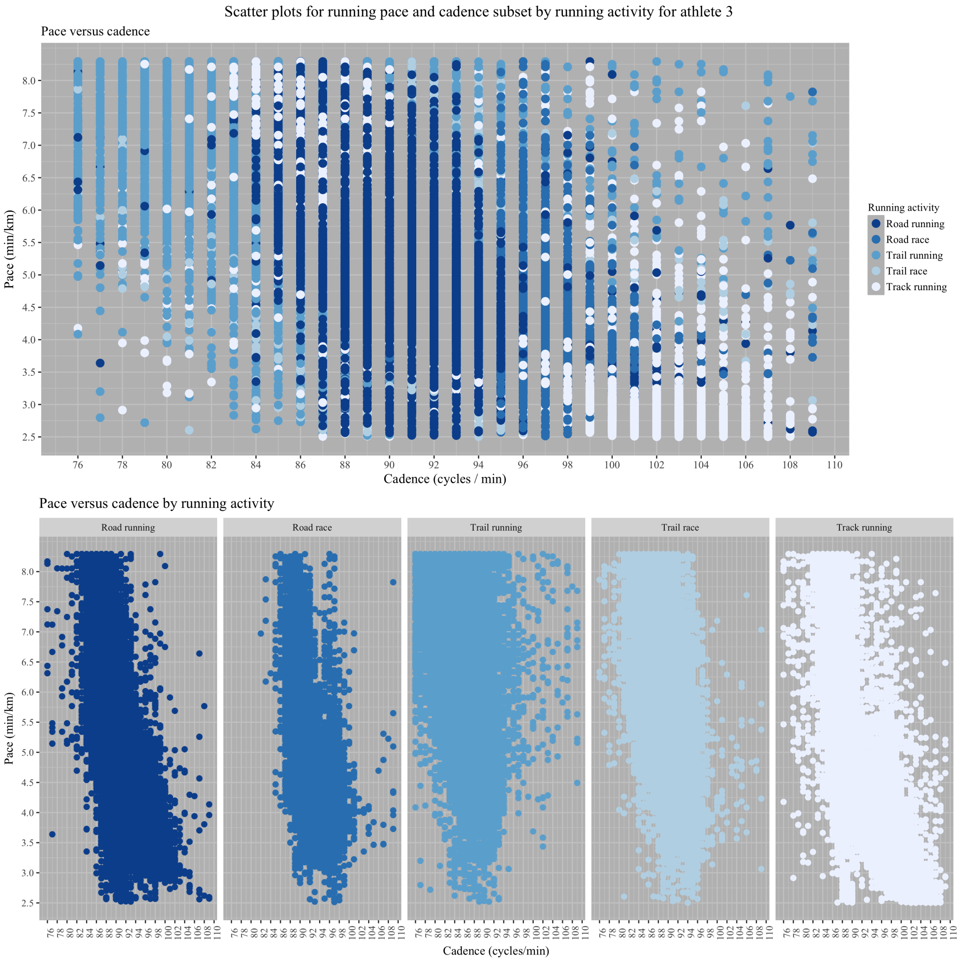


Figure 4.1: The raw data for each observed data point

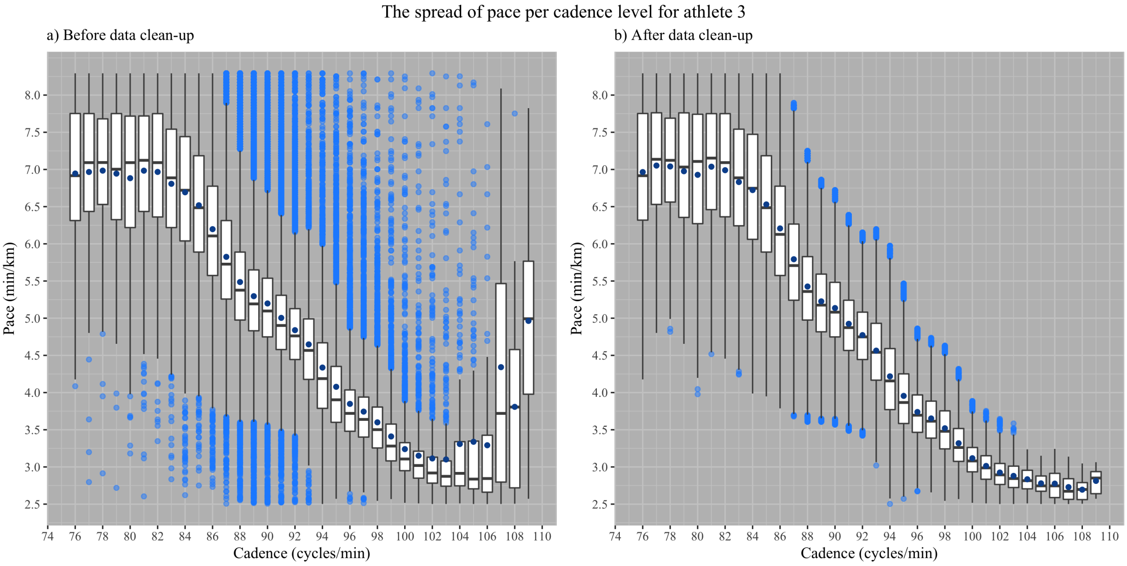


Figure 4.2: The spread of pace across cadence (median indicated by the line and mean indicated by the dot) for athlete 3

a skewness measure of 3.96.

The pattern in the data surfaces more clearly in the boxplots than the scatter plots. The medians of the boxplots in Figure 4.2 a) take on a curvi-linear pattern, starting out relatively flat before turning downward at a cadence of 83 and finally curving upwards again at a cadence of 103. The overlapping of the IQRs are apparent. For instance, a pace value of 7 min/km is associated with the range of cadence stretching from 76 to 85 before and after the clean-up. At the faster range of pace, a pace of 3 min/km is associated with cadences 100 to 103. In order to capture the pattern observed in boxplot a), the data had to be cleared from the outliers as per Equation 3.9 and using Algorithm 2. Data points higher or lower than the defined upper and lower ranges of pace for each cadence level were removed from the data set and re-plotted. Figure 4.2 b) shows the boxplot after the outliers have mostly been removed. The overall skewness of the data improved with the means now being closer to the medians. Nearly 92 % of the cadence levels showed a decrease in their skewness and the overall skewness saw a 70.4% reduction after removal of the outliers. There remains evidence of a curvi-linear relationship between cadence and running pace, albeit extent of the upward curve has now been reduced.

Table 4.1 contains the interaction coefficients from the ICM with the log-transformed response variable for the original data set. The base case is the road running activity and all the other activities' interactions are measured against road running. The interaction coefficients are indicated as the combination of the cadence and the running activity. For example the interaction between cadence and road racing is shown as "Cadence:road". The base intercept is the intercept for road running and "Cadence" is the slope for road running. The other running activities are the effects of the running activity on the base intercept. For the biological purpose of this analysis, the intercepts are not analysed as it implies a cadence of 0, which means the athlete is stationary and therefore cannot have a pace. All the coefficients are those in the linear function in the exponent when the response variable is back-transformed from the log-function.

$$Y = e^{\beta_0+\beta_1 x+\beta_{12}x_1x_2+\beta_{13}x_1x_3+\ldots\beta_{15}x_1x_5} \tag{4.1}$$

Depending on the running activity's presence, the associated coefficient's value is added to the slope of cadence, (β_1). For the base case of road running, the rate of change in pace is -0.03662 per unit increase in cadence. The variable that has the highest re-enforcing effect on the rate of change in pace per unit change in cadence, is track training (β_{15}). That is, for one unit change in cadence the linear coefficient in the exponential function for pace is further increased in negativity by -0.01. The slope of the straight line becomes -0.04662 ($\beta_1 + \beta_{12}$). This means that during track training, the expected change in pace per unit increase in cadence is more negative than during road running. The highest interference effect is seen with trail running where the slope of the straight line becomes less negative at -0.0303, meaning that the athlete's change in pace has slowed down when compared to road running. Both these changes in the slope is significant on the 5% level with their p-values lower than 0.05. The effect of trail racing is insignificant. The effect of cadence during road racing on pace is -0.0358, also a slower change than track training but slightly faster than trail running.

Table 4.1: Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 3 (original data)

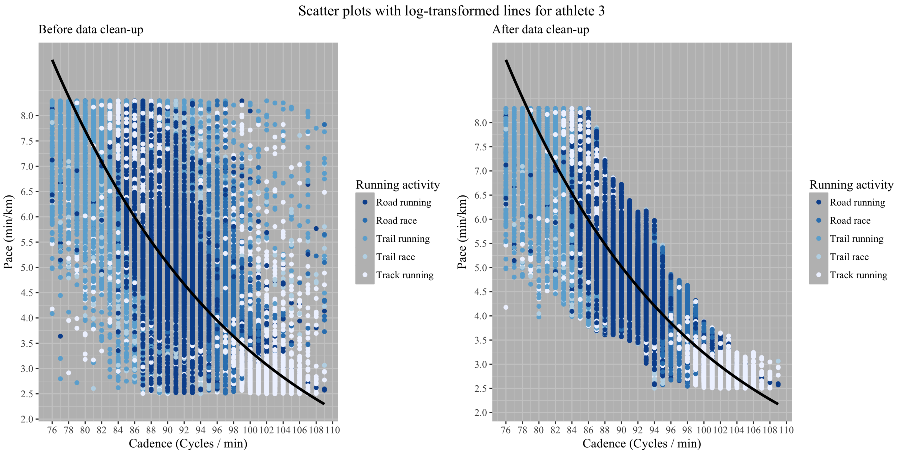
	Estimated parameter	P-value
Intercept (β_0)	4.89837	0.00000
Cadence (β_1)	-0.03662	0.00000
Road race	-0.06980	0.04278
Trail running	-0.75601	0.00000
Trail race	0.04797	0.06652
Track training	0.90369	0.00000
Cadence:road race (β_{12})	0.00134	0.00040
Cadence:trail running (β_{13})	0.00930	0.00000
Cadence:trail race (β_{14})	0.00021	0.46985
Cadence:track training (β_{15})	-0.01000	0.00000

The clean-up of the data had a significant impact on the interaction coefficients, seen in Table 4.2. The interaction between cadence and trail racing has become significant with the interaction having a negative effect on the slope. Track training has a more negative effect on the slope, whereas road racing changed from slowing the pace to increasing it.

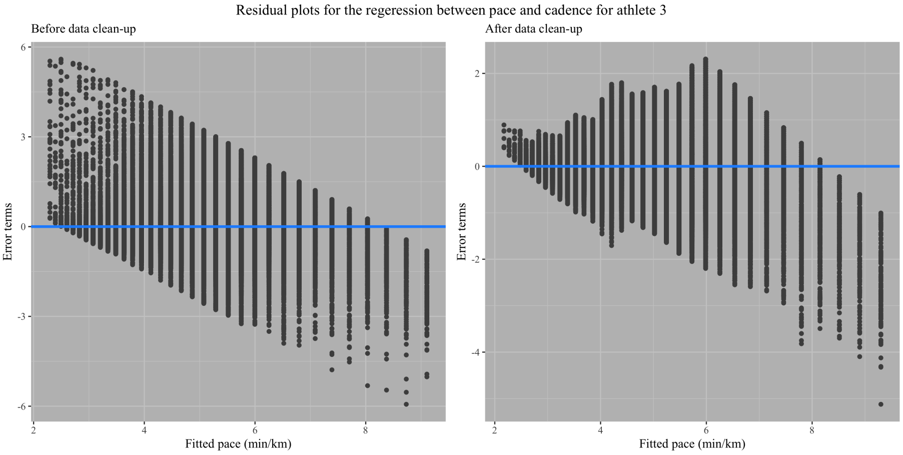
Table 4.2: Estimated parameters for the log-transformed interaction model from cadence and running activity on pace (cleaned data)

	Estimated parameter	P-value
Intercept (β_0)	4.96698	0.00000
Cadence (β_1)	-0.03743	0.00000
Road race	0.44882	0.00000
Trail running	-0.15696	0.00000
Trail race	0.38148	0.00000
Track training	0.94712	0.00000
Cadence:road race (β_{12})	-0.00449	0.00000
Cadence:trail running (β_{13})	0.00235	0.00000
Cadence:trail race (β_{14})	-0.00369	0.00000
Cadence:track training (β_{15})	-0.01050	0.00000

Table 4.3 contains the summary results from the regressions before and after the data had been cleaned. Figure 4.3 contains the fitted log-transformed lines and the associated residual plots for the models before and after the data clean-up.



(a) The log-transformed fitted lines



(b) Residual plots

Figure 4.3: Fitted lines and residual plots before and after the data clean-up for athlete 3

Table 4.3: Summary from the GLMs for the effect of cadence on pace for athlete 3

Variable	Value
R_a^2 (cubic polynomial)	0.466
R_a^2(cubic polynomial on cleaned data)	0.566
Change in R_a^2 (%) for cubic polynomes	21.565
R_a^2 (log transform)	0.520
R_a^2(log transform on cleaned data)	0.618
R_a^2 on ICM (log transform)	0.552
R_a^2 on ICM (log transform on cleaned data)	0.638
Change in ICM's R_a^2 (%) for log transforms	15.525
Change in overall skewness (%)	70.435
Total case improvements in skewness (%)	91.176

The models' adequacy improved by 21.5% for the cubic polynomial and with 15.5% for the log transformed model after the data had been cleared. This heterogeneity in the residuals confirms that the underlying data is not normally distributed and skewed.

Working backwards from a cadence of a 102, the boxplots are becoming larger, corresponding to the behaviour of the errors after 3.1 min/km. At the lower spectrum of cadence levels in the boxplots the removal of the errors did not change the behaviour of the data significantly, and therefore the same downward pattern observed in the residual plot before the cleanup is present after the clean-up. This cyclic pattern of inward and outward funneling of the errors is therefore related to how the boxplots behaved after the clean-up.

Despite the improvements in the models' adequacy, the R_a^2-values show mild strength in being able to capture the variability in the data.

4.2 Case study B: the trail specialist

Runner 4 is a trial specialist, especially over long ultra-marathon distances. A total of 105 206 data points covering 975.9 km were extracted from the GPS container files for this athlete. The mean duration between the data points is 3.67 seconds. Figure 4.4 shows the data points for pace and cadence subset by colour into four running activities (road running, trail running, trail racing and track training) before the data clean-up. Both the trail running and trail racing form clusters starting in the the upper left corner (lower cadence, slower pace) and emanate diagonally downward. Road running and track training have clusters more towards the middle of the diagram with scattered tails on in the upper-left side. It seems that track training may be split into two clusters, judging by the slight break-up in density of points at 4 min/km and a cadence between 86 and 92. The scatter plot also reveal the variability of pace generated per cadence level. For instance on a cadence of 89, the athlete generated paces across the entire range of pace.

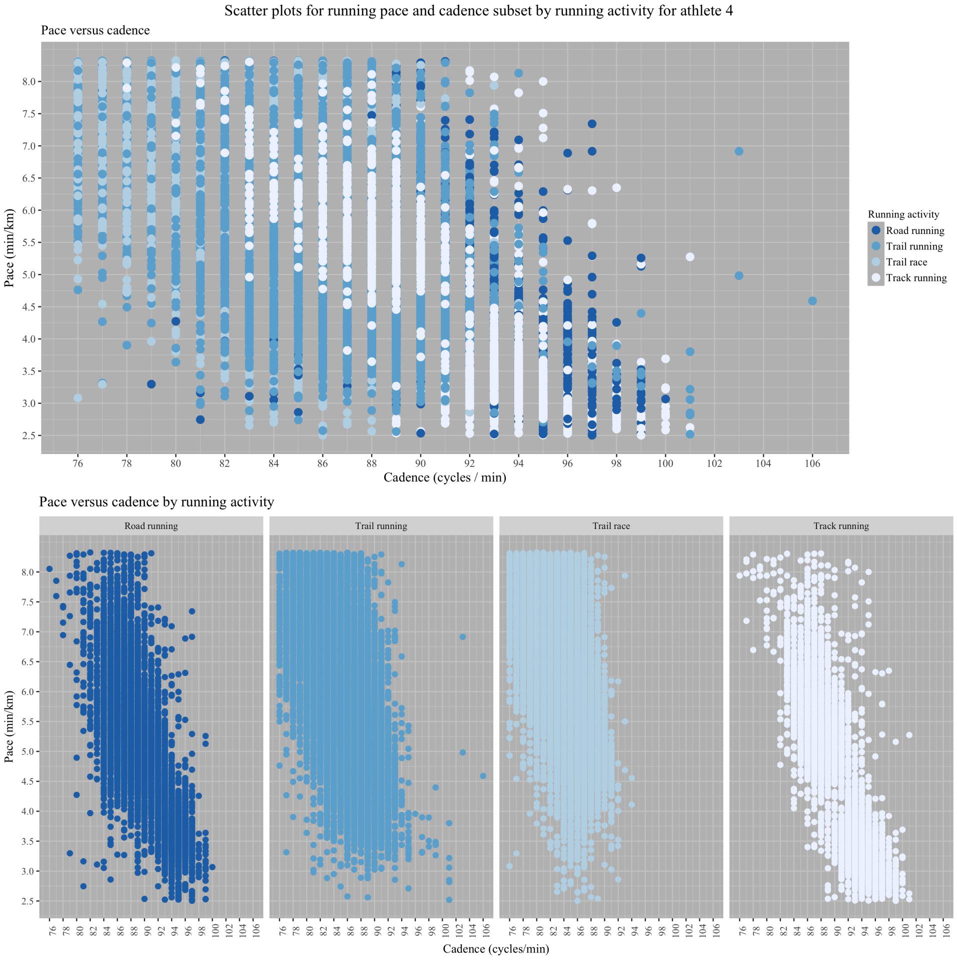


Figure 4.4: The raw data for each observed data point for athlete 4

A general pattern on the spread of the data is visible, with data points becoming scarce between a cadence of 76 and 83 below the 4.5 min/km mark. Data points in the upper right quadrant are becoming scarcer as well. The least variability is seen at the upper range of cadence.

The boxplots for pace per cadence level before and after the data clean-up is shown in Figure 4.5. The data is skew for each cadence level, going from being left skewed up to the cadence of 82 after which is becomes mostly skew to the right. For both figures the IQRs contract from a cadence of 76 up to 90, after which it expands again in size for two levels and then contracts again up to a cadence of 100. This contraction-expansion behaviour of the IQRs may be indicative of the runner's ability to concentrate his running pace for each increasing cadence level. The boxplot for the cadence at 103 and 106 in the data before the clean-up is completely out of sync with the general downward pattern of the IQRs for increasing cadence levels. Referring back to the scatter plot in Figure 4.4, these boxplots present only a few instances captured for a cadence of 103 and one at a cadence of 106. The data were cleaned from outliers as per Equation 3.9 and using Algorithm 2. The data then underwent another visual clean-up and points that are out of

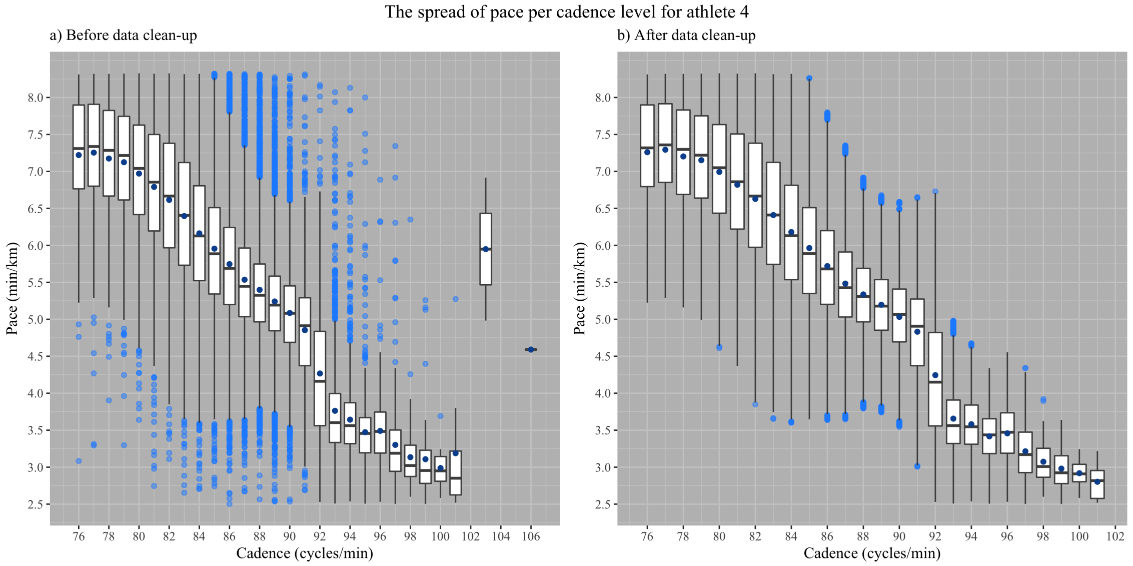


Figure 4.5: The spread of pace across cadence for athlete 4

sync with the general pattern were removed. The cleaned data set was re-plotted and is presented in Figure 4.5 b). The range of cadence now only goes up to 101. All the outliers between a cadence of 76 and 79 were removed. The spread of cadence levels for the range of pace is evident. Eight cadence levels are linked to a pace of 7 min/km, whereas 3 min/km is associated with four cadence levels. The pace of 3 min/km was connected to five cadence levels before the clean-up. There is improvement in the skewness of the data. Perhaps the most improvement in the skewness is seen at a cadence of 101, where the mean has shifted from the border of the 75^{th} percentile to just below the median. The overall skewness improved by 71.37% and 89.29% of the cadence levels became less skew. The contraction-expansion-contraction behaviour of the IQR remains, although the sizes of the boxes for the higher end cadences are now smaller. All the coefficients are those in the linear function in the exponent when the response variable is back-transformed from the log-function.

$$Y = e^{\beta_0+\beta_1 x+\beta_{12}x_1x_2+\beta_{13}x_1x_3+\beta_{14}x_1x_4} \tag{4.2}$$

The interaction between cadence and running activities and their effect on pace were investigated using the log-transform regression model. The interaction terms for the analyses before the clean-up are shown in Table 4.4. All the terms are significant on an α-value of 0.05. Both trail running and trail racing interfere with pace and slows down the rate of change, with net result for the slope on trail running being -0.0307 and for trail racing 0.03269. Track training re-enforces cadence and produces a more negative slope at -0.06369 increase in pace per unit increase in cadence.

Table 4.4: Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 4 (original data)

	Estimated parameter	P-value
Intercept (β_0)	5.88273	0.00000
Cadence (β_1)	-0.04779	0.00000
Trail running	-1.51601	0.00000
Trail race	-1.32429	0.00000
Track training	1.35182	0.00000
Cadence:trail running (β_{12})	0.01709	0.00000
Cadence:trail race (β_{13})	0.01510	0.00000
Cadence:track training (β_{14})	-0.01590	0.00000

Table 4.5 contains the coefficients for the ICM after the data had been cleaned. The coefficient for the slope has increased in negativity, implying a steeper downward slope. All of the interaction coefficients have changed slightly. Track running now has a lower re-enforcement effect and trail running and trail racing's interference is less.

Table 4.5: Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 4 (cleaned data)

	Estimated parameter	P-value
Intercept (β_0)	5.98146	0.00000
Cadence (β_1)	-0.04899	0.00000
Trail running	-1.47742	0.00000
Trail race	-1.31163	0.00000
Track training	1.34119	0.00000
Cadence:trail running (β_{12})	0.01665	0.00000
Cadence:trail race (β_{13})	0.01498	0.00000
Cadence:track training (β_{14})	-0.01574	0.00000

Table 4.6 contains the summary results from the regressions before and after the data had been cleaned. Figure 4.6 contains the fitted log-transformed lines and the associated residual plots for the models before and after the data clean-up. These lines are not subset by running activity and is only meant to showcase the general fit of the line over all the data. The line presents the continuous downward trend of the original observations (before the clean-up) but is not a good fit to the observed data with an R_a^2-value of 0.587 (from the log-transformed ICM). It also overfits the data before the lower cadences completely, but this may be due to the absence of data lower than 76 cycles/min. The scattered points at the cadences above 104 did not affect the trajectory of the line, as the logarithmic function is monotonic and tends towards 0. The errors are simply the difference between the fitted value (on the response scale, that is back-transformed from the exponential function) and the original observation of pace. The general trend of the residual plot is an outward funnel up to roughly a pace of 3.75 min/km after which there is a continuous diagonally downward trend.

The model from the cleaned data performed better with a 11.05% improvement of the R_a^2-value from the ICM, now at 0.651. This overall line is a better fit and does not under-estimate the pace at the highest cadence of 102. The pattern in the residual plot has also changed, taking a more wave-like form with increasing amplitudes from the 0 line. The error pattern changes between outward and inward funneling throughout the fitted range. The clean-up of the data did not stabilise the error terms' behaviour, but changed its patterns instead.

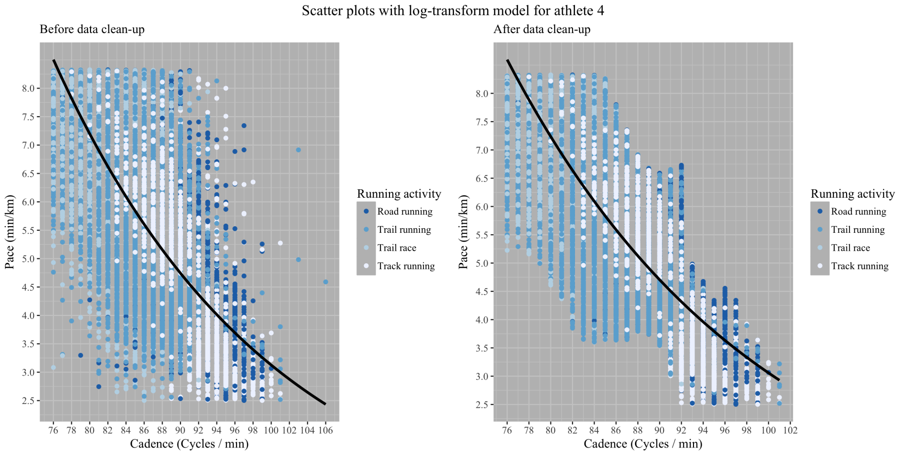
Table 4.6: Summary from the GLMs for the effect of cadence and running activity on pace for athlete 2

Variable	Value
R_a^2 (cubic polynomial)	0.536
R_a^2 (cubic polynomial on cleaned data)	0.603
Change in R_a^2 (%) for cubic polynomes	12.467
R_a^2 (log transform)	0.541
R_a^2 (log transform on cleaned data)	0.605
R_a^2 on ICM (log transform)	0.587
R_a^2 on ICM (log transform on cleaned data)	0.651
Change in ICM's R_a^2 (%) for log transforms	11.050
Change in overall skewness (%)	71.374
Total case improvements in skewness (%)	89.286

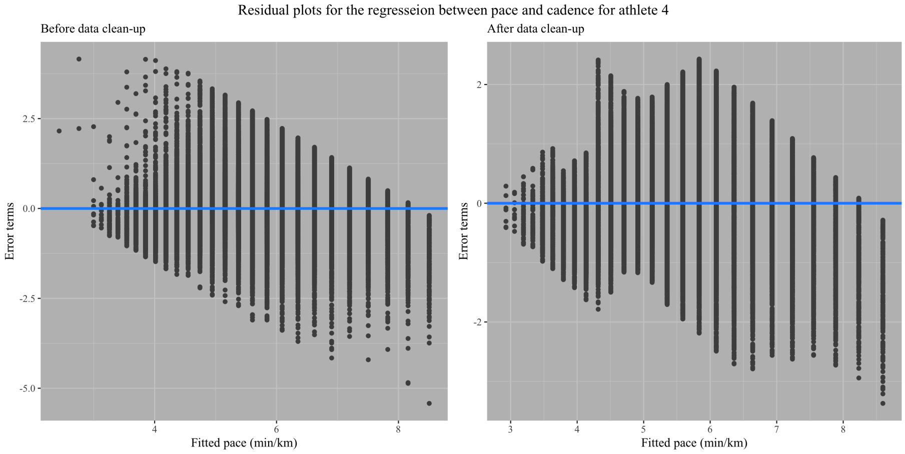
4.3 Case study C: the Comrades marathoner

Runner 2 prefers road running and is a sub 7 hours 30 minutes Comrades marathon finisher (a silver medalist). A total of 42 281 data points covering 684.41 km of running were recorded for this athlete. Data were recorded with a mean interval of 4.9 seconds between two points. The running activities for this athlete are road running, road racing and track training. Figure 4.7 shows the data points for pace and cadence subset by colour into the three running activities before the data clean-up. Road running forms a cluster between a cadence of 78 and 87 and within the pace range 4 to 6.5 min/km. Another smaller cluster is visible between a cadence of 86 and 91 and between paces 3 and 3.5 min/km. The spread of road racing across cadence is more condensed than road running. The data moves diagonally down for increasing cadence levels. The spread of the data across track training is extensive, with possible outliers at the higher cadences and the slower paces. Track training seem to have two clusters as indicated. The data points also move diagonally down from left to right extending beyond the scarce points observed for road racing at the higher cadences.

The boxplots for pace per cadence level before and after the data clean-up is shown in Figure 4.8. The IQRs follow a downward curvi-linear pattern with the spread of the IQRs



(a) The log-transformed fitted lines



(b) Residual plots

Figure 4.6: Fitted log-transformed lines (response scale) and residual plots before and after the data clean-up for athlete 4

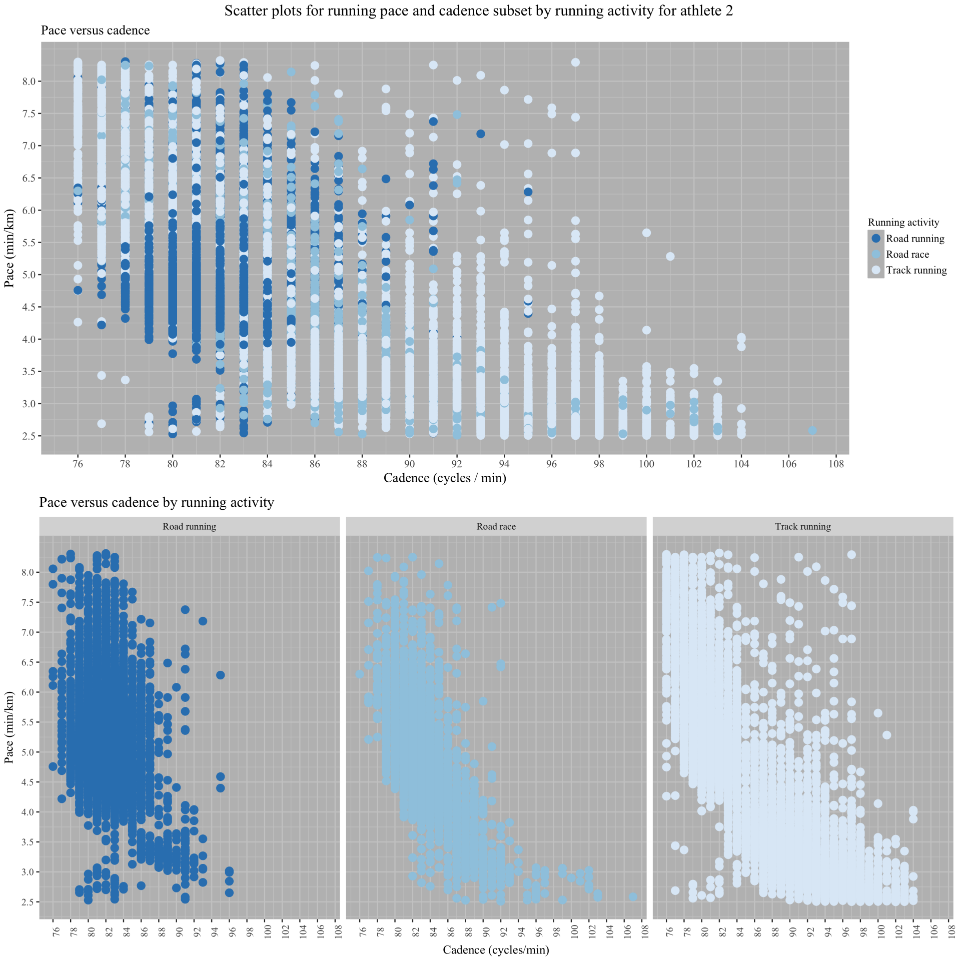


Figure 4.7: The raw data for each observed data point for athlete 2

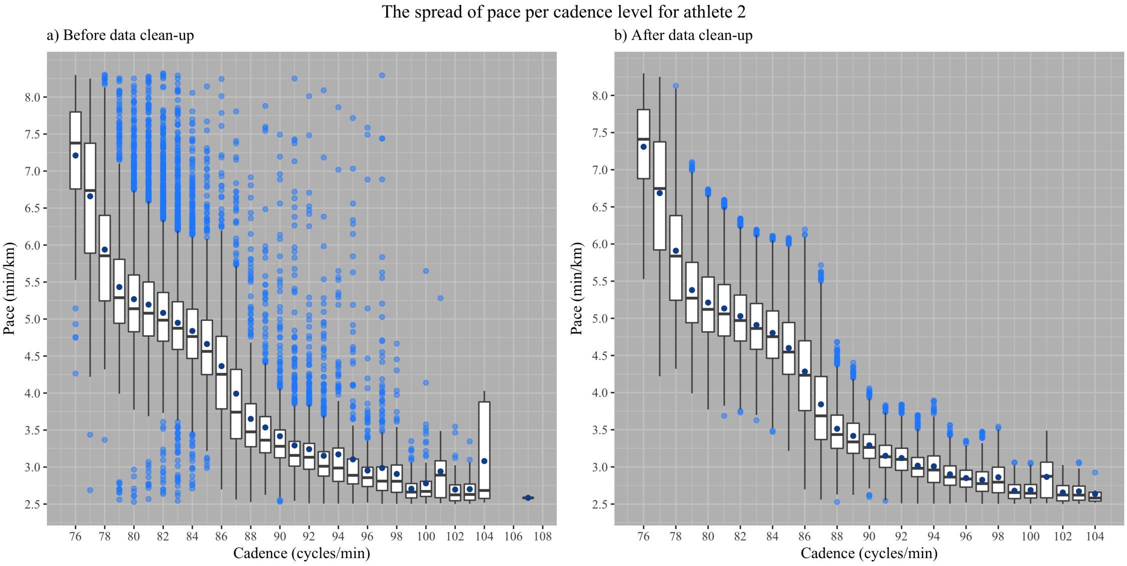


Figure 4.8: The spread of pace across cadence for athlete 2

also contracting with the exception of a cadence of 101 and 104. The spread per cadence level is skewed to the right for all levels, except cadences of 76 and 77. This behaviour of the data differs from athletes 3 and 4, where the skewness shifted from left skewed to right skewed for increasing levels of cadence. Outliers become less dense for cadences of 87 and higher. The data were cleaned from outliers as per Equation 3.9 and using Algorithm 2. The data then underwent another visual clean-up and points that are out of sync with the general pattern were removed. The cleaned data set was re-plotted and is presented in Figure 4.8 b). Outliers are now close to the ends of the whiskers. Another notable change is the contraction of the IQR for a cadence of 104. Skewness also improved for most of the cadences with the greatest shifts seen from a cadence of 90 and higher. The overall skewness improved by nearly 70% and 93.3% of the cadence levels became less skew (refer to Table 4.9). The overlapping of the cadence levels for pace values is clear, although the sizes of the boxplots are smaller than for the other athletes. The cadences between 79 and 84 are linked to 5 min/km before the clean-up and remains the same for after the clean-up. The spread of cadence for a pace value of 3 min/km reduced from a range between 92 and 98 to the range from 95.

The interaction terms from the ICM for the analyses before the clean-up are shown in Table 4.7. All the terms are significant on an α-value of 0.05. The generated function from the log-transformed ICM is as follows:

$$Y = e^{\beta_0+\beta_1 x+\beta_{12}x_1x_2+\beta_{13}x_1x_3} \tag{4.3}$$

Both road racing and track training re-enforces the slope of cadence, that is both increase the negativity of the slope of the fitted line and the pace becomes faster for every unit increase in cadence. Interesting to see, and it is not apparent from the scatter plots, that road racing interaction with cadence is actually slightly more re-enforcing than track training. Running on a track, the pace increases with -0.0406 units per unit increase in cadence. The pace increases with -0.0408 units per unit increase in cadence for road racing. This somewhat unexpected finding may well reveal some information to the athlete: he is able to translate the pace achieved during track training into his performances during a race or generate even faster paces during a race.

Table 4.7: Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 2 (original data)

	Estimated parameter	P-value
Intercept (β_0)	3.09650	0.00000
Cadence (β_1)	-0.01799	0.00000
Road race	1.85123	0.00000
Track training	1.84980	0.00000
Cadence:road race (β_{12})	-0.02284	0.00000
Cadence:track training (β_{13})	-0.02266	0.00000

Table 4.8 contains the interaction coefficients for the ICM after the data had been cleaned. Track training is now a stronger re-enforcer of pace than road racing, albeit only by 0.0035. This small difference may still be indicative to the runner that the pace he achieves on the track is being translated into increased pace achieved during road racing.

Table 4.8: Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 2 (cleaned data)

	Estimated parameter	P-value
Intercept (β_0)	3.24313	0.00000
Cadence (β_1)	-0.01989	0.00000
Road race	1.78235	0.00000
Track training	1.81575	0.00000
Cadence:road race (β_{12})	-0.02194	0.00000
Cadence:track training(β_{13})	-0.02229	0.00000

Table 4.9 contains the summary results from the regressions before and after the data had been cleaned (the R^2-values are the adjusted R^2). Figure 4.9 contains the fitted log-transformed lines and the associated residual plots for the models before and after the data clean-up.

Both models showed significant improvement in the R_a^2, with the cubic polynomial function benefiting the most. However, the log-transformed function from the ICM after the data clean-up has the highest R_a^2-value and at 0.752 shows a strong capability of the

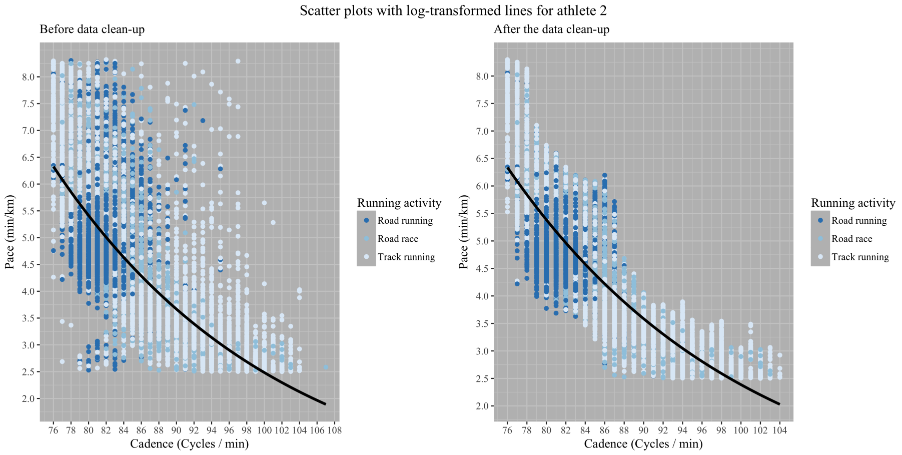
Table 4.9: Summary from the GLM for the effect of cadence and running activity on pace for athlete 2

Variable	Value
R_a^2 (cubic polynomial)	0.563
R_a^2 (cubic polynomial on cleaned data)	0.673
Change in R_a^2 (%) for cubic polynomes	19.428
R_a^2 (log transform)	0.640
R_a^2 (log transform on cleaned data)	0.732
R_a^2 on ICM (log transform)	0.660
R_a^2on ICM (log transform on cleaned data)	0.752
Change in ICM's R_a^2 (%) for log transforms	13.858
Change in overall skewness (%)	69.989
Total case improvements in skewness (%)	93.333

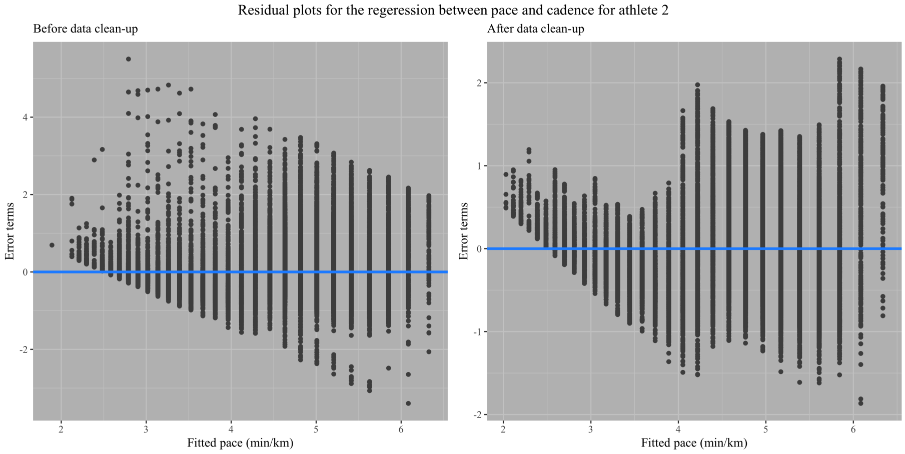
model to represent the data. The fitted line represents the overall log-transform of the data (not subset by running activity) and serves as an illustration of the general trend. This line's R_a^2-value is also stronger than the polynomials' R_a^2,-values both before and after the clean-up. The fitted lines do extend beyond the lowest values for pace though and fail to pick up the plateauing effect the data is exhibiting near the end. The residual plots for the log-transformed models show the errors after the back-transformation of the fitted values. Before the clean-up, the errors move along a downward path, albeit widely spread out. After the clean-up the residual plot forms a curvi-linear pattern. The spread of the error terms around the 0 line remains extensive but the range is less than the errors for the model based on the original data.

4.4 Case study D: the heart-rate runner

Runner 5 continuously monitors his heart rate via a chest-strap and stays within certain pre-determined heart rate limits irrespective of his pace. He is an 8-hour Comrades marathon finisher with a couple of Bill-Rowan medals. The running activities for this athlete are road running, trail running and road racing. A total of 760.8 km from 46 199 data points of running data were collected for this athlete. The mean time between data points is 5.12 seconds. Figure 4.10 shows the data points for pace and cadence subset by colour into the three running activities before the data clean-up. The horizontal spread of road racing is less than both road running and trail running and the data is more condensed. Faster paces (below 3 min/km) are observed than for road racing at cadences between 90 and 94, with some scattered faster instances below 3 min/km for trail running between cadences of 86 up to a 100. Less slower paces (above 7.5 min/km) are observed for road racing than for road running and trail running. The overall movement of the road running data is diagonally down from left to right, with the data clustering as indicated.



(a) The log-transformed fitted lines before and after the data clean-up for athlete 2.



(b) Residual plots

Figure 4.9: Fitted lines and residual plots before and after the data clean-up for athlete 2.

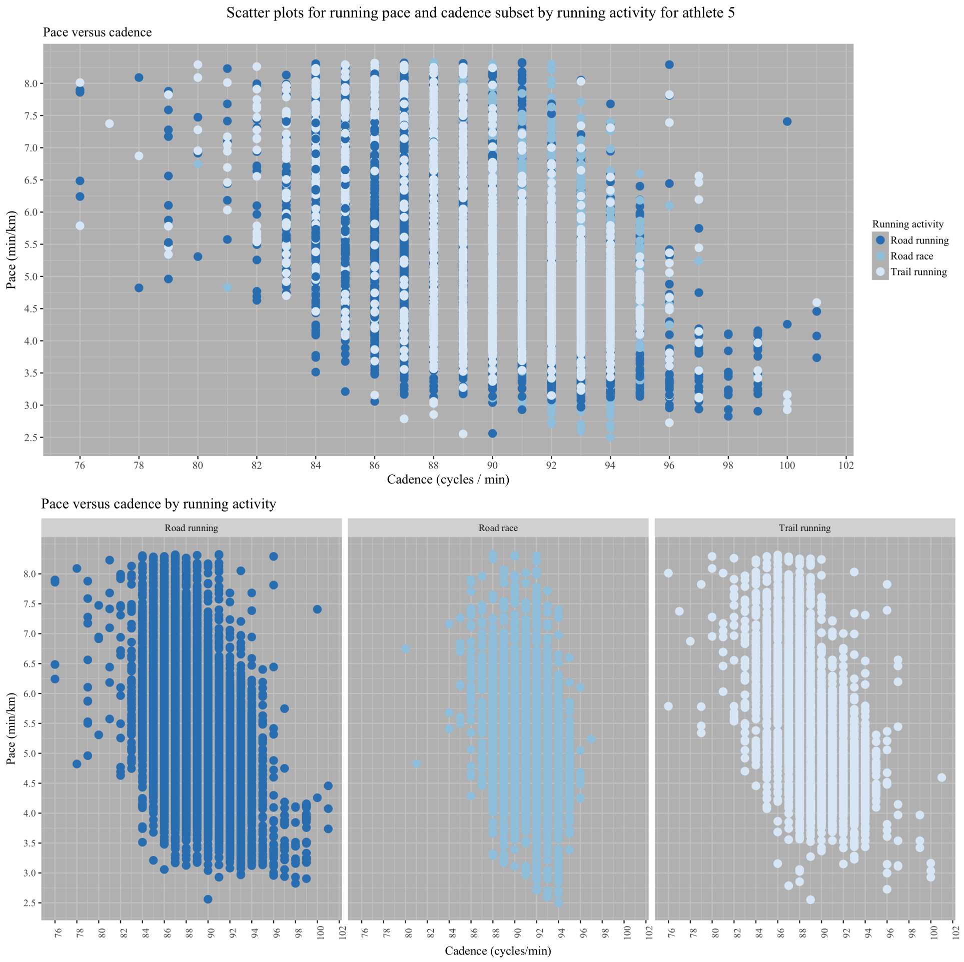


Figure 4.10: The raw data for each observed data point for athlete 5

The trail running data follows much the same pattern, with the major difference seen in the scarcity of data at the cadences of 96 and higher.

The boxplots for pace per cadence level before and after the data clean-up are shown in Figure 4.11. The general trend for the IQRs is unstable for the lower cadences with the trend first moving downward up to a cadence of 79, peaking at a cadence of 80 before it continues downward to a cadence of 98 before it fluctuates again. The direction of skewness per cadence level also changes hands often for the cadences up to 82, after which it remains mainly skewed to the right. As an exception, the cadence levels of 94 and 98 seem to be symmetrical. The data were cleaned from outliers as per Equation 3.9 and using Algorithm 2. The data then underwent another visual clean-up and points that are out of sync with the general pattern were removed. The cleaned data set was re-plotted and is presented in Figure 4.11 b).

The single record captured for a cadence of 77 was removed during the clean-up. The fluctuating pattern continues in the cleaned data set, however, the size of the IQRs for the upper cadences have been reduced. The spread of cadence levels for a pace of 5 min/km is the most apparent, with eight cadences being linked to 5 min/km before and after the

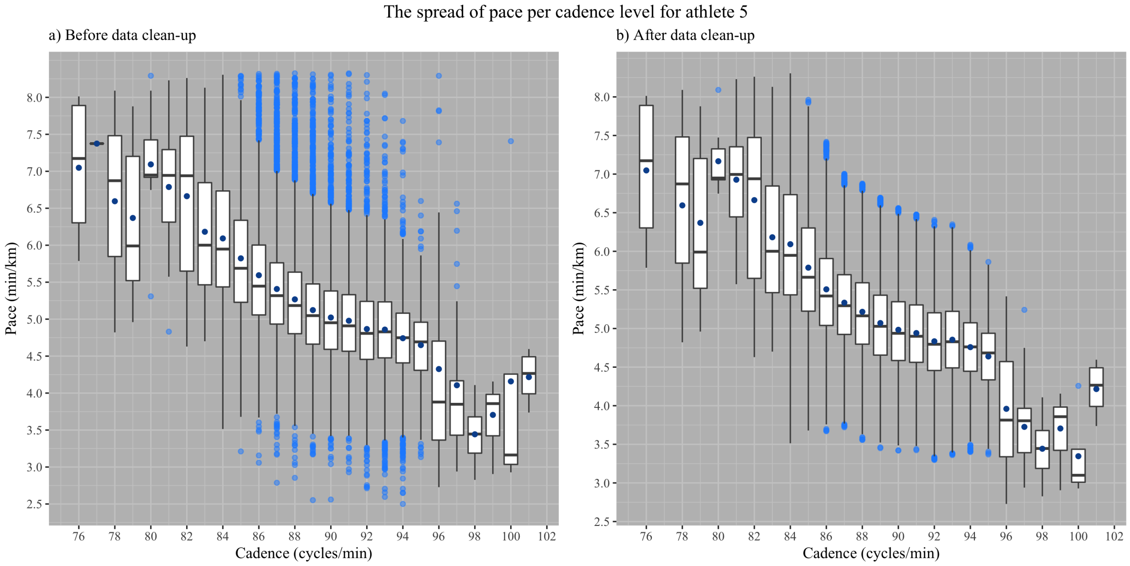


Figure 4.11: The spread of pace across cadence for athlete 5

clean-up. For a faster pace of 3.5 min/km, five cadences are linked to this pace before the clean-up with four cadence levels covering this pace after the clean-up. The overall skewness of the boxplots improved, especially in the ranges of 85 and 96. The skewness for a cadence of 97 reversed itself, going from skew to the right to skew to the left. The overall skewness of the pace per cadence spread improved by nearly 35% with 56% of the cadence levels became less skew (refer to Table 4.12). These figures are much lower than the improvements in skewness seen for the other athletes. This might be an indication that the runner does not really change his cadence and prefers to increase his step length to increase pace.

The back-transformed function from the ICMs with the interaction coefficients for this athlete is as follows:

$$Y = e^{\beta_0+\beta_1 x+\beta_{12} x_1 x_2+\beta_{13} x_1 x_3} \tag{4.4}$$

The interaction terms from the ICM for the analyses before the clean-up are shown in Table 4.10. All the terms are significant on a α-value of 0.05. Contrary to what was expected when analysing the scatter plots, the interaction coefficients suggest that road racing has an interfering effect on cadence whereas trail running has a re-enforcement effect. The linear slope component of the exponential function becomes -0.01661 for road racing and -0.03675 for trail running. That is, pace increase by -0.01661 units per increase in cadence for road racing and -0.03675 units for every increase in cadence during trail running.

Table 4.11 contains the interaction coefficients for the log-transformed ICM after the

Table 4.10: Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 5 (original data)

	Estimated parameter	P-value
Intercept (β_0)	3.62878	0.00000
Cadence (β_1)	-0.02248	0.00000
Road race	-0.51197	0.00000
Trail running	1.31960	0.00000
Cadence:road race (β_{12})	0.00587	0.00000
Cadence:trail running (β_{13})	-0.01427	0.00000

data had been cleaned. Still, road racing has an interfering effect and trail running has a re-enforcement effect, albeit less so than in the original data set. The reason for the adverse finding may be hidden in the density of the data, or by what measures road racing is represented when compared with trail running at the fastest observed paces. Before the data clean-up, paces faster than 4 min/km accounts for 6.643% during trail running and only 2.082% during road racing. After the data clean-up, paces faster than 4 min/km accounts for 6.889% during trail running and 1.662% during road racing. The "pull" towards faster paces is stronger in trail running than in road racing, despite the athlete being able to achieve paces faster than 3 min/km more often during road racing than trail running. The coefficients for cadence and the interaction term for cadence:trail are both less negative, whereas the interaction between cadence and road racing is more positive. The coefficients therefore moved in a positive direction. Pace increases with -0.0151 per increase in cadence for road racing and -0.0383 for every increase in cadence during trail running.

Table 4.11: Estimated parameters for the log-transformed interaction model from cadence and running activity on pace for athlete 5 (cleaned data)

	Estimated parameter	P-value
Intercept (β_0)	3.53891	0.00000
Cadence (β_1)	-0.02153	0.00000
Road race	-0.56471	0.00000
Trail running	1.20018	0.00000
Cadence:road race (β_{12})	0.00643	0.00000
Cadence:trail running (β_{13})	-0.01317	0.00000

Table 4.12 contains the summary results from the regressions before and after the data had been cleaned. Figure 4.12 shows the fitted log-transformed models and the associated residual plots before and after the data clean-up. The line represents the general movement of the data without any interaction. The output from the regression for this athlete differs greatly from the output for the other athletes. The R_a^2-values are all below 0.2 and the models therefore show poor strength in explaining the variability in the data. The model indicates that cadence and running activities are not sufficient

to explain his pace. It is possible that he makes use of stride length to generate pace and prefer to maintain a more constant cadence. This type of behaviour links with his performance philosophy to pace his run based on his heart rate, as a higher cadence comes at a metabolic costs to the runner (Heiderscheit et al., 2011).

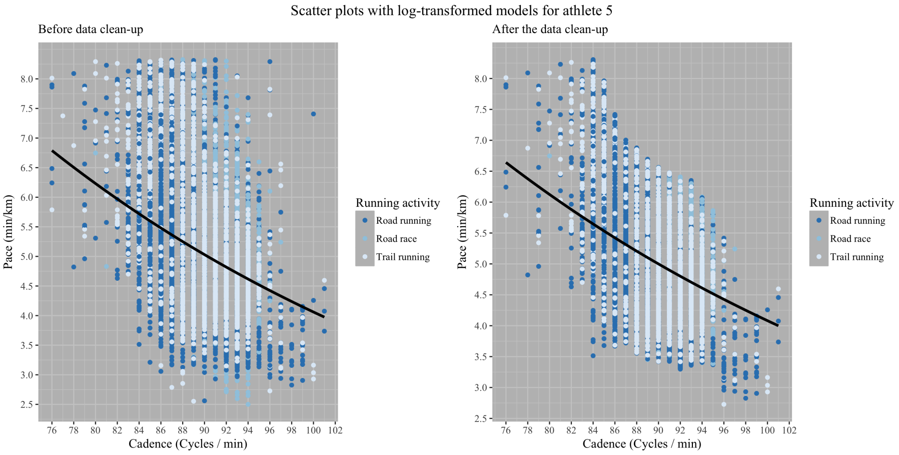
The best model fit is the log-transform ICM after the clean-up. The pattern of the residuals changes after the clean-up and is significantly different than the pattern observed before with the amplitude of the errors having been reduced. The residual pattern is curvi-linear with the widest error spread at a fitted value of roughly 5.6 min/km.

Table 4.12: Summary from the GLMs for the effect of cadence and running activity on pace for athlete 5

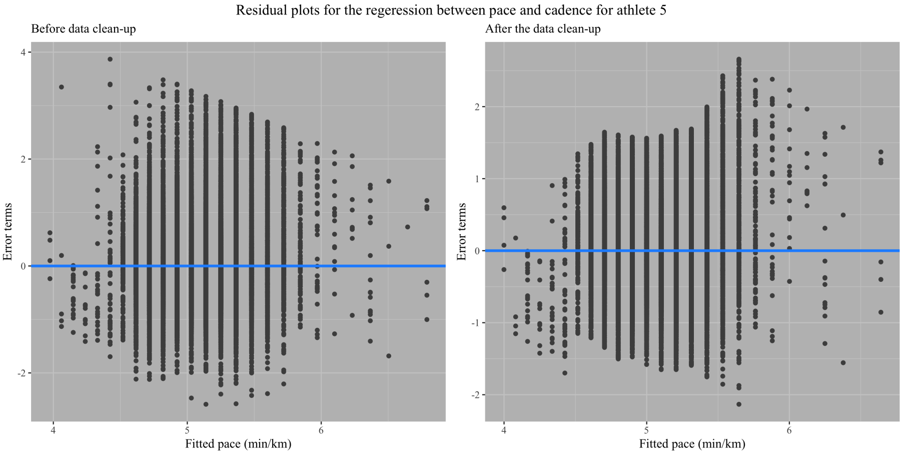
Variable	Value
R_a^2 (cubic polynomial)	0.130
R_a^2 (cubic polynomial on cleaned data)	0.153
Change in R_a^2(%) for cubic polynomes	17.640
R_a^2(log transform)	0.127
R_a^2 (log transform on cleaned data)	0.141
R_a^2 on ICM (log transform)	0.145
R_a^2 on ICM (log transform on cleaned data)	0.155
Change in ICM's R_a^2 (%) for log transforms	6.919
Change in overall skewness (%)	34.888
Total case improvements in skewness (%)	56

4.5 Conclusion

The main finding from these models are the surprising extent of the skewness in the data for pace per cadence. A reasonable explanation behind the skewness itself and the shift in skewness (from left to right or vice-verse) across the cadence levels might be that the algorithm responsible for the calculation of cadence does not respond fast enough when the pace is increased or decreased. By implication, a runner may have accelerated his pace on a higher cadence level but the algorithm has not yet adapted to capture the change in cadence. In truth he is running at a higher cadence but the lower cadence is still being captured. The faster pace then gets associated with the lower cadence level, resulting in the skewness to the left of the lower cadences. The same concept applies to the higher cadences: a runner have already decelerated to a slower pace from a high cadence to a lower cadence but the algorithm is yet to adapt. This leads to the skewness to the right for the higher cadences. The skewness and the decrease thereof after the clean-up procedures demonstrated the negative effect that the outliers might have on any regression model: the data itself is not a full representation of the truth and any prediction or explanatory function will be less accurate. Pace values are being linked to higher or lower cadence



(a) The log-transformed fitted lines before and after the data clean-up.



(b) Residual plots

Figure 4.12: Fitted lines and residual plots before and after the data clean-up for athlete 5.

values than that was actually achieved in the field and misguide the direction of the fitted line.

It was apparent that the IQR gave away the general behavioural pattern of pace for increasing cadence levels. Finding the pattern or trend in the scatter plots alone proved to be challenging and inconclusive, more so before the clean-up. The overlapping of the IQRs also adds to the challenge to fit accurate values per cadence level, as a range of cadences are associated with the same pace value. Although the clean-up of the data did reduce this spread of cadence levels for pace values, most of the reduction occurred at the faster paces.

Despite the improvements in the models' adequacy after the clean-ups, the R_a^2-values from the log-transformed models show mild to moderate strength in its ability to capture the variability in the data in three of the case studies. The ICMs from athlete 5 (case study D) showed to be incompetent to explain the variability of pace when cadence and running activity were considered. It is expected that the model for this athlete would require the step length as another variable to improve the model's adequacy. Interactions were distinguishable between the different running activities, with some being coherent with running logic and others not so much. The behaviour of the errors in the residual plots proved that the data are not normal and the margin of error changed throughout the fitted values. The relationships between the response and the standard error are heterogeneous for all the models. The margin of error changes and fluctuates across the fitted values of pace and are not homoscedastic around the 0 line. The pattern of the residual plots change after the clean-up, mostly from a downward linear pattern to a curvi-linear or wave-like pattern. Most of the errors seem to become larger for the middle range of pace and decrease again towards the faster and slower paces. This behaviour of the error terms may be linked to the wide ranges of the cadence associated with pace values, particularly in the middle ranges of cadence (refer to the boxplots). The large ranges added to the complexity of the model and its ability to fit accurate pace values per cadence level. The log-transformation of the response is justified, but other transformations should be considered in future work. There might also still be missing variables not accounted for that limit the capabilities of the models but alter the performance of the runner. Examples are:

1. Changes in temperature or time of the day at which the running activity took place. The body responds differently to colder and higher temperatures which impacts performance.

2. Total activity of the day, which may have lead to pre-mature fatigue and subsequent poorer performances.

3. Fatigue linked to unrelated running factors (such as poor sleep, an insufficient diet, general life-stress etc.).
4. The running execution approach of the athlete, as demonstrated by athlete 5 (case study D). Inclusion of his heart rate as an explanatory variable might lead to some interesting results and be able to improve the model's adequacy.

Stride length is not included in the data set and therefore was not be added to the model. It is a reasonable assumption that stride length will improve model adequacy, as it is part of the equation to calculate running speed.

Much more work is needed to present a accurate and reliable estimate of pace from cadence alone. Data from the device need verification and validation by evaluating it against experimental studies in the real world. Further work on algorithms to clear the data from inconsistencies is required. This way outliers and incorrect data can be removed from the data set before it is stored in the GPS container files. The following avenues are to be considered for further exploration on the data and the modeling of running form:

1. Documentation and recording of the missing or other explanatory variables to incorporate them into the interaction models.
2. Capturing stride length as part of the data set.
3. Instead of providing the cadence for each instance of the data set, it will perhaps be better to simply provide the total step count up to the captured instance. Data can then be re-worked to calculate the stride frequency per passing time unit and deliver a more accurate stride frequency to the data set.

Another approach to aggregation may also be considered, such as modeling only on the movements of IQRs for cadence levels per running activity as a unit in stead of working with the individual data points or instances. In general the current model is considered a good start and parsimonious enough to form the basis of a model to capture and explain the interaction between cadence and running activity (and by implication running surface) and its overall effect on the pace that can be generated.

Chapter 5

Harnessing the big data from fitness trackers to visually analyse overground running [1]

The techniques of data mining and in particular data visualisation on a large scale to identify patterns, outliers and associations have the potential to shift the boundaries of graded running analyses and research across the spectrum in the running community. This chapter analysed the plausibility of using the large scale spatio-temporal data generated from a running watch to visually analyse graded running form. Research found in the literature is mainly executed under controlled conditions with well-trained athletes who were asked to control their pace and cadence to some degree during UR and DR within the time frame set for the execution of the research. On the contrary, this research focused on overground running covering LR, UR and DR that represent the athlete's response to conditions in the real world over an extended period of time. The research environment is therefore uncontrolled and the runner is free to decide on running pace and routes. Graded running is defined as running on either a level surface (0% grade), uphill (positive grades) or downhill (negative grades). Grades are calculated as per Equation 3.6. The gradient of the running surface was included as a regressor variable to analyse the interaction between the runner's cadence and his environment and their combined effect on the pace.

The starting point of the analysis is on the cleaned data set inherited from the clean-up procedure from Chapter 4 for the modeling of cadence, the running activity and pace. However, a further clean-up was required to remove outliers related to the grades that were run at per cadence level. The clean-up is based on Equation 3.10 and Algorithm 3. Figures 5.1, 5.4, 5.7 and 5.10 present the scatter plots for grade versus cadence and pace versus grade, both before and after the data clean-ups. The grade versus cadence plots are overlaid with a directional colour scale for pace and the pace versus grade plots are overlaid with a directional colour scale representing the cadence. The intensity of the

---

[1] A modified version of this chapter was communicated to the IEEE Transactions on Big Data.

shades of the colours are used to visually present a third variable (either the change in pace or cadence) for the 2D versions of the visual analysis.

The interactions between grade and cadence on pace for the visualised outputs were modeled as a multivariate regression model using a parametric second degree polynomial function and a non-parametric TPS model built on the cleaned data set. Third degree interactions polynomials were initially fitted in order to catch the expected curvy surface, but the fitted values became extreme due to the cubing of already large variables. The cubic polynomials were therefore abandoned. The output from these models presented the 3D version for the visual analysis for overground running.

The interaction effects between cadence and slopes were analysed using a multivariate regression ICM with pace as the response variable (log-transformed) and cadence and the type of slope as the independent variables. The types of slopes are level (LR), downhill (DR) and uphill (UR). The ANOVA showed significant changes in pace for between the types of slopes for all four athletes.

5.1 Case study A: the semi-professional all-rounder

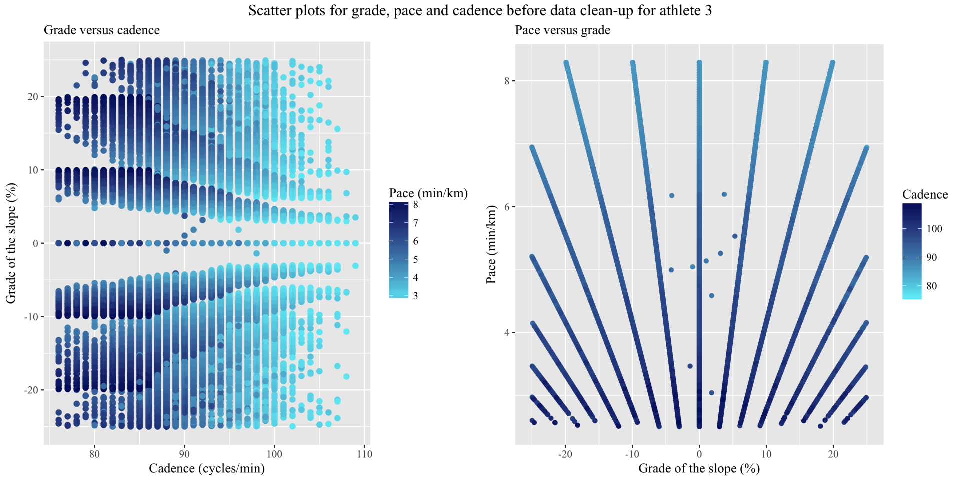
Table 5.1 contains the correlation coefficients between cadence and pace, grade and cadence and grade and pace before and after the data clean-up. The correlation coefficients serve as a starting point to evaluate the linearity between the variables and the possible environmental impact that surface gradient has on both cadence and pace.

Table 5.1: Correlation coefficients between cadence, grade and pace for athlete 3

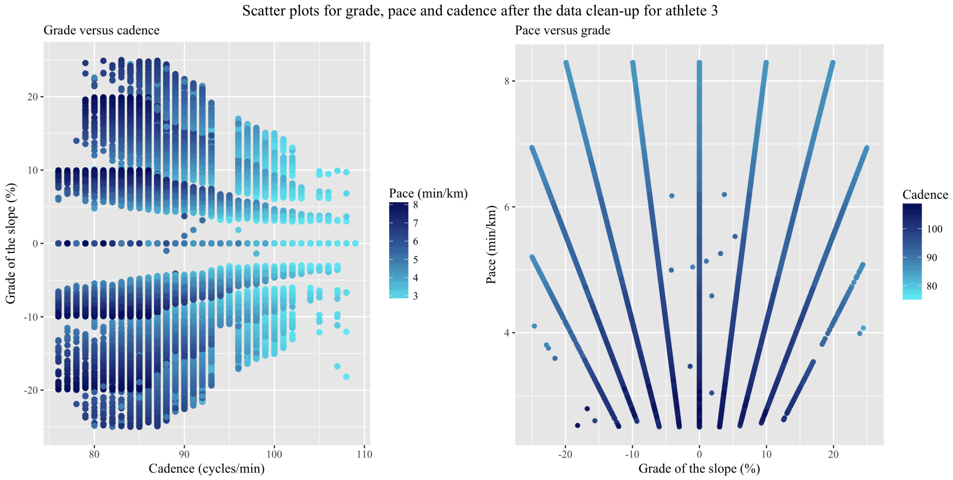
	Data status	Cadence - pace	Grade - cadence	Grade - pace
1	Original	-0.750	-0.007	0.056
2	Cleaned	-0.746	-0.008	0.060

In both sets of data the correlations between grade and cadence and grade and pace are weak to almost non-existent. It may be that neither the athlete's cadence nor his pace is influenced by changes in the slope of the running surface. There is a strong negative association between cadence and pace, indicating that pace becomes faster as the cadence increases.

The colour intensity for pace in the grade versus cadence plots in both Figures 5.1 a) and b) changes from dark to lighter shades indicating that the pace becomes faster at a higher cadence. This is in accordance with literature that an increase in cadence should preferably be used to generate faster pace as it lowers the risk to injury Heiderscheit et al. (2011). There are also more light shaded instances nearer the 0% grade line, implicating that the runner achieves faster speeds at a grade near or at 0%. The weak correlation



(a) Before data clean-up



(b) After data clean-up

Figure 5.1: The 2D representations for the combinations of pace, cadence and grade for athlete 3

between gradient and cadence is reflected in the same figure. There does not seem to be a strong visual indication that increases in gradient of a slope results in a higher selected cadence. This poor correlation agrees with the literature which showed inconclusive results pertaining to the relationship between cadence and grade. The graph shows that cadences below 100 are used across almost all grades. What might be of some concern is the stark symmetry in the colour intensity of the data around the 0% grade line in both the cleaned and original data sets. The pace values near or below 3 min/km seems extremely fast and unlikely for uphill running. These faster paces seem to mirror the gradient and pace distribution for the same cadence below the 0% grade line.

The pace versus grade plot from Figure 5.1 present with some structure in the data resulting in straight lines towards the 0% gradient line. Despite the straight line structures in the pace versus gradient plots from Figure 5.1, it still provides some interesting information. It seems that the pace tends towards the 0% grade line, that is pace becomes faster as the road gradient nears 0%. This corresponds to the weak association between pace and gradient. Pace slows with the absolute changes in slope towards the 0% line, and not in a uniform-direction from negative to positive.

The investigation into the spread of gradient surrounding each cadence level supports the symmetrical structure observed in the scatter plots. Figure 5.2 shows the boxplots for the grade per cadence level before and after the data had been cleared from outliers. The data seems to be evenly spread around the median which in most cases is the 0% grade line, however some of median values coincide with either 25^{th} or the 75^{th} percentiles. This occurs when the exact same value is recorded more than once in the data, again pointing to the possibility that the altitude changes are captured in set interval per unit distance moving forward. The outliers are identified as the blue points beyond the whiskers of the boxplot. The pattern of the whiskers of the boxplot shows an ebb and flow of outward and inward funneling, although the IQRs seem to taper towards the end of the cadence scale. This might imply that the runner prefers to use the higher cadences for grades closer to 0%.

The data underwent a second cleaning procedure, this time removing the outliers based on gradient. All data points outside of the upper and lower limits of grade per cadence level (from Equation 3.10) were removed and the data re-plotted. The behaviour of the IQRs did not really change after the outliers were removed, probably due to the underlying symmetry in the gradient data. Nonetheless, the grade versus cadence plot in Figure 5.1 b) may now reveal some new information. The tapering of the range of grade across the cadence levels is more prominent. There are some explanations behind this observation. The occurrence of cadences above 96 may be becoming scarcer or that these high cadences are truly not used as often during DR or UR. The same explanation

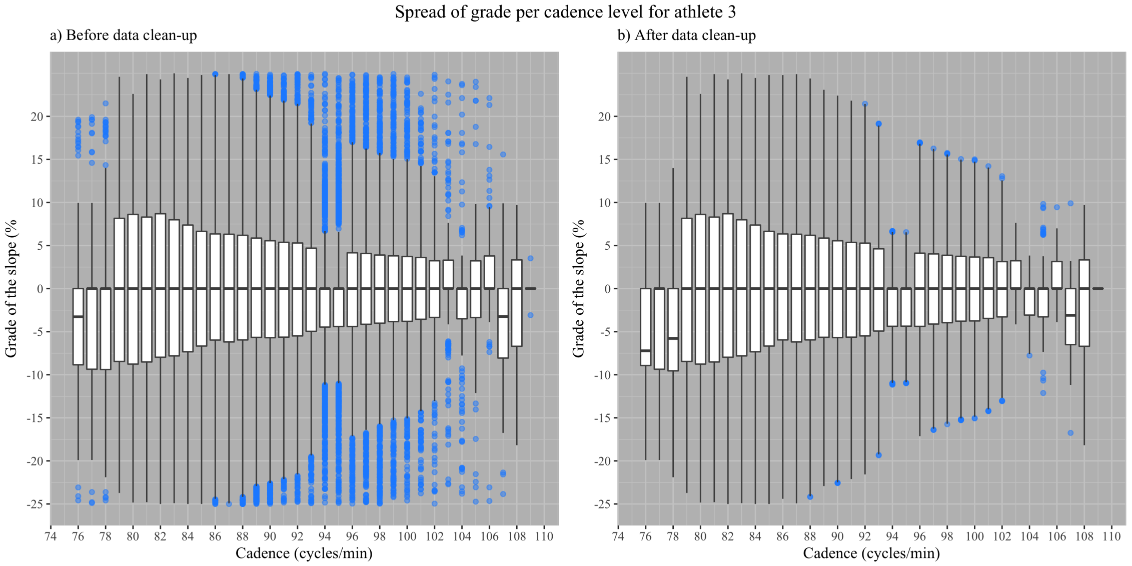


Figure 5.2: The spread of grade across cadence

behind the skewness of the pace around each cadence level is presented for this reasoning. The athlete may have already changed their cadence when running at the instantaneous grade, but the algorithm to change the cadence reading is yet to adapt to the true level.

Table 5.2 provides an overview of the pace that is achieved during DR, UR and LR after the data had been cleaned. The runner achieves his fastest mean pace during DR and slowest pace during UR. The mean pace for LR lies in between. More than 38% of the data points are linked to LR with the UR and DR taking a bit more than 30% of the total data points. This may also explain the symmetry of the scatter plots, as the time spent going up and the time spent going down is almost equal in magnitude. The runner's routes' geographical compositions may be equally distributed between inclining and declining surfaces. In essence, "what goes up must come down": a route that starts and ends at the same point should have a 0 meters netto elevation, owing to the almost equal amount of time spent running uphill and downhill.

Table 5.2: Descriptive statistics for the effect of gradient on pace for athlete 3 (pace in min/km)

Variable	Value
Mean pace (DR)	5.098
St.dev pace (DR)	0.952
Mean pace (UR)	5.271
St.dev pace (UR)	0.966
Mean pace (LR)	5.193
St.dev pace (LR)	0.953
Data points % (DR)	30.223
Data points % (UR)	30.917
Data points % (LR)	38.860

The results from the log-transformed multiple linear regression ICM ($R_a^2 = 0.614$) to measure the re-enforcement and interfering effects of slopes on cadence and pace are presented in Table 5.3. Both interactions with DR and UR are re-enforcing cadence on pace, albeit UR has a stronger effect than DR. These re-enforcement actions do not correspond with the pattern of mean pace values observed for the slope levels as in Table 5.2. It was expected that DR would have a stronger re-enforcement effect on cadence than UR.

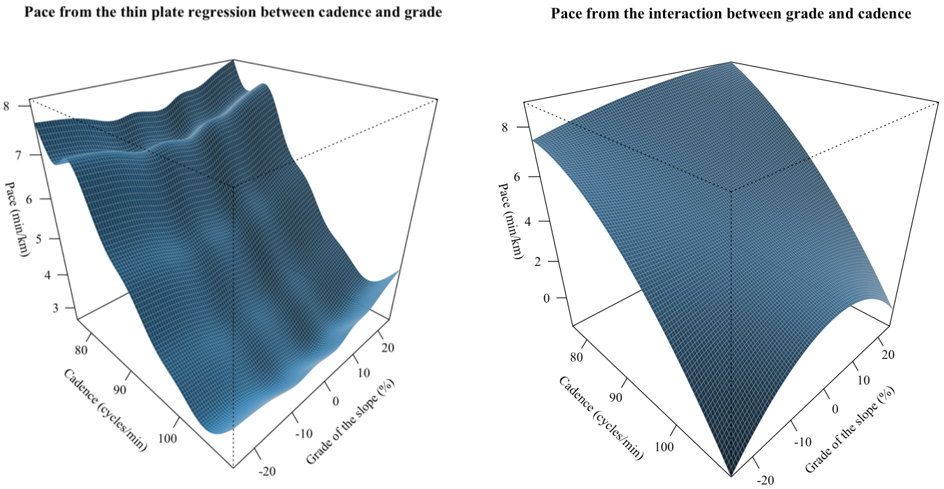
Table 5.3: Estimated parameters for the log-transformed interaction model from cadence and slopes on pace for athlete 3 (cleaned data)

	Estimated parameter	p-value
Intercept (β_0)	5.51586	0.00000
Cadence (β_1)	-0.04334	0.00000
Downhill running	0.04298	0.00092
Uphill running	0.13583	0.00000
Cadence: downhill running (β_{12})	-0.00063	0.00001
Cadence: uphill running (β_{13})	-0.00141	0.00000

Table 5.4 contains the summary from the adequacy measures for the TPS model and the polynomial ICM. The lower AIC for the TPS model after the data has been cleared is lower than the AIC-score for the polynomial ICM and is therefore closer to the truth. For this reason the TPS model is considered to be the model that is a better representation of the interaction between cadence and grade and their combined effect on pace. The 3D outcomes of the two models are shown in Figure 5.3. The TPS is clearly more curvier than the polynomial ICM. The two models seem to contrast each other: where the ICM bulges upward around the 0% grade line, the TPS decreases but maintains a wave-like structure throughout the plate across the range of grades. The symmetry around the 0% grade line of the underlying data structure remains evident in both the models.

Table 5.4: Summary from the thin plate and polynomial interaction models for cadence, grade and pace for athlete 3

Variable	Value
AIC (TPS)	629822.550
R^2 (TPS)	0.590
AIC (TPS) after clean-up	624295.099
R^2 (TPS) after clean-up	0.585
Change in AIC for the TPS	5527.451
AIC (ICM)	640566.488
AIC (ICM) after clean-up	634851.546
Change in AIC for the ICM	5714.942
R_a^2 (ICM)	0.577
R_a^2(ICM) after clean-up	0.571
Change in R_a^2 for the ICM (%)	-0.954


Pace from the thin plate regression between cadence and grade

Pace from the interaction between grade and cadence

(a) The thin plate spline model (b) The polynomial interaction model

Figure 5.3: The 3D curves as a result of the interaction between grade and cadence for athlete 3

Table 5.5 is a direct comparison on the extreme points from the TPS model and the actual data. At face value the actual and fitted values from the TPS model do not seem to reach an agreement. The slowest fitted pace is 0.61 min/km faster than the actual slowest pace. The fitted pace values at the highest and lowest grades also differ by a significant margin. The fitted and actual cadences at the fastest pace differs considerably by three cycles. Cadences above 101 presents only 0.43% of all the data, which may render the model's accuracy at these higher cadences.

Table 5.5: Actual and fitted values from the thin plate model for athlete 3

Variable	Value
Fastest actual pace	2.50
Fastest fitted pace	2.64
Slowest actual pace	8.29
Slowest fitted pace	7.68
Actual pace at min grade	5.21
Fitted pace at min grade	7.21
Actual pace at max grade	6.94
Fitted pace at max grade	7.68
Actual cadence at fastest pace	102
Fitted cadence at fastest pace	105
Actual grade at fastest pace	3.00
Fitted grade at fastest pace	3.19

The models in Figure 5.3 are the 3D representation of the 2D components from Figure 5.1. The scatter plot of grade versus cadence from Figure 5.1 is the top view of the TPS and the plot of pace versus grade is the side view of the TPS. From the 3D image of the TPS model the lowest point on the bent plate (i.e. the fastest pace) is found roughly at a cadence between 100 and 108, with the exact figure given in Table 5.5 as 105. This corresponds to the lighter shading of a fast pace on the grade versus cadence plot in Figure 5.1 b) found between a cadence 100 and 110. The darker shaded instances for cadence in the plot of pace versus grade found below 4 min/km agrees with the cadence of 100 or higher. The colour changes intensity to a lighter shade for higher values of pace and a grade away from 0% which corresponds to the wave-like patterns across the grades in the TPS image.

5.2 Case study B: the trail specialist [2]

Table 5.6 contains the correlation coefficients between cadence and pace, grade and cadence and grade and pace. There is almost no linearity between grade and pace and a weak positive linear relationship between grade and cadence, implying that grade may have some influence on the selected cadence when the athlete is running on slopes. The positive correlation is an indication that the athlete may increase their cadence as grade increases. The negative linearity between cadence and pace is strong.

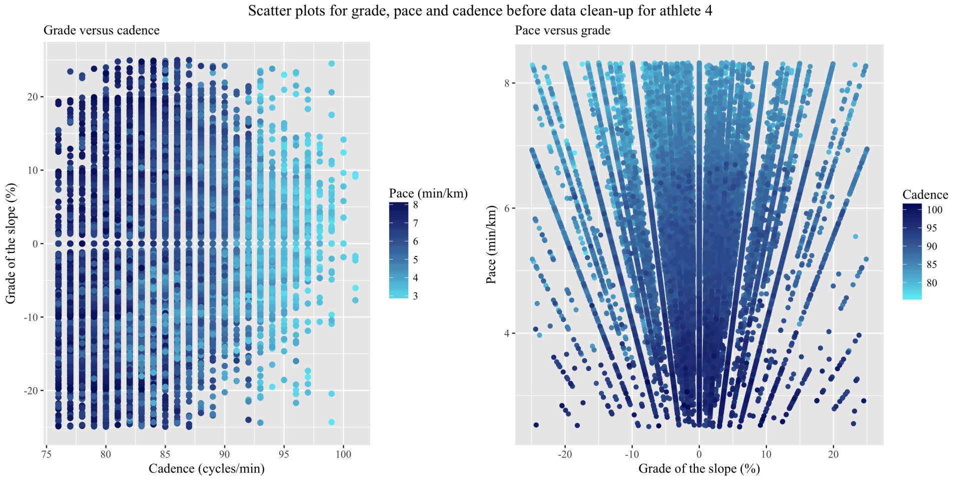
Table 5.6: Correlation coefficients between cadence, grade and pace for athlete 4

Data status	Cadence - pace	Grade - cadence	Grade - pace
Original	-0.772	0.231	-0.038
Cleaned	-0.772	0.291	-0.094

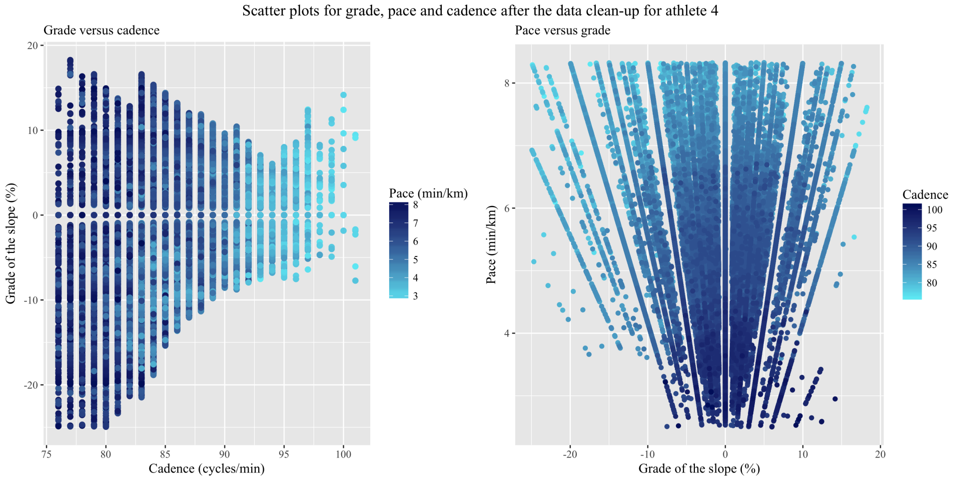
Figure 5.4 contains the 2D graphs used to visually analyse running form on slopes. There does not seem to be any linear relationship between grade and cadence, however, the pattern does start to taper from both the positive and negative grades after a cadence of 87. Cadences up to the point of tapering are used across the entire spectrum of grade. Judging from the colour scale, the pace starts to pick up from a cadence of 90 onwards. The colour also becomes a lighter shade roughly at the point where the tapering pattern starts for increasing values of cadence. The colour change from dark shades to light shades across the cadence spectrum supports the good correlation coefficient between pace and cadence. This is a good indication to the athlete that he might be mainly using cadence to increase his pace. Noticeable is the symmetry in the data around the 0% line, both in scatter of the data points and some in colour intensity. The cluster of lighter shades for negative grades between a cadence of 82 and 87 is not mirrored in the same scope of positive grades. There is some concern regarding the symmetry in the colour intensity of the data points around the 0% line from a cadence of 93 upwards. Running a pace near or below 3 min/km for extreme positive grades seem highly unlikely, but this level of makes sense when going downhill where gravity may be a helping hand towards faster paces.

---

[2] A modified version of this section was communicated to the IEEE Transactions on Big Data.



(a) Before data clean-up



(b) After data clean-up

Figure 5.4: The raw data for the combinations of pace, cadence and grade for athlete 4

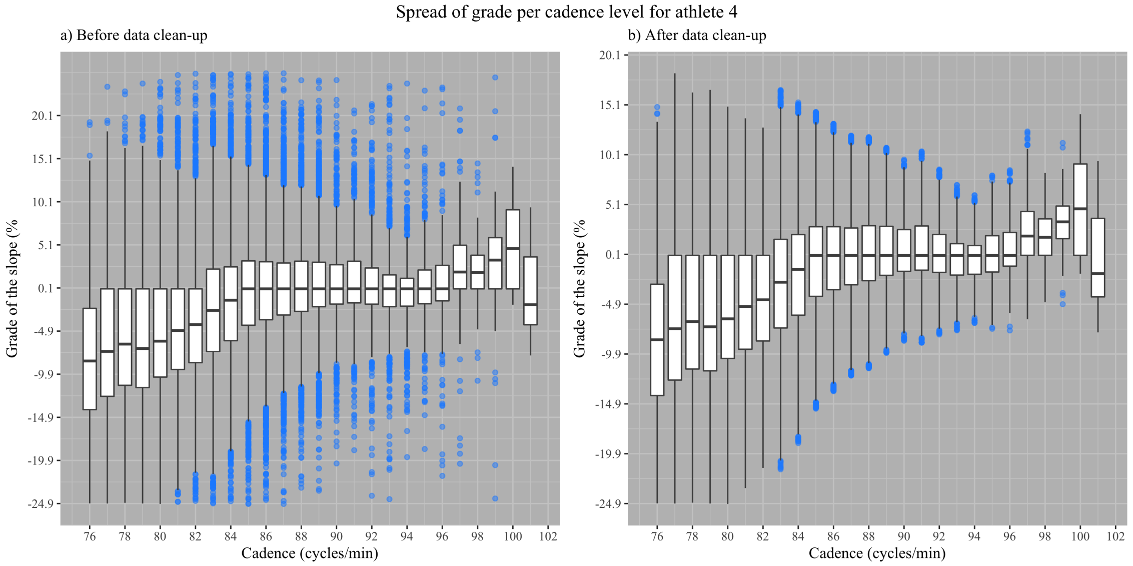


Figure 5.5: The spread of grade across cadence for athlete 4

There is visible structure in the plots for pace versus grade in Figure 5.4. Straight lines have formed which tend towards the 0% line from both positive and negative grades. There is no visual indication of linearity between grades and pace. The symmetry around the 0% line is again apparent. However, the directions of the data may be of significance. The lines tend to the 0% line, indicating that the pace increases as the runner nears level surfaces. It might therefore be that the absolute value of the grade (i.e. the distance from 0%, irrespective whether the grade is negative or positive) has a positive association with pace: the closer the grade to 0%, the faster the pace. The darker shaded points in the fast pace region below 4 min/km corresponds to the lighter shades of pace with increasing cadence levels.

The analyses of spread of the grade per cadence level revealed the pattern in the data. Figure 5.5 presents the grades per cadence level as boxplots. The pattern of the boxplots differ from those seen in case study A from athlete 3. The existing (although weak) correlation between grade and cadence is evident in these graphs. The IQRs shift upward for increasing cadence levels up to a cadence of 85, then stabilise around a grade of 0.1% and shift upward again at a cadence of 100. There are no outliers present for a cadence of 100 and up. The total spread of the boxplot (from upper to lower whisker) funnels inward across the cadence levels for both the original and cleaned data. The clean-up procedure removed almost all the data higher than roughly 17% grades. The behaviour of the IQRs did not show any real change after the clean-up. The scatter plot in Figure 5.4 b) for grade versus cadence now clearly forms a tapering pattern from left to right for increasing cadence levels, but shows scattered points that break the downward tapering pattern from a cadence of 96. This break coincides with the unchanged pattern of the boxplots for cadence of 100 and 101. As there were no outliers for these cadence levels the pattern at these levels in the scatter plot did not change. Evaluating the density of these cadence levels provide some insight. The data on the cadence level of 100 and 101 represent 0.018% and 0.0127% respectively of the total size of the data set. It is a very small sample size when compared to the rest of the cadence levels and may be insufficient to truthfully present the behaviour of the data. The mirroring of the colour scale for the higher cadences around the 0% grade is still evident. The symmetry around the 0% line in the plot for pace versus grade is broken with the maximum grade now below 20%. The tendency of the data towards the 0% is still the same.

From Table 5.7 his mean pace achieved during DR is slower than the pace during UR. This does not correspond to running logic. The behaviour of the data in Figure 5.4 may explain the unexpected means of pace during UR and DR. The mirroring of the colour intensity scale in a) at the higher cadences indicate that the runner achieves the same (or faster) paces during UR as duringDR. The athlete has also spent more of his time running uphill than downhill (45.3% versus 34.46%). The data points for UR outnumber the data points for DR and may therefore be responsible for either a more accurate mean or still be influenced by large outliers. The runner achieves his fastest pace during LR, which corresponds to the tendency of the straight lines towards the 0% gradient in the pace versus grade plots. The faster pace achieved during LR than DR may make sense when reviewing the literature. More energy is required to run up a slope and the pace slows down, but coming down the slope a runner is forced to dissipate energy to facilitate braking and may eventually slow the pace Vernillo et al. (2016).

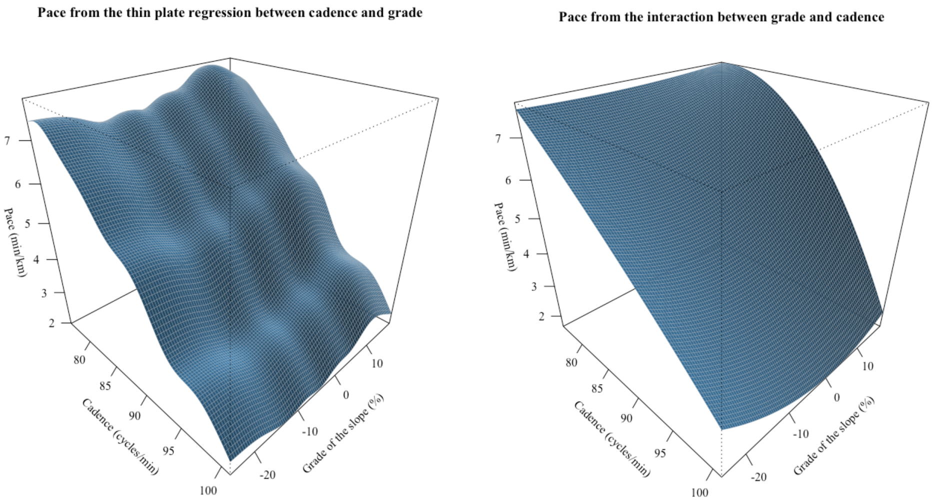
Table 5.7: Descriptive statistics for the effect of gradient on pace for athlete 4

Variable	Value
Mean pace (DR)	5.614
St.dev pace (DR)	1.182
Mean pace (UR)	5.560
St.dev pace (UR)	1.114
Mean pace (LR)	5.285
St.dev pace (LR)	1.205
Data points % (DR)	34.464
Data points % (UR)	45.306
Data points % (LR)	20.230

The results from the log-transformed multiple linear regression ICM ($R_a^2 = 0.628$) to measure the re-enforcement and interfering effects of slopes on cadence and pace are presented in Table 5.8. Both levels of UR and DR interfere with cadence and slows the pace, with DR providing the most interference. This interference corresponds to the mean values of pace obtained for the slope levels in Table 5.7, where the athlete achieves his fastest pace at level running and his slowest pace during DR.

Table 5.8: Estimated parameters for the log-transformed interaction model from cadence and slopes on pace for athlete 4 (cleaned data)

	Estimated parameter	p-value
Intercept (β_0)	5.91014	0.00000
Cadence (β_1)	-0.04861	0.00000
Downhill running	-0.71539	0.00000
Uphill running	-0.11450	0.00252
Cadence: downhill running (β_{12})	0.00786	0.00000
Cadence: uphill running (β_{13})	0.00171	0.00007



(a) The thin plate spline model (b) The polynomial interaction model

Figure 5.6: The 3D curves as a result of the interaction between grade and cadence for athlete 4

The 3D output from the models after the data had been cleaned is shown in Figure 5.6. The summary of the models in terms of their adequacy is shown in Table 5.9. There is a stark difference in the 3D models. The TPS is much more curvier than the bulging polynomial ICM and picks up the varying behaviour of the data seen in Figure 5.4. The grade versus cadence plot in Figure 5.4 b) is the top view and the pace versus grade is the side view of the TPS 3D model. The changes in colour intensity in the grade versus cadence plot corresponds with the downward, but wavy, trend seen in the 3D plate. The ICM reaches its minimum value (i.e. the fastest pace) close to the 0% line (-1.69%), whereas the TPS model reaches its lowest point at -6.05% gradient. The wave-like pattern of the plate across the grade represents the changing cadence levels used across the grade scale. This corresponds to the changes in the colour intensity of the flat 2D pace versus grade plot.

The AIC value for TPS for both before and after the clean-up is lower when compared to the AIC score for the ICM. From the AIC theorem, this implies that the TPS models are closer to the truth than the ICMs. The R^2-values for the TPS model are also slightly higher than the ICM's R^2-values. However, both the models' R^2 values actually decreased somewhat after the data clean-up. The TPS model is the better 3D representation of the interaction between cadence, grade and pace.

Table 5.9: Summary from the thin plate and polynomial interaction models for cadence, grade and pace for athlete 4

Variable	Value
AIC (TPS)	118880.100
R^2 (TPS)	0.637
AIC (TPS) after clean-up	113211.252
R^2 (TPS) after clean-up	0.634
Change in AIC for the TPS	5668.849
AIC (ICM)	120405.544
AIC (ICM) after clean-up	114682.296
Change in AIC for the ICM	5723.248
R_a^2 (ICM)	0.627
R_a^2 (ICM) after clean-up	0.624
Change in R_a^2 for the ICM (%)	-0.479

Table 5.10 shows the direct comparison between the fitted values for the TPS model and the actual values from the data. Pace is given in min/km. The fitted and actual cadences at the fastest pace differ by two cycles. Considering the importance of an accurate cadence representation, a difference of two cycles can be considered as significant. However, it may be that the actual fit is a better presentation as the cadences above 100 accounts for only 0.032% of all the data. There is a substantial difference of 0.46 min/km difference between the slowest fitted and actual pace. The fastest paces also differ by 0.44 min/km. This difference in the fastest pace equates to a difference of 85 meters over a 1-minute interval and might be the difference between winning and losing, or achieving the race goal or not. It must be understood by the athlete that these fitted values are

Table 5.10: Actual and fitted values from the thin plate model for athlete 4

Variable	Value
Fastest actual pace	2.50
Fastest fitted pace	2.06
Slowest actual pace	8.33
Slowest fitted pace	7.87
Actual pace at min grade	6.93
Fitted pace at min grade	7.18
Actual pace at max grade	7.61
Fitted pace at max grade	7.41
Actual cadence at fastest pace	99
Fitted cadence at fastest pace	101
Actual grade at fastest pace	6.01
Fitted grade at fastest pace	-6.05

the expected pace values for a combination of grade and cadence and should not be taken as his ultimate capability. Chapter 6 deals with ultimate and overall capabilities. The actual and fitted grades at the fastest pace are on opposite sides of 0% gradient line. This is an inconsistent finding to present to an athlete.

5.3 Case study C: the Comrades marathoner

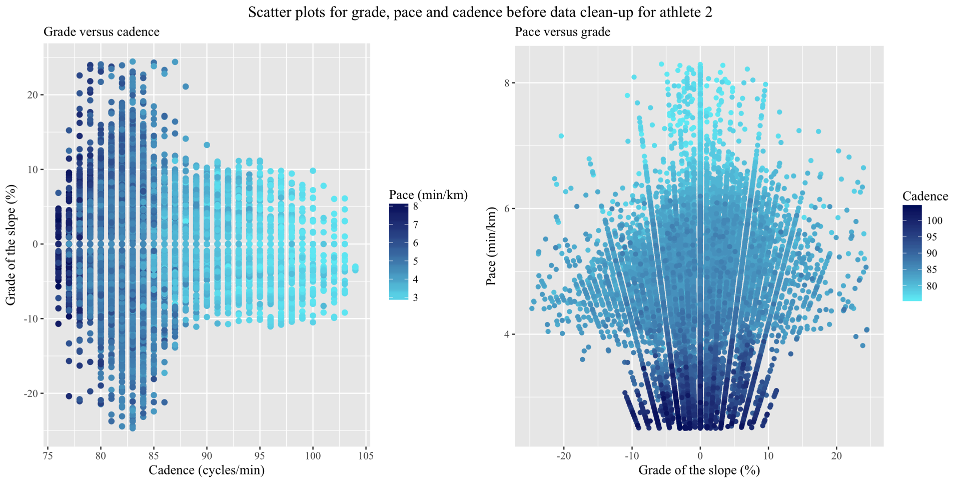
Table 5.11 contains the correlation coefficients between cadence and pace, grade and cadence and grade and pace. All the correlations improved after the data clean-up. There is almost no linearity between grade and cadence and some weak positive correlation between grade and pace, which implies that increases in grade may decrease pace (the runner slows down). There is a strong negative correlation between cadence and pace, indicating that that pace becomes faster (i.e. the numerical value decreases) as cadence increases. This strong correlation is a good sign to the athlete that he uses cadence to run faster.

Table 5.11: Correlation coefficients between cadence, grade and pace for athlete 2

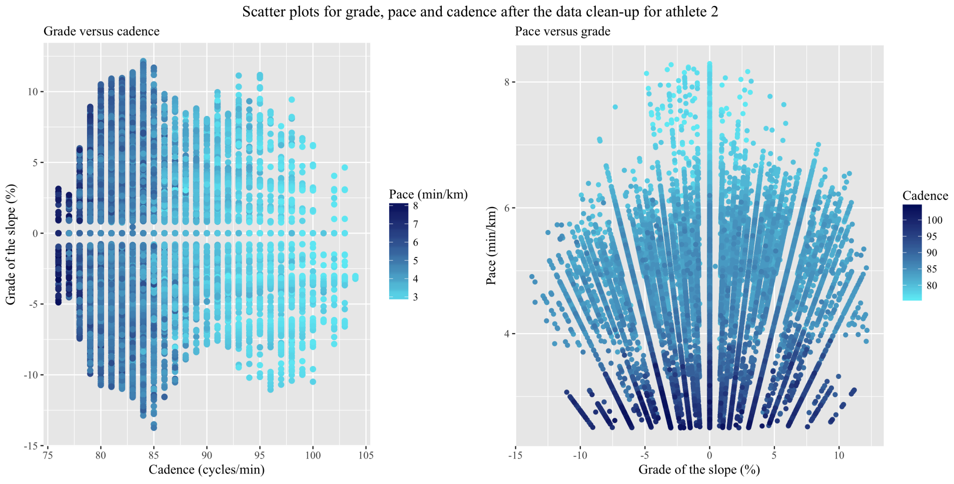
Data status	Cadence - pace	Grade - cadence	Grade - pace
Original	-0.810	-0.025	0.103
Cleaned	-0.811	-0.031	0.109

Figure 5.7 contains the 2D graphs used to visually analyse running form on slopes. In agreement with the high correlation between cadence and pace, the intensity of the colour of the points change to a lighter shade as cadence increases in the grade versus cadence plot. The data funnels outward from the beginning up to a cadence of 83, after which it funnels inward and then remain at a constant spread before a small inward funnel from a cadence of 101. There is symmetry in the data, although the points below the 0% line are generally a lighter shade that those above the 0% line. The same straight line structure from the previous case studies is seen in the pace versus grade plot with scattered data points in-between the lines. The darker shaded points at the faster paces (roughly ≤ 3.75) correspond with cadences of 95 and above. There are almost no data points for grades beyond 15% (positive or negative) and below 3.75 min/km. After the clean-up the grade versus cadence plot now exhibits a wave-like pattern of outward and inward funneling throughout the range of cadence. Data points with grades beyond 15% on either side have been removed. The pace versus grade plot is now less densely populated between the structural lines which makes the structured lines more pronounce.

Figure 5.8 shows the boxplots for the spread of grade across the cadence levels for both original and cleaned data sets. Interesting to note that in both data sets the IQRs



(a) Before data clean-up



(b) After data clean-up

Figure 5.7: The raw data for the combinations of pace, cadence and grade for athlete 2

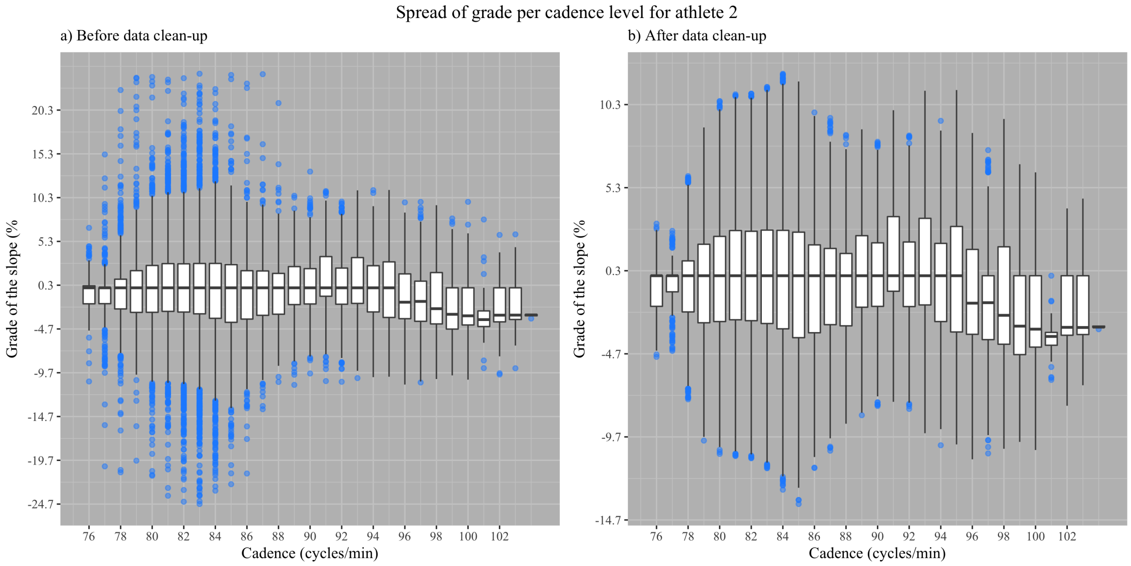


Figure 5.8: The spread of grade across cadence for athlete 2

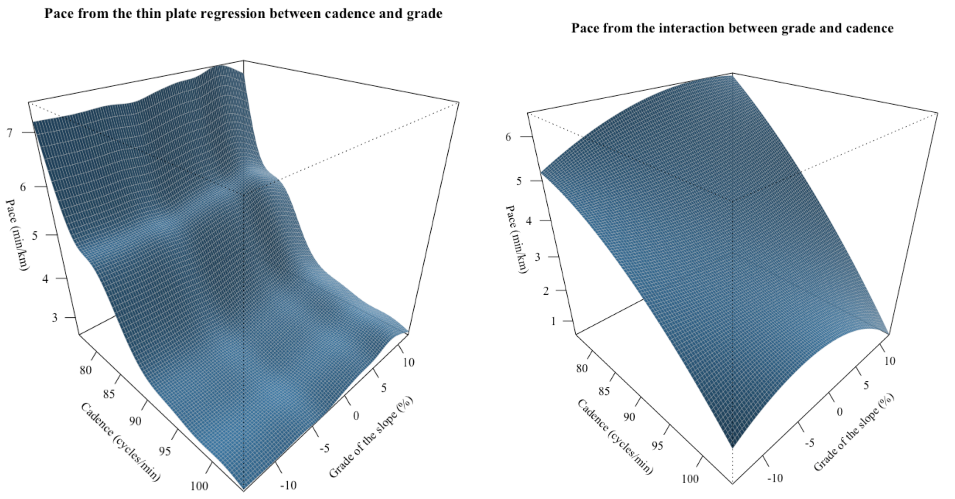
shift below the 0% line at cadences from 96 and upward. This corresponds to the negative (albeit weak) correlation between grade and cadence. The decrease in gradient (i.e. going downhill) is associated with higher cadences. The entire IQR for cadences from 99 and higher are below the 0% line. Although the entire range of cadence is used during DR, it seems that the cadences of 99 and higher are exclusively used for DR.

From Table 5.12 the runner's mean pace at DR is the fastest, followed by LR and the slowest mean pace is achieved for UR. This type of spread of the mean running paces across the type of slope is coherent with running logic. However, this figure might still be biased towards the larger proportion of the time spent at DR compared to LR. Only 18.1% of the data points represent LR and more than double the amount is spent at DR.

Table 5.12: Descriptive statistics for the effect of gradient on pace for athlete 2

Variable	Value
Mean pace (DR)	4.645
St.dev pace (DR)	0.822
Mean pace (UR)	4.863
St.dev pace (UR)	0.839
Mean pace (LR)	4.716
St.dev pace (LR)	1.058
Data points % (DR)	39.441
Data points % (UR)	42.414
Data points % (LR)	18.145

The results from the log-transformed multiple linear regression ICM ($R_a^2 = 0.742$) to measure the re-enforcement and interfering effects of slopes on cadence and pace are



(a) The thin plate spline model

(b) The polynomial interaction model

Figure 5.9: The 3D curves as a result of the interaction between grade and cadence for athlete 2

presented in Table 5.13. It seems that DR actually interferes with cadence (despite having a faster mean pace during DR) and slows the pace down, more so than UR.

Table 5.13: Estimated parameters for the log-transformed interaction model from cadence and slopes on pace for athlete 2 (cleaned data)

	Estimated parameter	p-value
Intercept (β_0)	5.21711	0.00000
Cadence (β_1)	-0.04400	0.00000
Downhill running	-0.60027	0.00000
Uphill running	-0.06929	0.01967
Cadence: downhill running (β_{12})	0.00696	0.00000
Cadence: uphill running (β_{13})	0.00100	0.00496

The 3D output from the TPS model and polynomial ICM after the data had been cleaned is shown in Figure 5.9. There is quite a remarkable difference in the shape of the two models. The summary of the models in terms of their adequacy is shown in Table 5.14. The AIC score is lower for the TPS model than for the ICM with the the R^2-values being higher. There is not much change in the R^2-values for the ICM after the data had been cleaned. The R^2-value for the TPS model actually decreased after the clean-up. The AIC for the ICM reduced more than the AIC for the TPS model. The 3D output from the TPS model shows that the data is not symmetrical around the 0% line, which was difficult to distinguish in the flat 2D version of grade versus cadence. The plate is bent upwards for the combination of cadences above 100 and positive grades. Pace is therefore expected by the model to be slower for UR than for DR, which corresponds with the faster average pace seen in Table 5.12.

Table 5.15 shows the direct comparison between the fitted values for the TPS model

Table 5.14: Summary from the thin plate and polynomial interaction models for cadence, grade and pace for athlete 2

Variable	Value
AIC (TPS)	48792.554
R^2 (TPS)	0.7179
AIC (TPS) after clean-up	46707.209
R^2 (TPS) after clean-up	0.7170
Change in AIC for the TPS	2085.345
AIC (ICM)	53309.972
AIC (ICM) after clean-up	50966.978
Change in AIC for the ICM	2342.994
R_a^2 (ICM)	0.679
R_a^2 (ICM) after clean-up	0.681
Change in R_a^2 for the ICM (%)	0.193

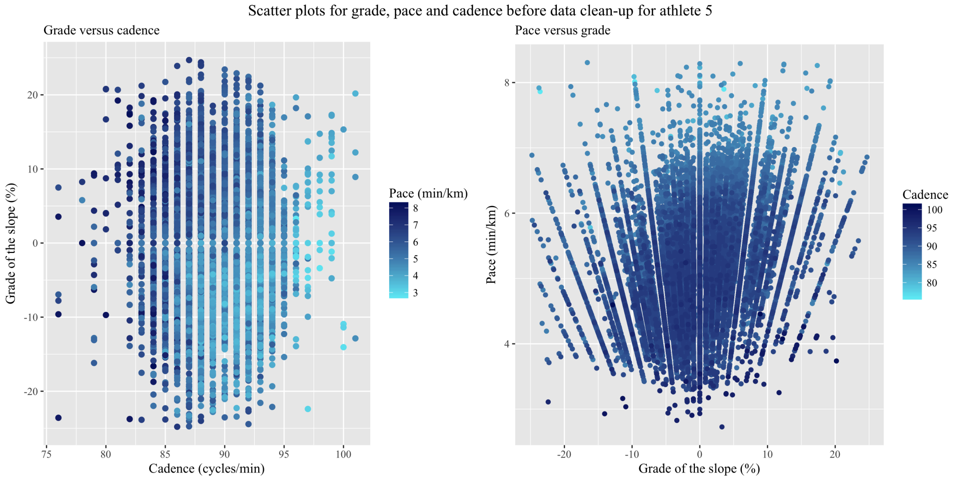
Table 5.15: Actual and fitted values from the thin plate model for athlete 2

Variable	Value
Fastest actual pace	2.50
Fastest fitted pace	2.53
Slowest actual pace	8.29
Slowest fitted pace	7.34
Actual pace at min grade	4.91
Fitted pace at min grade	4.45
Actual pace at max grade	4.05
Fitted pace at max grade	4.75
Actual cadence at fastest pace	102
Fitted cadence at fastest pace	104
Actual grade at fastest pace	-2.25
Fitted grade at fastest pace	-3.22

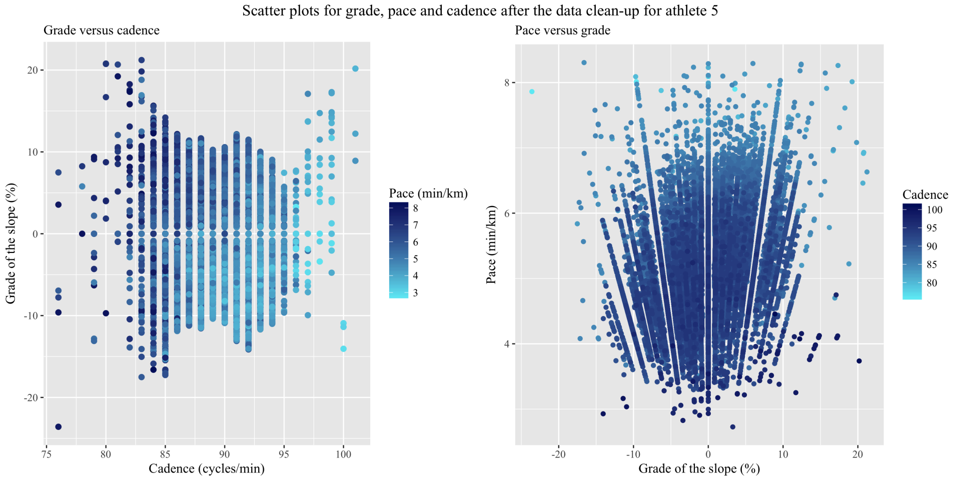
and the actual values from the data. Pace is given in min/km. The fitted cadence at fastest pace (the lowest point in the bent plate) is two cycles more than the actual cadence for the fastest pace. Considering the importance of an accurate cadence representation, a difference of two cycles may prove to be substantial. Cadences equal to or above 102 represent 0.343% of the entire data set, so the model's ability to accurately predict the true cadence at the fastest pace is limited.

5.4 Case study D: the heart-rate runner

Table 5.16 contains the correlation coefficients between cadence and pace, grade and cadence and grade and pace. There is weak negative linearity between cadence and pace, indicating perhaps that the athlete does not use cadence to increase his pace. The linearity between grade and cadence is almost non-existent. There is a weak positive relationship



(a) Before data clean-up



(b) After data clean-up

Figure 5.10: The raw data for the combinations of pace, cadence and grade for athlete 5

between grade and pace, indicating that increases in grade perhaps slows the runner down.

Table 5.16: Correlation coefficients between cadence, grade and pace for athlete 5

Data status	Cadence - pace	Grade - cadence	Grade - pace
Original	-0.378	-0.052	0.337
Cleaned	-0.381	-0.076	0.345

Figure 5.10 a) and b) shows the 2D graphs representing running form on slopes before and after the second data clean-up respectively.

In the grade versus cadence plots before the data clean-up, the colour shade for pace becomes lighter over the range of cadence, but not with the same intensity as with the other athletes. This pattern of the slighter change in the intensity agrees with the weak correlation between cadence and pace. What is more apparent from this plot is the

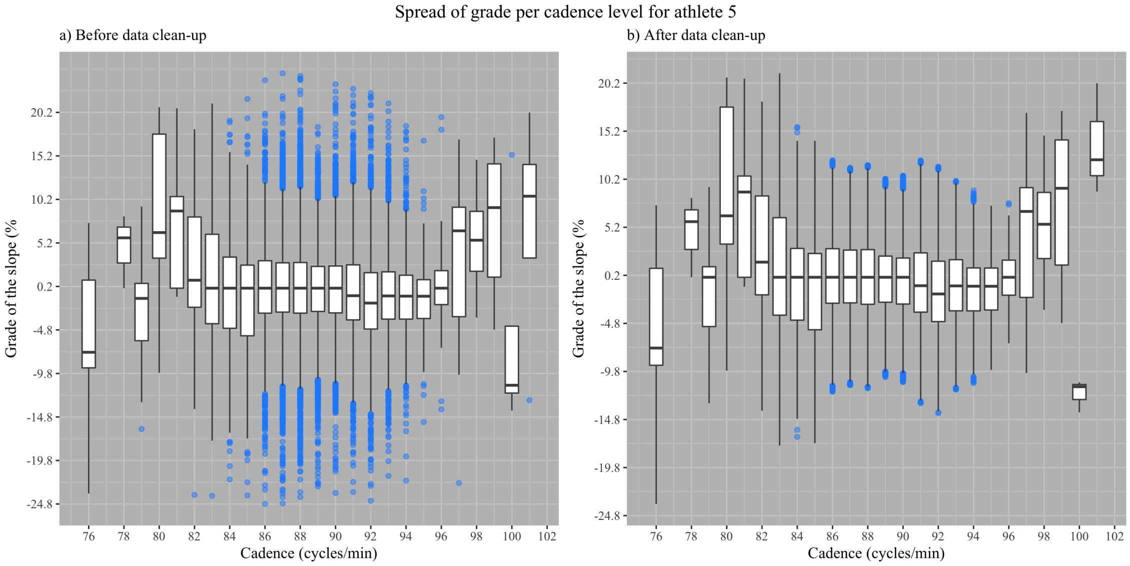


Figure 5.11: The spread of grade across cadence for athlete 5

majority of the lighter shaded points are beneath the 0 line, which corresponds to the existing (albeit weak) correlation between grade and pace: the athlete achieves faster paces during DR than UR for the range of cadence between 87 and 95. The pattern of the colour shades is not symmetrical as observed with the other runners. The pace versus grade plot shows the same structural lines. A clear difference is seen with regards to the intensity of cadence: it remains darker for the vast majority of the data (between a cadence of 90 to 95). Some lighter shades are seen at the slower paces, roughly from 6.5 min/km. The data also tend towards the 0% line for faster paces from both positive and negative grades.

Figure 5.11 shows the boxplots for the spread of grade across the cadence levels before and after the data clean-up. Almost all of the blue coloured outliers are found in the cadence range between 84 and 94, with some scattered outliers for the other cadence levels. The IQRs' behaviour presents with a curvi-linear trend around the 0% line. The IQRs for the lower and higher cadences (below 80 and above 96) are skew and almost all the data for these levels are contained within the whiskers of their boxplots. Subsequently, these boxplots remained almost the same after the data clean-up procedure. The almost non-existent linearity between grade and cadence may be explained in the curvi-linear shape of the IQRs.

After the data clean-up the scatter plot for grade versus cadence now presents with a cyclic pattern of inward and outward tapering. The asymmetry of the colour intensity of pace around the 0% line is still present. The data from a cadence of 97 and higher remain almost as-is after the clean-up, as these data points were all contained in the IQR or whiskers of the boxplots. The main difference seen in the pace versus grade plot is the contraction of the range of grades. Almost all grades below -20% were removed with a few points still scattered above 20%.

From Table 5.17 the runner achieves his fastest mean pace during DR, followed by LR and the slowest mean pace is during UR. The colour intensity of pace over the grade versus cadence from the scatter plots from Figure 5.10 supports these values. The athlete spends most of his time doing UR and the least of his time at LR (only 12.1%). The low percentage of LR may be due to the absence of track training as seen in the other runners, which are done predominantly at level ground.

Table 5.17: Descriptive statistics for the effect of gradient on pace for athlete 5

Variable	Value
Mean pace (DR)	4.908
St.dev pace (DR)	0.592
Mean pace (UR)	5.336
St.dev pace (UR)	0.618
Mean pace (LR)	5.117
St.dev pace (LR)	0.624
Data points % (DR)	41.003
Data points % (UR)	46.911
Data points % (LR)	12.086

The results from the log-transformed multiple linear regression ICM ($R_a^2 = 0.228$) to measure the re-enforcement and interfering effects of slopes on cadence and pace are presented in Table 5.18. For this athlete the re-enforcement effect of UR is insignificant ($p - value > 0.05$), while the DR has an interfering effect on cadence.

Table 5.18: Estimated parameters for the log-transformed interaction model from cadence and slopes on pace for athlete 5 (cleaned data)

	Estimated parameter	p-value
Intercept (β_0)	3.42207	0.00000
Cadence (β_1)	-0.02017	0.00000
Downhill running	-0.31017	0.00000
Uphill running	0.12761	0.05928
Cadence: downhill running (β_{12})	0.00308	0.00003
Cadence: uphill running (β_{13})	-0.00097	0.20257

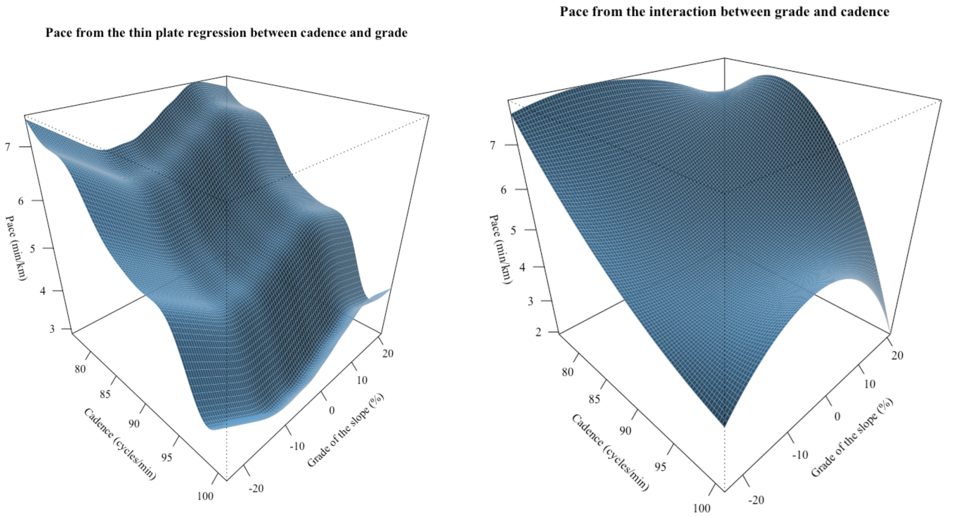
A parametric second degree polynomial function and a non-parametric TPS model were used to model the visualised output from the interaction between grade and cadence on pace using the cleaned data set. The 3D output from the models is shown in Figure 5.12. There are distinct differences between the two models. Where the ICM bulges upward around the 0% grade line, the TPS curve continues downward in a wave-like fashion. The asymmetry around the 0% line is clearer in the TPS model, especially at the higher cadences.

The summary of the models in terms of their adequacy is shown in Table 5.19. The AIC score for the TPS model is lower for both the data sets (before and after clean-up) than the AIC for the ICM. Interesting to note are the negative changes seen for both models' R^2-values. Where the AIC improved, their coefficient of determination actually decreased. The cadence versus grade plot from Figure 5.10 b) is the top view of the 3D models. The TPS model is a better representation of the flat 2D view than the ICM, as the colour intensity from the 2D image corresponds with the downward curvi-linear pattern observed for the 3D TPS model. The pace versus grade plot from Figure 5.10 b) is the side view for the 3D models. Again, the TPS model is a better visual presentation of the 2D version than the 3D curve from the ICM. The faster paces that converge to the 0% line agree better with the downward trend of the bent plate than the bulging trend across the grades observed in the ICM.

Table 5.19: Summary from the thin plate and the polynomial interaction models for cadence, grade and pace for athlete 5

Variable	Value
AIC (TPS)	69215.390
R^2 (TPS)	0.287
AIC (TPS) after clean-up	65875.896
R^2 (TPS) after clean-up	0.280
Change in AIC for the TPS	3339.495
AIC (ICM)	70129.094
AIC (ICM) after clean-up	66534.556
Change in AIC for the ICM	3594.538
R^2_a (ICM)	0.271
R^2_a (ICM) after clean-up	0.268
Change in R^2_a for the ICM (%)	-1.275

Table 5.20 directly compares the fitted values for the TPS model and the actual values from the clean data. Pace is given in min/km. The difference between the fastest and actual and fitted paces presents 22.6 meters over a 1-minute interval. The only values that closely agree are the actual and fitted pace at the maximum grade. The fitted and actual cadence at the fastest pace differ by three cycles, which is substantial. Cadences above 96 represents 0.154% of the total data set, which may make it difficult for the model to accurately fit these high cadence values for the faster paces. The actual and fitted grade at the fastest pace are on opposite sides of the 0% line which is a contradictory outcome to present to an athlete.



(a) The thin plate spline model

(b) The polynomial interaction model

Figure 5.12: The 3D curves as a result of the interaction between grade and cadence for athlete 5

5.5 Conclusion

The analysis of overground running with the focus on slopes proved to be more challenging than initially expected. The symmetry of the colour intensity for pace surrounding the 0% grade line in the grade versus cadence scatter plots are cause for some concern. The accuracy of the altitude readings are questionable, or perhaps even the generated pace at the instances of UR and DR. The straight lines in the pace versus grade plots also gave rise to some doubt regarding the structure of the underlying altitude data. This structural representation may be the result of how the device actually records the runner's altitude. It may be that in most cases altitude is recorded in set interval sizes for the given geographical region which gives rise to a constant change in elevation for each data point in that region, and is therefore not presenting the actual instantaneous change in gradient of the slope. The tendency of the straight lines from the pace versus grade plots may explain the non-existent to weak associations between grade and pace. The absolute value of the grade, i.e. the scalar distance from the grade to 0, may be the driver towards faster paces as the grades near 0%. During UR a runner must use more energy to propel his body against gravity, however, during DR a runner must facilitate braking that slows the pace. This might explain the tendency of the straight lines towards the 0% grade line from both positive and negative grades, even though the mean pace for three athletes were the fastest when going downhill.

As the case with the models for cadence and running activity from Chapter 4, the spread of gradients per cadence level is extensive and skew for all the athletes. The same

Table 5.20: Actual and fitted values from the thin plate model for athlete 5

Variable	Value
Fastest actual pace	2.73
Fastest fitted pace	2.91
Slowest actual pace	8.31
Slowest fitted pace	7.47
Actual pace at min grade	7.86
Fitted pace at min grade	7.47
Actual pace at max grade	6.63
Fitted pace at max grade	6.67
Actual cadence at fastest pace	96
Fitted cadence at fastest pace	99
Actual grade at fastest pace	3.27
Fitted grade at fastest pace	-4.76

explanation behind the skewness of the pace around each cadence level is presented for this reasoning. The athlete may have already changed their cadence when running at the instantaneous grade, but the algorithm to change the cadence reading is yet to adapt to the true level. Higher or lower grades (and paces) are therefore associated with the wrong cadence levels that were achieved in reality. The IQRs in the boxplots for the grades per cadence level gave away the pattern of the data. Athlete number 4 was the only athlete whose general pattern of the IQRs concurred with the research findings of Vernillo et al. (2016): he uses higher cadences for increases in slope gradients. The IQRs from athlete 3 reflected the symmetry found in the 2D visualisation. The IQR behaviour of athlete 2 is the opposite to athlete 4 – it seems that the negative grades are run at the higher cadences. The IQRs from athlete 5 showed a curvi-linear trend and might be a reflection of his preferred self-monitoring technique during running (to use heart rate as a measure to gauge pace, with possible variation in step length and not cadence for increasing grades). The 3D TPS curves were able to correct some of the observed visual symmetry around the 0% grade line that was apparent in the 2D images, especially at the upper range of the cadences.

Runners can use the 3D image or the results from the TPS model to make decisions regarding the management of pace. Together with a race plan that stipulates the pace they ought to manage in order to finish the race within the goal time, they can use the 3D model to match their capabilities regarding the management of pace on the slopes. For instance if it is required that a pace of 5.5 min/km be maintained up a slope of 5%, the runner can use the 3D image and first evaluate whether they are reasonably capable of achieving the required pace. If they are, they can find the corresponding required cadence. Otherwise they must adjust the race plan or their goal.

A draw-back from the 3D image is that the figure requires intuitive rotation to find the best vantage point for meaningful understanding and a user with little or no aptitude for 3D image processing will struggle to interpret the figure. A single vantage point does not enable the viewer to see all the data and some points are hidden behind the curves of the plate. The 3D image also does not capture the variability in pace for each combination of gradient and cadence but only provides the expected value for each combination. In order to understand their pace variation, a runner will have to be familiar with some statistical principles such as a curve's goodness of fit. The flat 2D image requires no rotation and the colour scales immediately apprehends patterns in the data. It allows the viewer to self-generate a mental image of their running interaction with the environment but requires some interpretation of the colour scales and axes. The runner can also immediately appreciate the variability of their pace for each cadence level and on slopes. These visualisation models may advance a runner's understanding of their running form on slopes, but the underlying data still requires verification and validation, especially the altitude data used to calculate the instantaneous grade of the slope. The recreational or lifestyle athlete set to benefit probably the most from the visual analysis proposed here, as it does not require in-depth understanding of all the mechanics involved but shows potential to communicate the runner's running form during graded running. The spatio-temporal data from the running watch may not always be reliable and subsequently the analyses thereof must be approached with some caution.

A suggestion on future work follows from the categorisation of contniuous clinical data as done by Barrio et al. (2013) using a GAM. The categorisation of continuous grade on a ordinal scale ranging from "very steep uphill" through to "level" and on to "very steep downhill" may reduce the load on a model to estimate running form on slopes by replacing the highly variable continuous grade with a limited categorical variable. It might also be considered to isolate areas with known and validated altitude data and substitute those for the altitude data from the GPS container file. The analysis will then only be valid within the isolated area, but results (running form on slopes) would be more reliable and representative of the truth.

Chapter 6

The distribution curves to monitor running form

The analysis of the distribution curves continues on the second cleaned data from Chapter 5, i.e. it is built on the data that has also been cleaned from the outliers of grade per cadence level. The distribution curve is the integral of the density curve (or the histogram in the case of cadence) and presents the proportion of time the athletes spend at a certain cadence or lower and at a certain pace or faster. The direction in which the distribution curves must be read is important. The cadence curve must be read from left to right (going from low cadences to higher cadences) and the pace curves must be read from right to left (going from slower paces to faster paces). The critical regions in the distribution curves are the changes in the slope of the curve, as these areas imply changes in the direction of the density curve (where it starts to move towards a secondary peak or present with a tail). A decrease in the slope (i.e. becoming more flat) indicates an improvement and a tendency towards higher cadence or faster pace. The changes in the slope of the distribution curves may indicate the following:

1. Bi-modality in the original data (pace or cadence), which shows that the athlete is capable to run at higher cadences or faster paces other than their mean for a considerable portion of the time. The bi-modality is sometimes visible as a clear "bump" in the distribution curve. This change in the slope is an indication of a step-wise improvement.

2. An overall shift of the distribution curve towards higher cadences or faster paces points to holistic improvement in performance.

3. Sudden increases in the slope (when it becomes steeper) may reveal a regression towards slower paces or lower cadences, which are unwanted changes. The athlete's

performance may be decreasing and if monitored over time must be investigated to find the cause.

Four non-parametric models for cadence and pace were constructed per athlete (two models per variable). Two GAMs – a Gaussian distribution and a Gauss-log transform for cadence and pace, as well as two glsscms, for a Gaussian distribution and a Gauss-log transform of the calculated percentages. Four measures were used for model adequacy:

- The AIC score;
- the number of negative fitted values;
- the fitted curve's monotonicity;
- the fitted curve's behaviour in the critical region.

The fitted curve's behaviour and sensitivity to change in the critical regions proved to be the most important criterion and is considered a superior quality to pure monotonicity and the curve's ability to not have any negative values fitted.

6.1 Case study A: the semi-professional all-rounder

Table 6.1 contains the summary of the model adequacy measures extracted from the GAMs and SCMs in *R*. The non-parametric models are compared using the AIC values.

Table 6.1: Model adequacy measures for the distribution regressions for athlete 3

Variable	AIC score	R^2	Negative values	Monotonicity score	Technique	Distribution
Cadence	114.80	0.9994	3	8	gam	gaussian
Cadence	126.47	0.9991	0	5	gam	gauslog
Pace	2.87	1.0000	0	0	gam	gaussian
Pace	-140.12	1.0000	0	0	gam	gauslog
Cadence	204.09	0.9900	4	0	scm	gaussian
Cadence	164.86	0.9969	0	0	scm	gauslog
Pace	268.26	0.9997	5	0	scm	gaussian
Pace	-91.57	1.0000	0	0	scm	gauslog

6.1.1 Analyses of the GAM and SCM on cadence

Figure 6.1 shows the histogram, the fitted distribution curve and the diagnostic plots for both the Gaussian model family and the Gauss-log transformed model on the distribution for cadence from the GAM. The fitted line is overlaid with the original calculated percentages. Clearly the distribution of cadence in the histogram is bi-modal with a small secondary peak at a cadence of 97. For the Gaussian model, the fitted curve (top right graph in Figure 6.1a) show some instability and wiggliness at the start up until a cadence of 86, which after it stabilises and seem to follow the calculated distribution fairly well. This secondary peak is reflected in the distribution curve, where the slope of the line decreases and a small bulge forms starting at a cadence of 96. The residual plot reflects the curviness of the fitted line, especially in die middle region. The errors are more closely scattered around the endpoints.

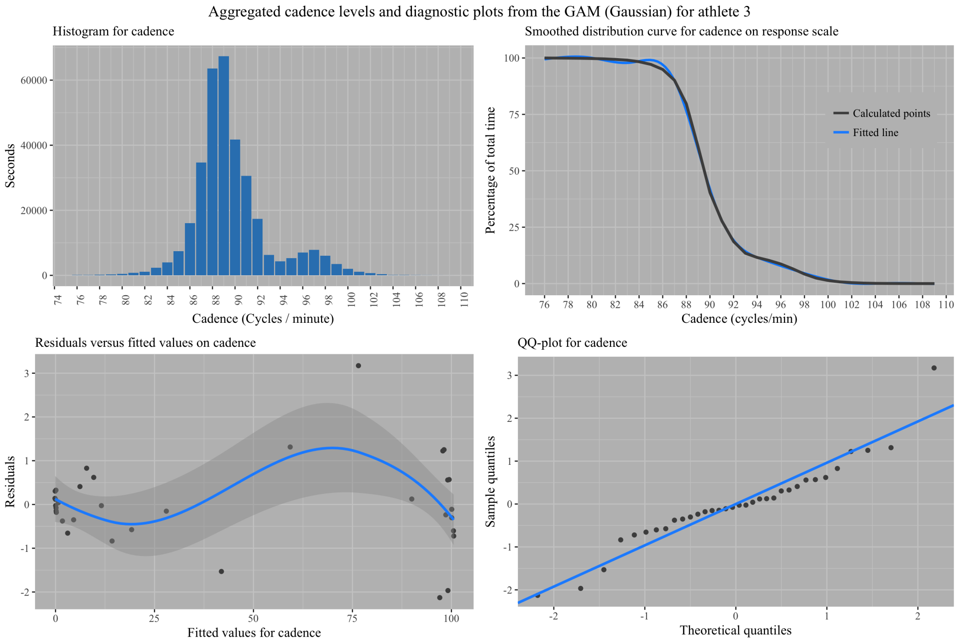
The fitted line for the Gauss-log model (top right graph in Figure 6.1b) also displays wiggliness at the start, with a slight departure from the original data throughout the middle region. The main difference from the Gaussian model is seen at the end spectrum of cadence, where the fitted line becomes wiggly again and dips below the original data at a cadence of 94. This wiggly behaviour continues but stabilises at a cadence of 101. The fitted line completely fails to pick up the behaviour in the critical region. The amplitude of the errors in residuals plot is lower than the Gaussian model, but this might be due to the transformation of the data that reduces the extent of the error terms.

Figure 6.2 contains the relevant plots for the SCM on cadence. The fitted line for the Gaussian model (top right in Figure 6.2a) shows departure from the original data almost throughout the range of cadence. The bump resulting from the secondary small peak in the histogram is somewhat picked up in the fitted distribution line. The residual plot for the Gaussian model shows amplified curvature when compared to the same model from the GAM.

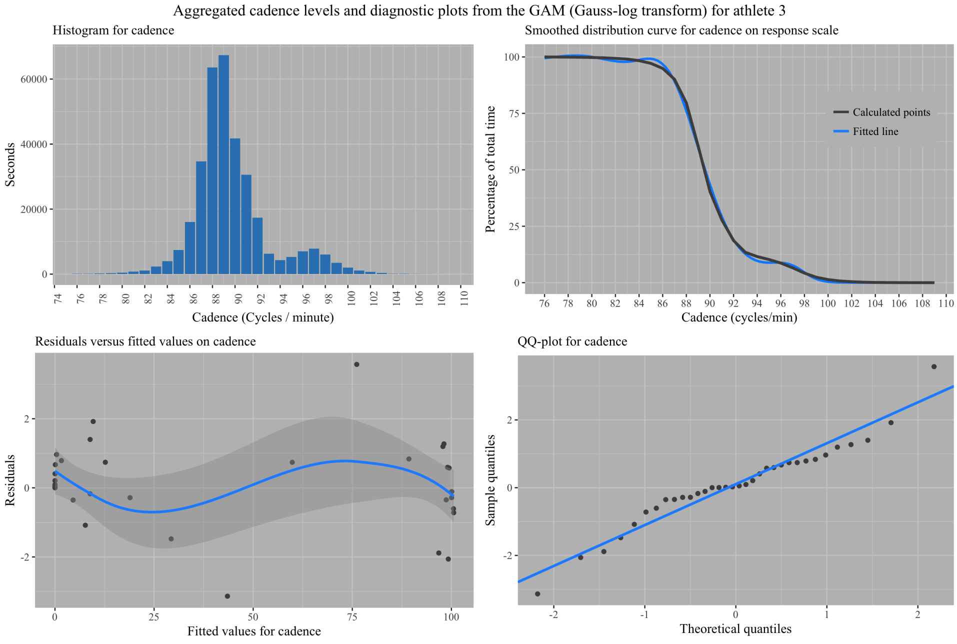
The Gauss-log model (top right in Figure 6.2b) from the SCM on the other hand agrees better with the original calculated distribution percentages than the Gaussian model. The waves in the residual plot have a lower amplitude from the 0 line. Although perhaps a better visual fit to the original calculated percentages of the distribution, this model too fails to pick up the “bump” between cadences 94 and 96.

Neither the SCMs were successful in fitting the slope in the critical region and is therefore excluded as candidates for selection. The lowest AIC value for the cadence models is 114.8 from the GAM based on the Gaussian family distribution. The next closest AIC score is the GAM with the Gauss-log transform. However, this model shows instability in the critical region and failed to follow the change in the slope satisfactory. Despite the Gaussian model’s three negative values and that the curve is not monotonic, meaning that the slope changes sign during the course of the fitted line, it is the superior model. It was the only model that picked up the decrease in the slope in the critical region, albeit underestimating the the change in the original data.

Figure 6.3 is a continuation of the selected GAM with the Gaussian model, where the data have now been subset into the different running activities. The sub-setting of the

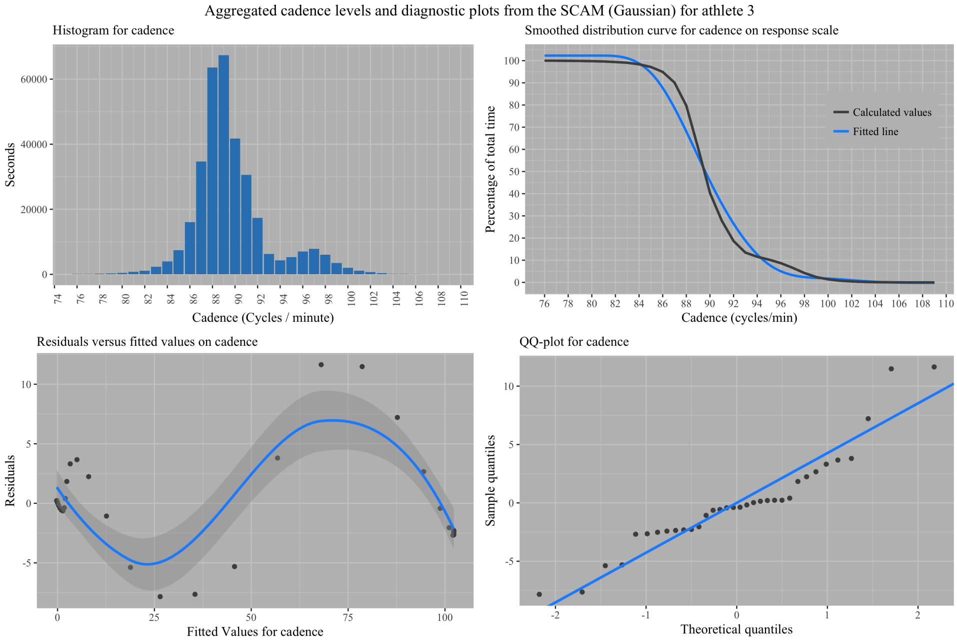


(a) Gaussian model

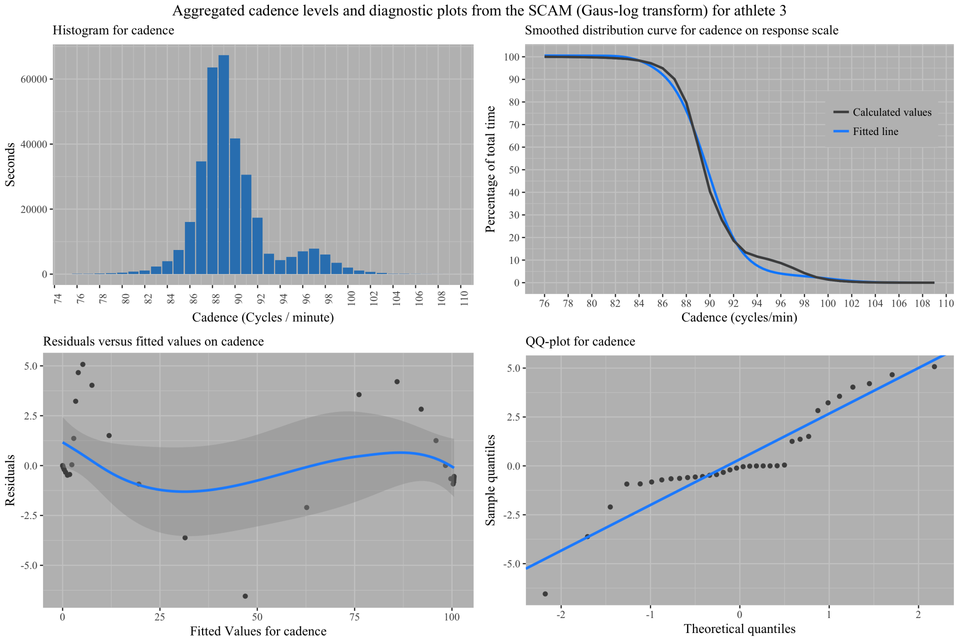


(b) Gauss-log transformed model

Figure 6.1: The distribution curves for cadence from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 3.



(a) Gaussian model



(b) Gauss-log transformed model

Figure 6.2: The distribution curves for cadence from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 3.

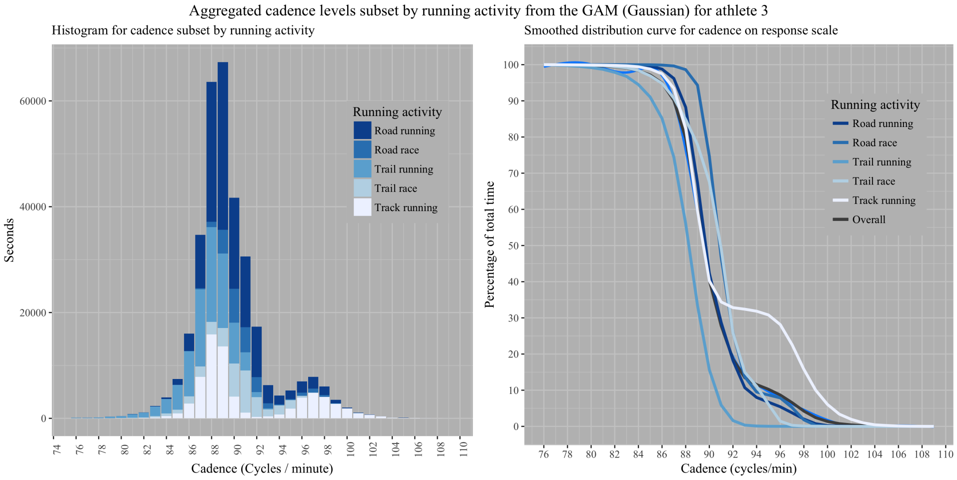


Figure 6.3: The distribution of cadence subset by running activity for athlete 3

original data into the respective running activities as done in Chapter 4 clearly shows how the time spent at the various cadence levels differ. It is now apparent what caused the secondary peak in the histogram. The distribution for track training is bi-modal, with a small bi-modality picked up for road running. As expected, the curve for track training bulges to the right, as the athlete is running faster during track training and is using a higher cadence for the pace generated. The curve for road running has a small bulge to the right. The bi-modalities in the track- and road running distributions may be explained when exploring the different subsets of of track training and road running regimes. Track training consists of shorter distances covered in minimal time, ranging from 300m to 800m. The longer distances in track training involves 1000m and 2000m, also covered at the fastest pace possible. The athlete will probably run the longer distances at a somewhat slower pace and a lower cadence and cover the shorter distances with a faster pace and a higher cadence. Road running consists of longer runs and shorter tempo runs. During the tempo runs the athlete runs sections of the route at a pace slightly above his normal pace and the rest of the sections at normal pace or rest by walking. The same concept applies – the athlete will probably use a higher cadence to generate the faster pace during those intervals and a lower cadence during his normal running pace.

6.1.2 Analyses of the GAM and SCM for the distribution on pace

Figure 6.4 shows the density- and distribution curves, residual plots and the qq-plots for pace from the GAMs for both the Gaussian family and the Gauss-log transform. Starting with the Gaussian family from the GAM, the fitted line almost perfectly follows the original data points (top right in Figure 6.4a). The right-skewness of the pace is also evident in the pace density curve (top left in Figure 6.4a), with the mean being right of the curve's peak. There is an indication of bi-modality in the density curve with a small peak at at around 3.5 min/km. This bump is reflected in the distribution curve where the rate of the descent of the line has decreased from around the 4.25 min/km mark. The residual plot shows clustering of the error terms into three groups in the mid-range between 25% and 90% on the fitted values. The errors in upper cluster has a u-shaped pattern.

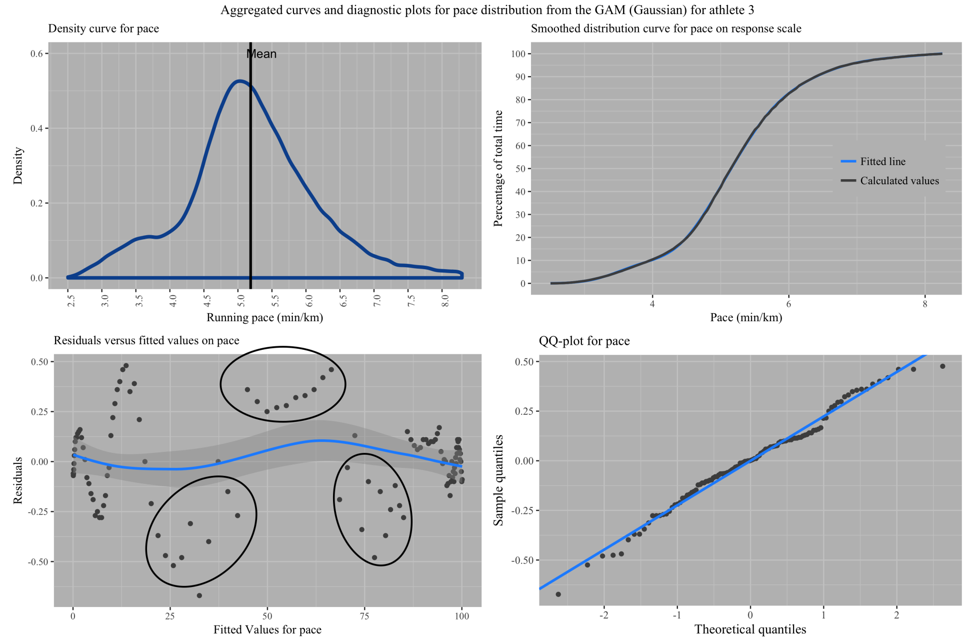
The fitted line from the model for the Gauss-log transform from the GAM seems to match the calculated data points even more closely. The bump in the distribution curve is also reflected by the decrease in the rate of change of the distribution plot at roughly 4.25 min/km. The same type of wave-like pattern is shown in the residual plot, albeit now the error terms follow a continuous rise and fall cyclic pattern.

Figure 6.5 shows the density-, distribution curves and the diagnostic plots for the distribution on pace from the SCMs. The fitted line from the Gaussian model (top right in Figure 6.5a) swirls around the true calculated data points in small amplitudes and moves below the 0% line, which is highly undesirable for the prediction of running form and simply cannot be true. The fitted line reacts poorly to the bump in the density curve, first dipping below the actual data points and then above the points.

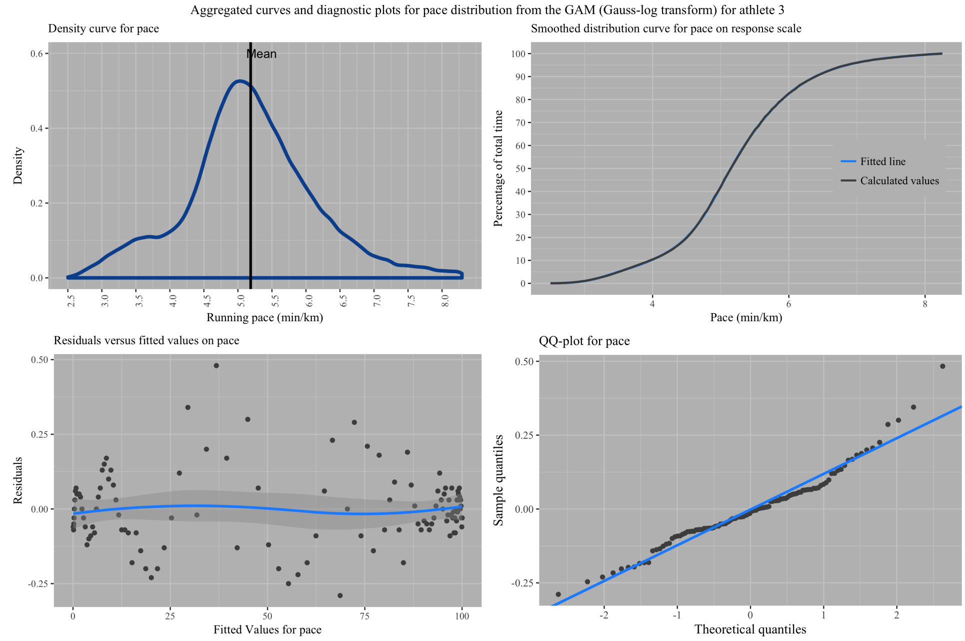
The Gauss-log model (top right in Figure 6.5b) did not fit any negative values. It did not pick up the change in the slope of the true distribution points at the bump in the density plot, but completely underestimates the rate of change in the original curve.

Table 6.1 contains the model adequacy measures for the pace distributions. The Gaussian SCM has fitted negative values (which was also reflected in the visualisation of the curve in Figure 6.4) and is therefore discarded from further consideration. All the other models are perfectly monotonic without any negative fitted values. The AIC score for the Gauss-log transformed GAM is the lowest of the three remaining candidates followed by the same model version of the SCM. This model also reflected the bump in the density curve. The Gauss-log version of the GAM is therefore the preferred model from all the candidates. An added benefit to the Gauss-log transform is that it will never fit negative values. This benefit is however inherit to the mathematical nature of the natural logarithm, in that it tends to infinity for decreasing values of y but never crosses the $x = 0$ line.

Figure 6.6 shows how the selected model is broken down into the running activities. The smaller peak on the left side of the density curve for pace is caused by the first peak of the density curve for track training. The pace for track training is bi-modal. The right sided skewness for the trail racing is apparent. This bi-modality seen in track training has the same theory behind the bi-modality seen in the histogram for the cadence during track training. The distinctions between the distances covered in track training

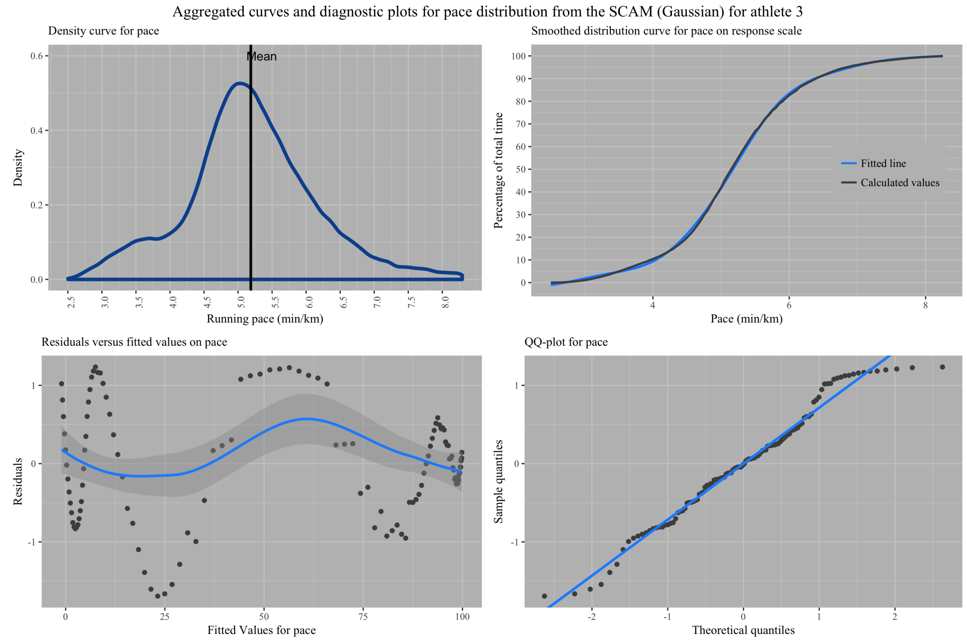


(a) Gaussian model

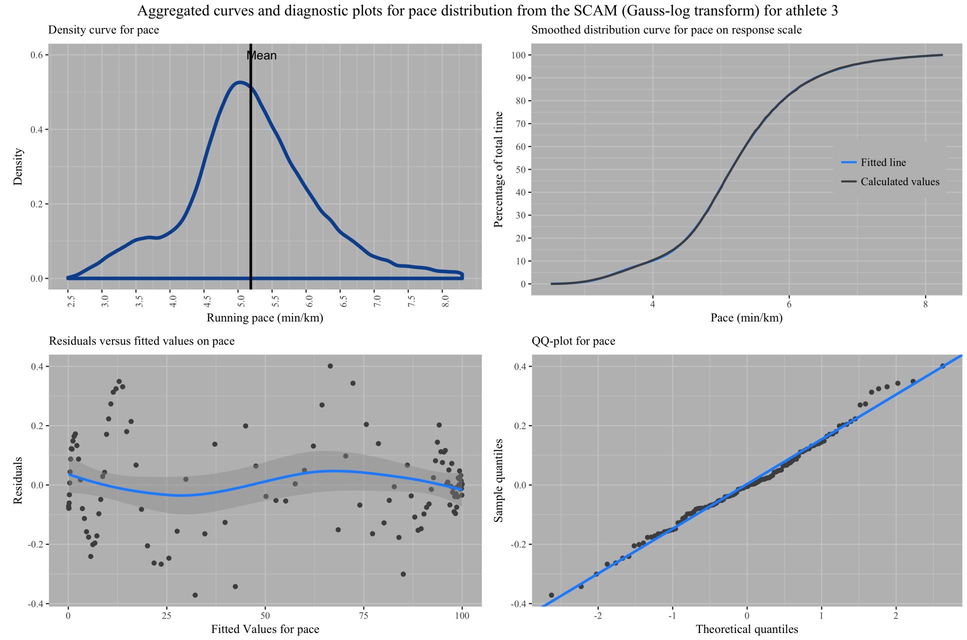


(b) Gauss-log transformed model

Figure 6.4: The distribution curves for pace from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 3.



(a) Gaussian model



(b) Gauss-log transformed model

Figure 6.5: The distribution curves for pace from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 3.

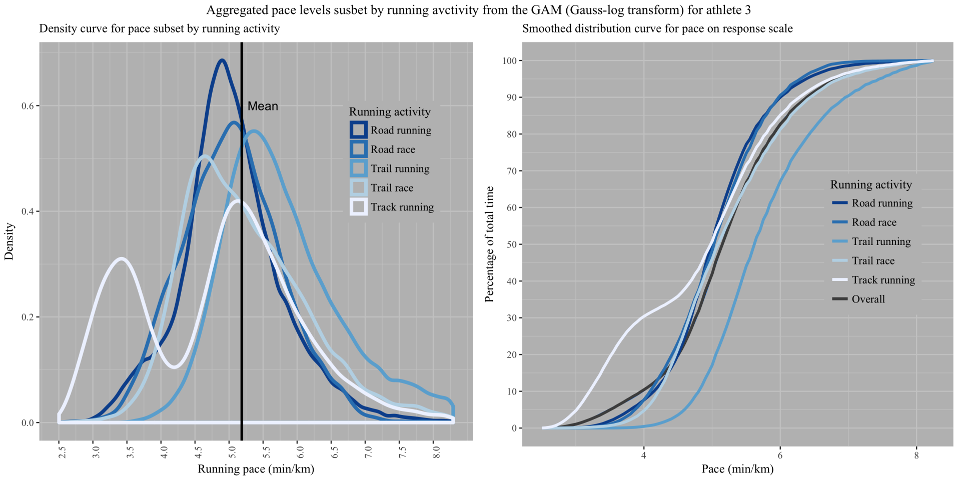


Figure 6.6: The distribution of pace subset by running activity for athlete 3

and the subsequent faster paces associated with the shorter distances are most probably responsible for the two peaks seen in the density curve. Interesting to note is that the average for trail racing is slightly faster than the average for road racing, and the average for racing is slower than the average for road running The bi-modality of track training is reflected in the distribution curves with a shorter peak at faster cadences and the secondary peak somewhat slower than road running. The sharp change in the slope of the distribution curve in Figure 6.6 for track training from roughly the 4.5 min/km mark relates to the shorter peak to the left in the density curve. There is some crossing of the curves for road racing, road running and trail racing at the 4.25 min/km mark. The athlete is capable to spend between 15% and 20% of his time running at this pace or faster for these three running activities. The curve for trail running remains below the rest of the curves.

6.2 Case study B: the trail specialist

Table 6.2 contains the summary of the model adequacy measures for the distribution regression models on cadence and pace.

6.2.1 Analyses of the GAM and SCM on cadence

Figure 6.7 shows the histogram, the fitted distribution curve and the diagnostic plots for both the Gaussian model family and the Gauss-log transformed model on the distribution for cadence from the GAM. The fitted line is overlaid with the original calculated percentages. The histogram shows a bi-modal distribution, although not as distinct as with athlete 3 (case study A). A small secondary peak is observed at a cadence of 93 and 94. The fitted distribution line in the Gaussian model (top right in Figure 6.7a) matches

Table 6.2: Model adequacy measures for the distribution regressions for athlete 4

Variable	AIC score	R^2	Negative values	Monotonicity score	Technique	Distribution
Cadence	3.49	1.0000	1	0	gam	gaussian
Cadence	36.00	0.9999	0	0	gam	gauslog
Pace	12.16	1.0000	0	1	gam	gaussian
Pace	-45.58	1.0000	0	0	gam	gauslog
Cadence	69.63	0.9996	3	0	scm	gaussian
Cadence	55.24	0.9998	0	0	scm	gauslog
Pace	156.62	0.9999	4	0	scm	gaussian
Pace	130.34	0.9999	0	0	scm	gauslog

the original data almost perfectly. The rate of change in the line starts to decreases at a cadence of 90 and becomes even slower at a cadence of 92. This corresponds to the start of the secondary peak seen in the histogram. The residual plot shows the dispersion of the error terms in the middle and closer grouping of error terms at the ends.

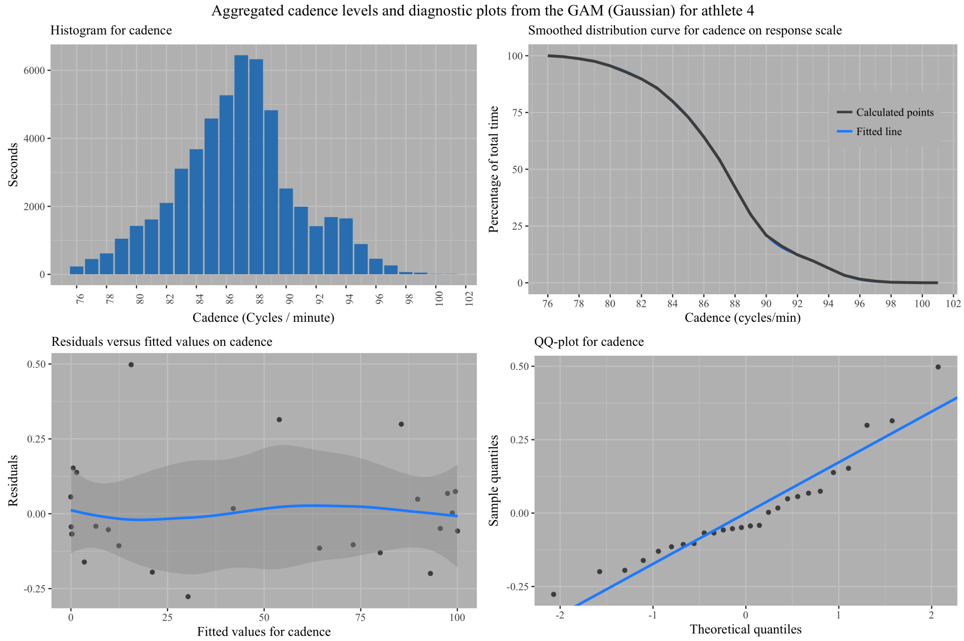
The fitted distribution line from the Gauss-log transformed model (top right in Figure 6.7b) departs slightly away from the original data. It does not pick up the decreased slope at a cadence of 92 as well as the Gaussian model does. The residual plot shows that in addition to the dispersion of the errors, the error terms have been amplified and is now spread further away from the 0.0 line. The qq-plot, however, shows closer adherence to the straight line (intercept of 0.01238 and a slope of 0.3359). Although the intercept is further away from 0 than the Gaussian model, the slope is closer to 1.

Figure 6.8 shows the density, distribution and diagnostic plots for cadence from the SCM for both the Gaussian and the log-transformed models. Neither of the two models' fitted lines pick up the decreased sloped at the cadence of 92. Therefore these models will not be considered any further for selection.

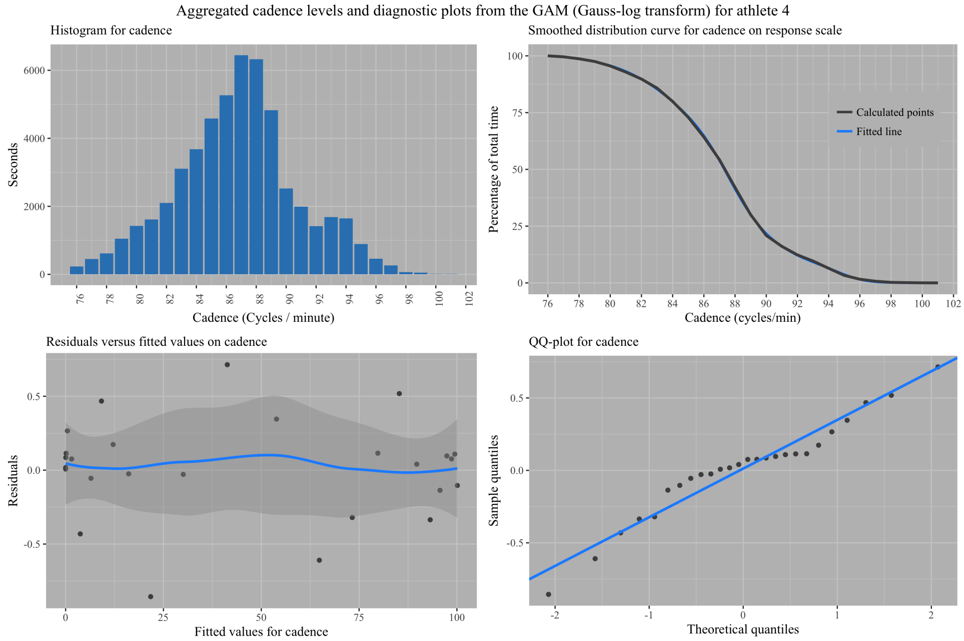
Table 6.2 contains the AIC- and other model adequacy scores for the distribution models. All the fitted lines for cadence is monotonic. Neither of the SCMs picked up the decrease in the slope of the actual distribution curve, which is a crucial quality that the selected model must be able to present to the athlete.

The lowest AIC score obtained is 3.49 for the Gaussian model from the GAM. It also picked up the decreased slope in the actual distribution curve. However, this model fitted a negative value, which is not possible in reality. The Gauss-log transformed GAM did not fit any negative values and picked up the change in the slope of the original data. Irrespective of the negative fitted value, the preferred model is the Gaussian GAM. This decision is based on the difference in capabilities of the model to pick up the change in the slope in the critical region, which is where the Gaussian model performed better. A single negative fit can be adjusted to equate 0 in future work on the model.

The selected Gaussian model for athlete 4 is broken down into the separate running

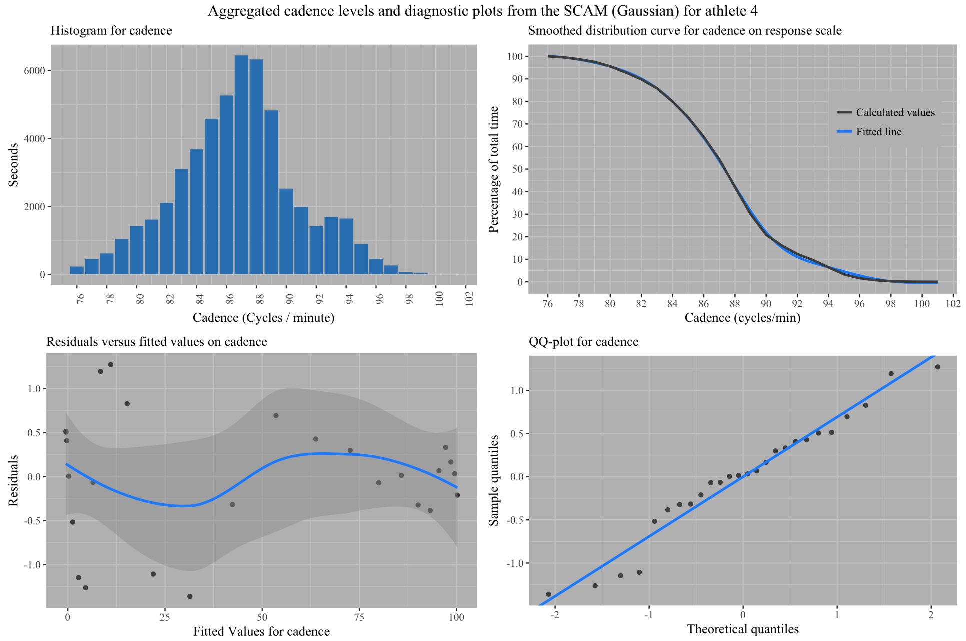


(a) Gaussian model

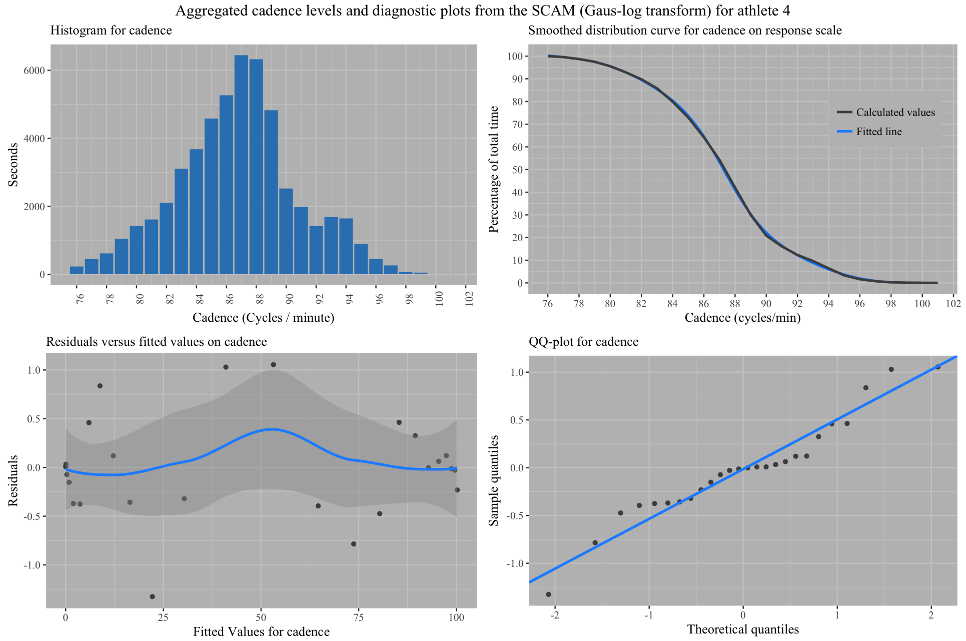


(b) Gauss-log transformed model

Figure 6.7: The distribution curves for cadence from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 4.


Aggregated cadence levels and diagnostic plots from the SCAM (Gaussian) for athlete 4

(a) Gaussian model


Aggregated cadence levels and diagnostic plots from the SCAM (Gaus-log transform) for athlete 4

(b) Gauss-log transformed model

Figure 6.8: The distribution curves for cadence from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 4.

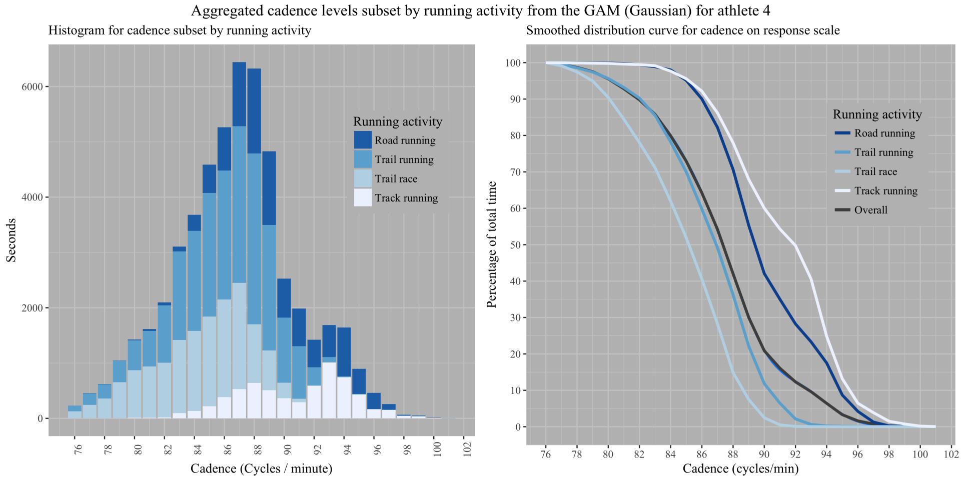


Figure 6.9: The distribution of cadence subset by running activity for athlete 4

activities in Figure 6.9. Track training is largely responsible for the secondary peak seen at a cadence of 92. Track training is also bi-modal in itself with its first peak actually being the smaller peak of the two. This is opposite to what was observed for athlete 3 (case study A). The first peak for track training is also shifted to the right of the main peak for all the data at a cadence of 87. Road running contributes to the bi-modality of the data to a smaller extent. The same explanation as for athlete 3 is applicable for this bi-modality of track and road running. The peaks for trail running and trail racing coincide, but trail racing seems to have a tail to the left. The distribution curve for track training reflects the bi-modality with the sudden rate of change of the slope decreasing at a cadence of 92. Trail racing and trail running falls below the fitted and overall lines which implies that the athlete is generally using the lower cadences during trail running and trail racing. This pattern distinguishes the athlete's capabilities to how he uses cadence during the running activities and underscores the importance of being able to observe running form holistically and not just a single figure of averages. The change in the slope of the distribution curve (right graph in Figure 6.9) implies the runner is improving towards obtaining a higher cadence, whereas a sudden decrease means the athlete might either struggling to move to the higher cadence and prefers staying on the current cadence or that the running environment does not allow the shift to a higher cadence. An example would be a technical ascend or descend during trail running. In order to avoid injury due to missteps, the runner might prefer a slower cadence in order to better control their movement and negotiate the environmental obstacles such as stones, logs, water pools etc.

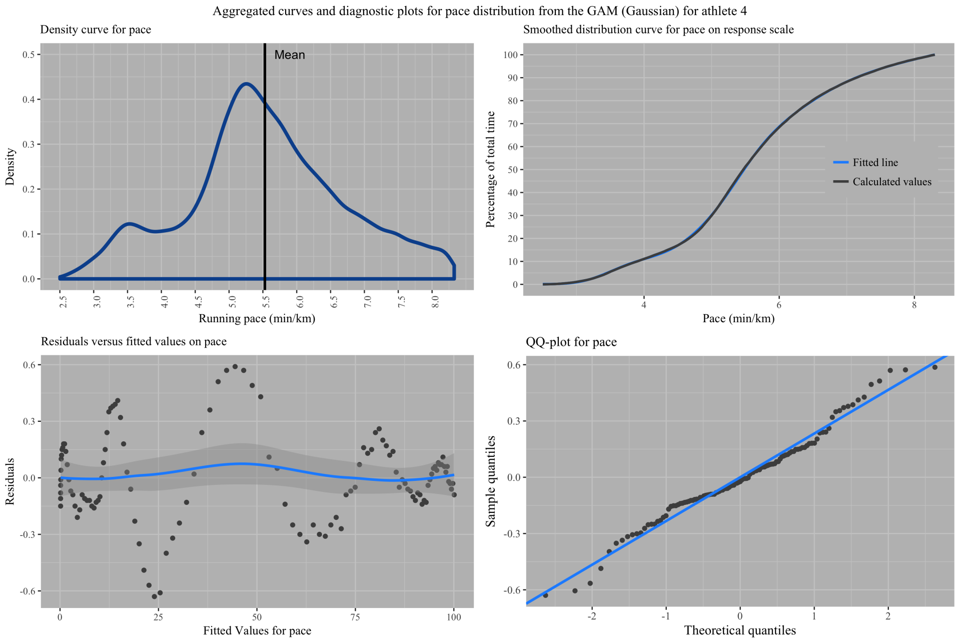
6.2.2 Analyses of the GAM and SCM for the distribution on pace

Figure 6.10 shows the density- and distribution curves, residual plots and the qq-plots for pace from the GAMs for both the Gaussian family and the Gauss-log transform. There is evidence of bi-modality in the pace density plot (top left in Figure 6.10a) with a short peak at a pace of 3.5 min/km and the taller peak at 5.25 min/km. The distribution for pace is still skewed to the right with the mean being to the right of the peak. The fitted distribution curve from the Gaussian model (top right in Figure 6.10a) does well to pick up the bi-modality with a decrease in the slope at 4 min/km. The fitted line fits the original calculated values rather well. The residual plots for both models are curvi-linera and the errors follow a cyclic pattern, with the errors being clustered closer together at the end points and increasing amplitudes toward the middle of the fitted range. The error margins for the Gauss-log model (Figure 6.10b) is smaller than the error margins for the Gaussian model. The qq-plots show adherence to the straight line in the middle, however the dots stay on the same side of the line for consecutive points and then move holistically to the opposite side of the line before departing from it near the ends. This pattern indicates bi-modality in the error terms' distribution with tails on either side for the Gaussian model the Gauss-log model. The Gauss-log model does have a stronger tail to the right. The fitted distribution line for the Gauss-log model (Figure 6.10b) also follows the calculated points very closely and picks up the bump in the density plot at 4 min/km, although it runs slightly above the calculated line.

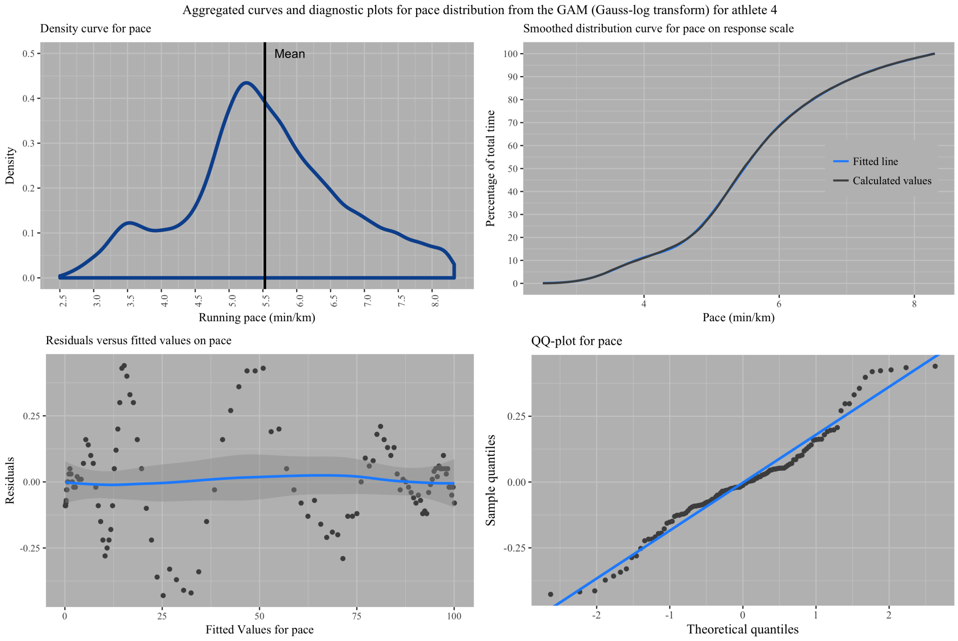
Figure 6.11 shows the density-, distribution curves and the diagnostic plots for the distribution on pace from the SCMs. The fitted Gaussian model (top right in Figure 6.11a) fails to pick up the decreased rate of change in the calculated values at 4 min/km and underfits the data. The line also dips below the 0% line, which is highly undesirable.

The Gauss-log model (top right in Figure 6.11b) does initially pick up the slower rate of change in the originally calculated data, but it first overestimates the line and then underestimates the values in the critical region.

Reviewing the model adequacy table (Table 6.2), the GAM on the Gauss-log transform has the lowest AIC score, followed by the Gaussian GAM. The Gaussian GAM is not entirely monotonic. The Gaussian model for the SCM fitted negative values and has the highest AIC score. This model is eliminated as a candidate. The Gauss-log SCM has a much higher AIC score than those for the GAMs and it slightly overfits the calculated values in the critical region between 4 and 3 min/km. The Gauss-log transform version of the GAM is the prefered model, as it best represents the truth and picked up the bump in the density curve. This model has a mathematical asymptote at the 0% line, which

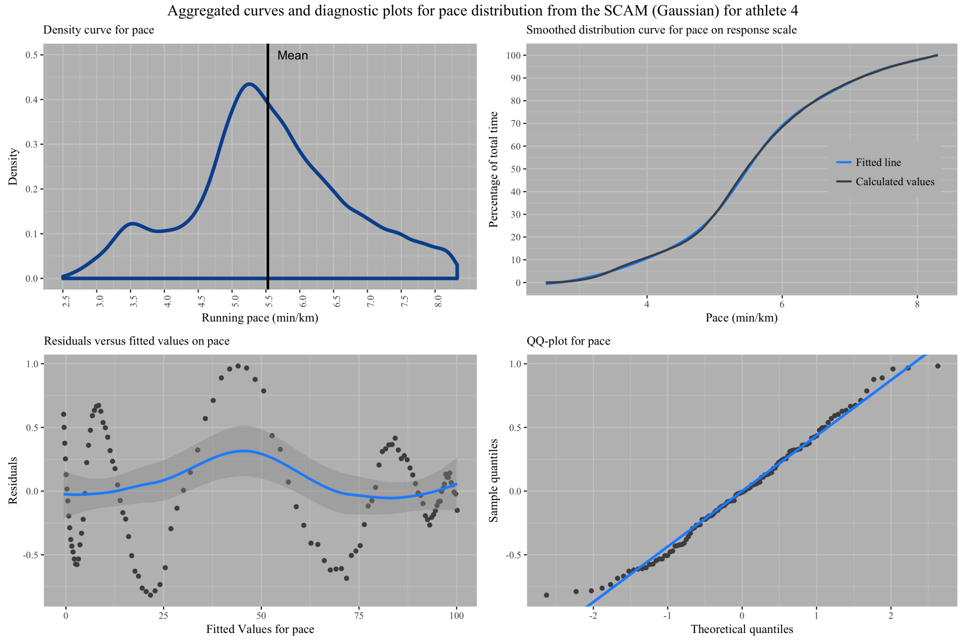


(a) Gaussian model



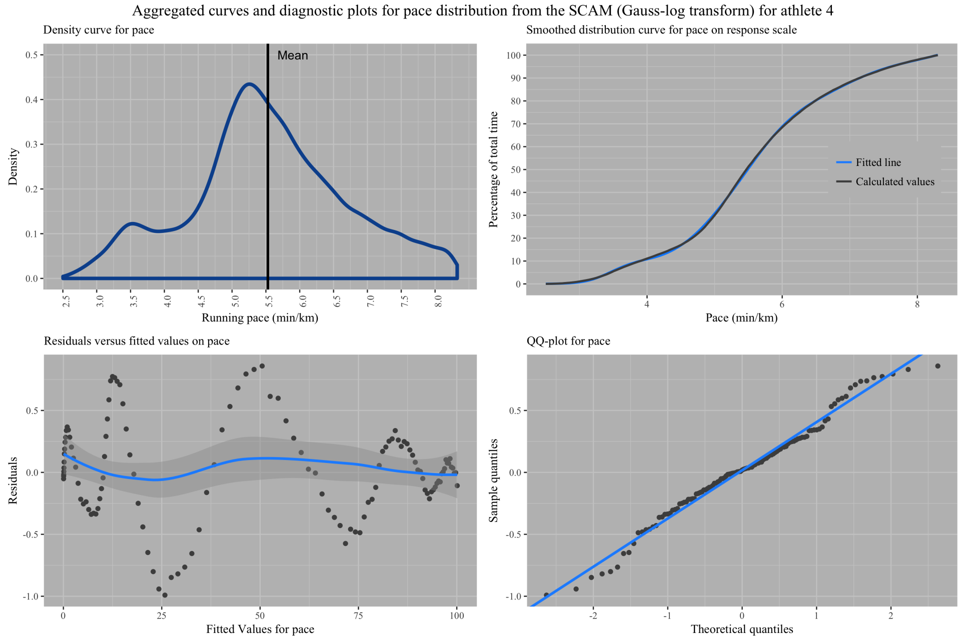
(b) Gauss-log transformed model

Figure 6.10: The distribution curves for pace from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 4.



Aggregated curves and diagnostic plots for pace distribution from the SCAM (Gaussian) for athlete 4

(a) Gaussian model



Aggregated curves and diagnostic plots for pace distribution from the SCAM (Gauss-log transform) for athlete 4

(b) Gauss-log transformed model

Figure 6.11: The distribution curves for pace from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 4.

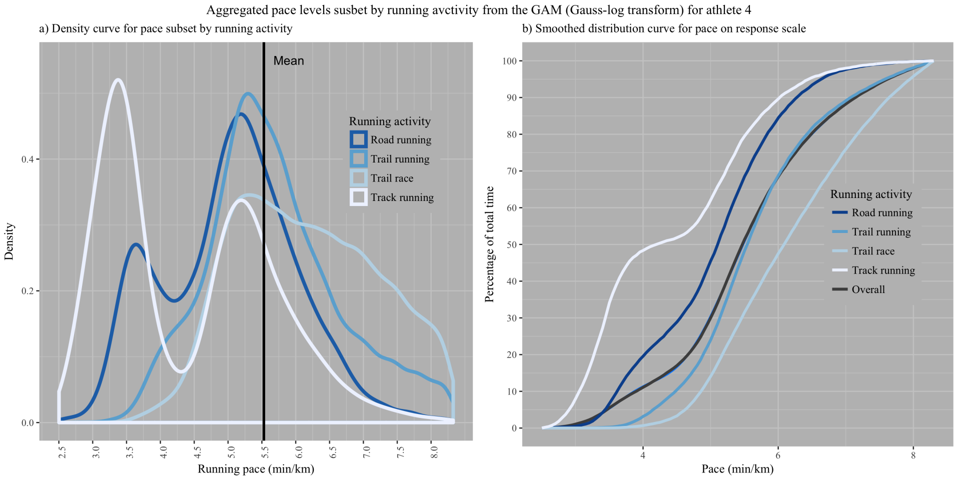


Figure 6.12: The distribution of pace subset by running activity for athlete 4

will prevent negative values.

Figure 6.12 shows how the selected model is broken down into the running activities. The bi-modality of the pace is apparent in the density plot. Both track and road running have two peaks. The positioning of the peaks mirror those of the histogram of cadence for track training: the left peak (at the faster pace) is higher than the secondary peak at slower paces. The histogram from Figure 6.9 showed the initial peak lower than the second. The higher peak for faster paces and the higher peak for higher cadences is an indication that the athlete is using cadence to increase running pace. The mirroring in the order of the peaks shows that the athlete increases his cadence to run faster. The distribution curve picks up the bi-modality of pace, with the change in the slope is the most apparent for track training. Trail racing is skewed to the right. The distribution curves for trail racing is well below the overall and the fitted line and its slope is a gradual descent with almost no change in the slope until the 5 min/km mark. At this point the slope rapidly descends and the rate of change increased. This change in slope matches with the thin tail seen in the density plot towards the faster paces. The athlete may therefore not be able to achieve this faster paces during trail racing. However, keeping in mind the environmental variability of trail running itself, it may not be physiological feasible or economical to run at these fast paces during a trail race. Shifts of the curves in its entirety towards the left will be indicative of general and holistic improvement on pace.

6.3 Case study C: the Comrades marathoner

Table 6.3 contains the summary of the model adequacy measures for the distribution regression models on cadence and pace.

Table 6.3: Model adequacy measures for the distribution regressions for athlete 2

Variable	AIC score	R^2	Negative values	Monotonicity score	Technique	Distribution
Cadence	114.63	0.9984	0	3	gam	gaussian
Cadence	74.28	0.9996	0	0	gam	gauslog
Pace	-14.65	1.0000	2	0	gam	gaussian
Pace	197.85	0.9998	0	10	gam	gauslog
Cadence	144.74	0.9951	1	0	scm	gaussian
Cadence	154.90	0.9931	0	0	scm	gauslog
Pace	476.87	0.9979	2	0	scm	gaussian
Pace	189.84	0.9998	0	0	scm	gauslog

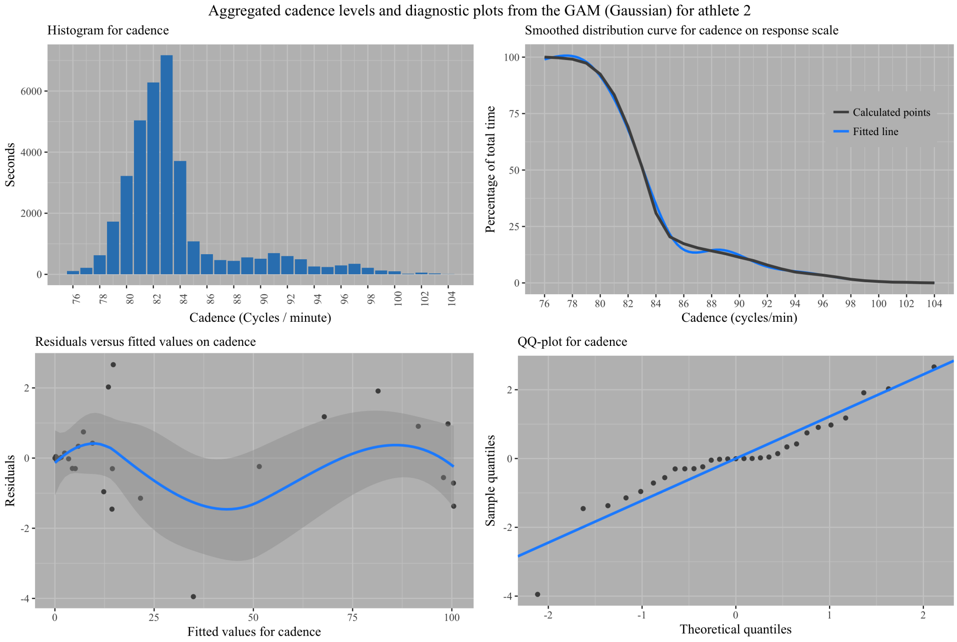
6.3.1 Analyses of the GAM and SCM on cadence

Figure 6.13 shows the histogram, the fitted distribution curve and the diagnostic plots for both the Gaussian model family and the Gauss-log transformed model on the distribution for cadence from the GAM. The histogram (top left in Figures 6.13 a and b) has a long right tail with small secondary peak seen at a cadence of 91. There is clear slow down of the slope of original line in the distribution curve at a cadence of 85 (top right in Figure 6.13a). This corresponds to the start of the tail in the histogram. The fitted line from the Gaussian model (top right in Figure 6.13a) shows some instability at the lower cadences and goes above the 100% mark, which can be regarded with the same mathematical implausibility as a negative fitted percentage. The fitted line fluctuates around the original data points with some increased instability and higher amplitude in the deviation from the original data at the 85 cadence mark. The line does recover and match the original values from a cadence of 94 and higher. The errors are not normally distributed and show a curvi-linear pattern in the residual plots and evidence of bi-modality in the qq-plot.

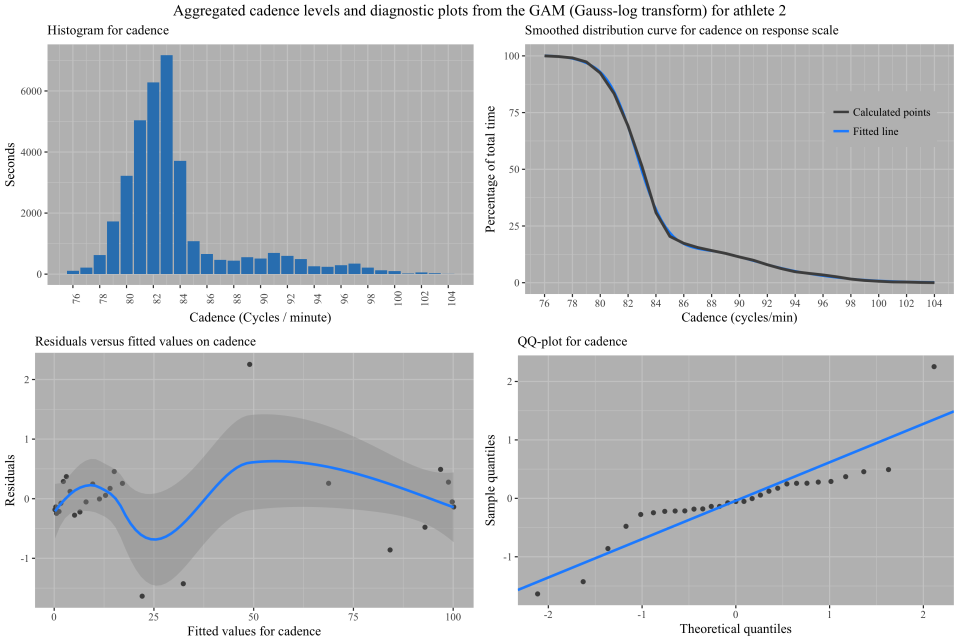
The fitted line from the Gauss-log transformed model (top right in Figure 6.13b) is more stable and does pick up the change in the slope, although a bit delayed at a cadence of 88. It also does not over-fit the data for the lower cadences as the Gaussian model did. The error margin from 0 of the Gausslog-transform is smaller than the Gaussian model as shown in the amplitude if the errors in the residual plot.

Figure 6.14 shows the density, distribution and diagnostic plots for cadence from the SCM for both the Gaussian and the log-transformed models. Both these models over-fitted the original data at the lower cadences to above 100% and failed to pick up the decrease in the slopes in the critical regions. They are therefore not considered any further.

The Gauss-log transform GAM is the superior model with the lowest AIC score and its ability to pick up the change in the slope in the critical region. The selected Gauss-log model for athlete 2 is broken down into the separate running activities in Figure 6.15. It is clear that track running is responsible for the secondary peak at a cadence of 91 and

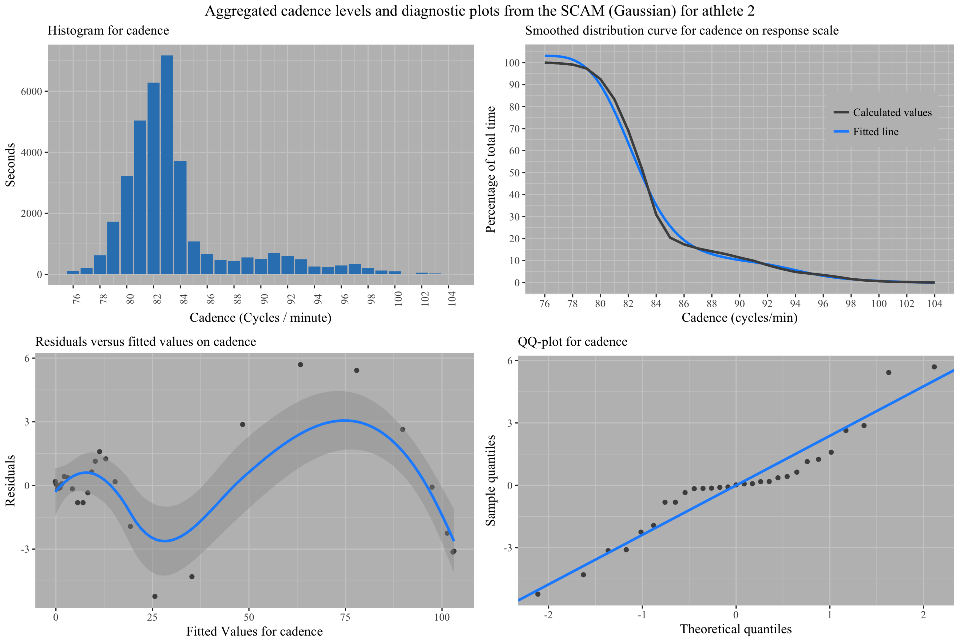


(a) Gaussian model

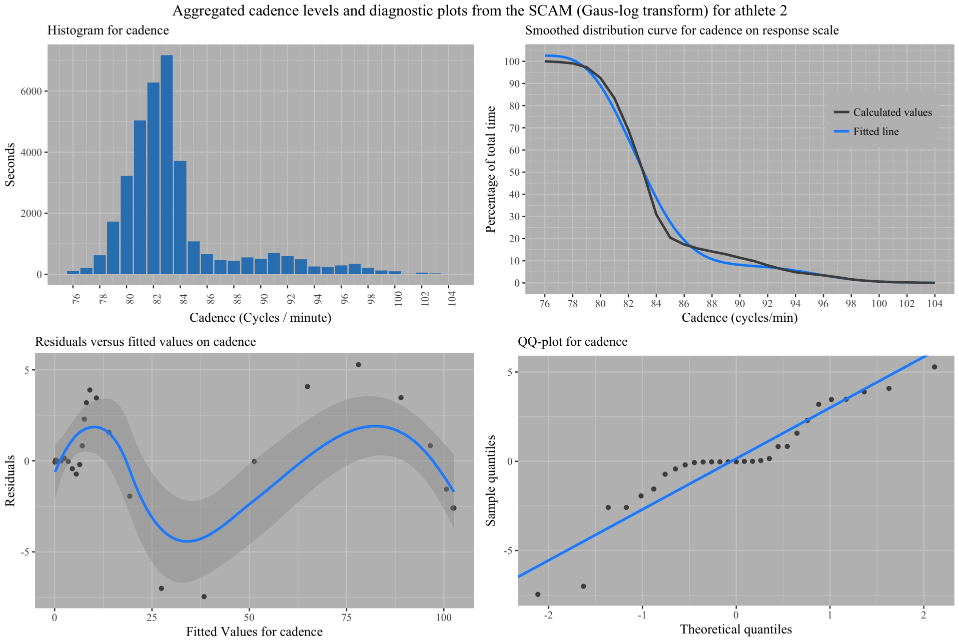


(b) Gauss-log transformed model

Figure 6.13: The distribution curves for cadence from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 2.



(a) Gaussian model



(b) Gauss-log transformed model

Figure 6.14: The distribution curves for cadence from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 2.

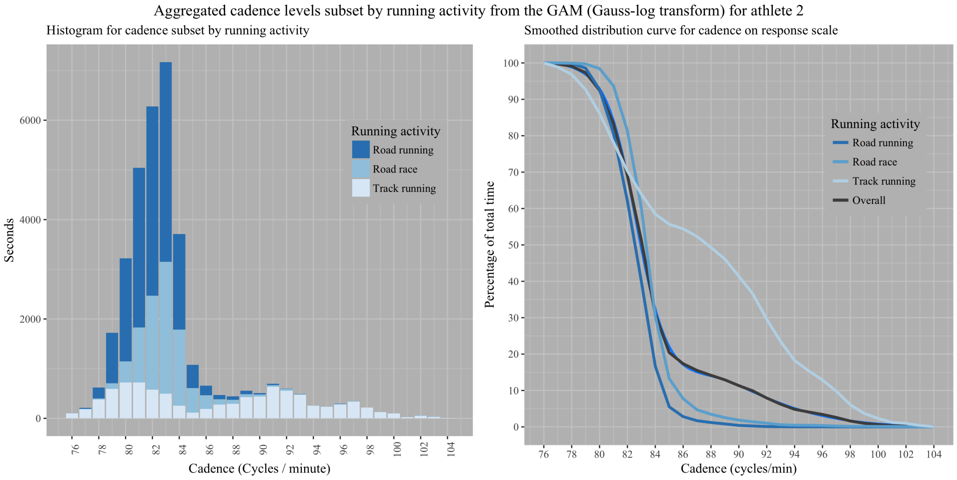


Figure 6.15: The distribution of cadence subset by running activity for athlete 2.

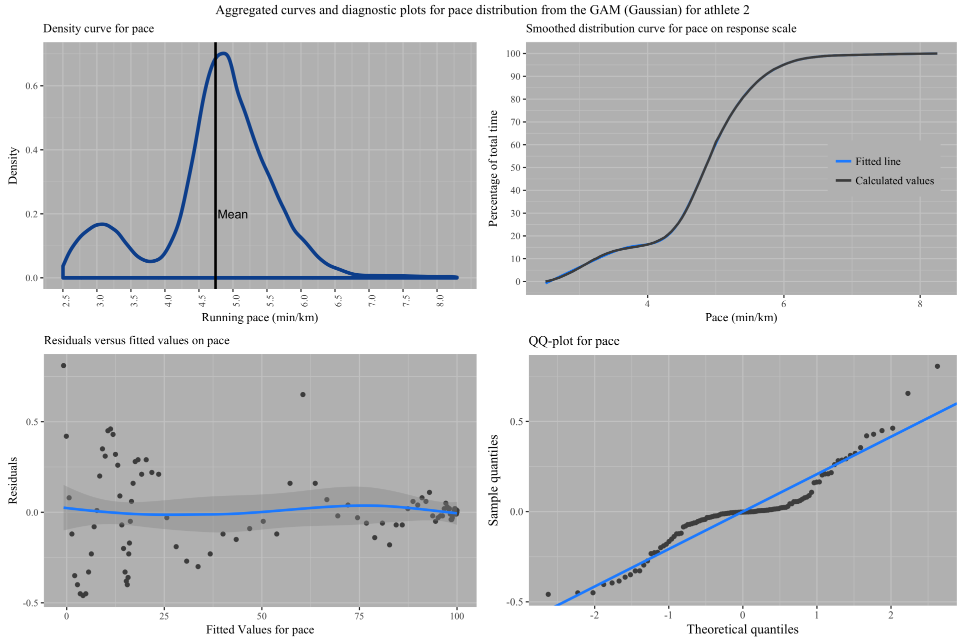
the subsequent changes in the slope. There also seems to be a third smaller peak at a cadence of 97. The first and secondary peaks are 10 cycles apart. These smaller peaks are picked up in the distribution curve where the slope changes (decreases) at the cadences of 84 and 94 on the original data (right graph in Figure 6.15). The fitted curve dips below the first change in the slope but recovers to pick up the second change in slope.

6.3.2 Analyses of the GAM and SCM for the distribution on pace

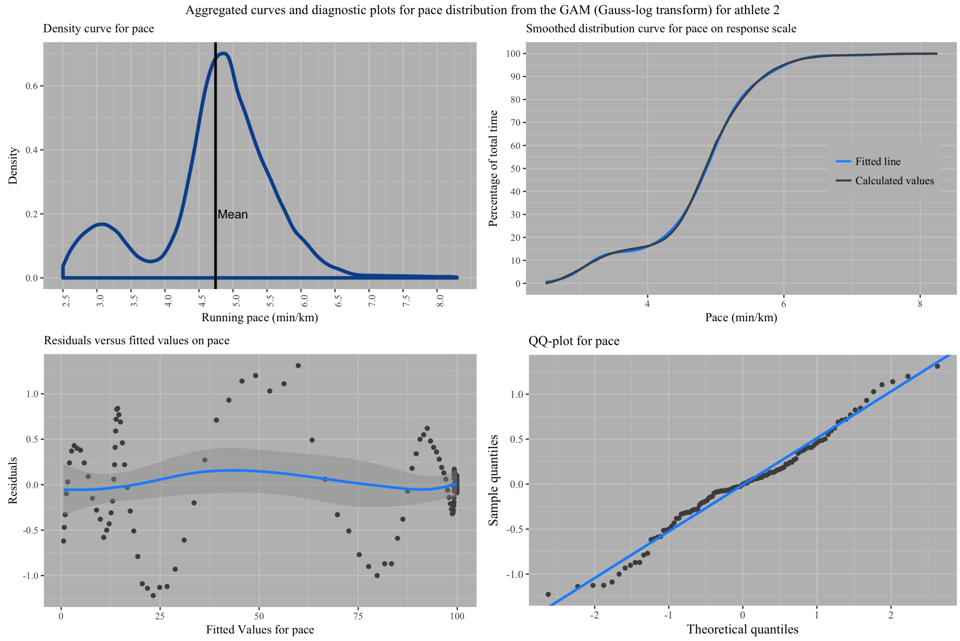
Figure 6.16 shows the density- and distribution curves, residual plots and the qq-plots for pace from the GAMs for both the Gaussian family and the Gauss-log transform. The density plot is bi-modal and skew to the left. The initial peak is just after the 3 min/km mark and the main peak is close to 4.8 min/km. Although the margin is small, the fitted Gaussian line fails to pick up the initial first change in the slope at 3 min/km but performs better at the second change in the slope from where it matches the original data well (top right in Figure 6.16a).

The Gauss-log transform model (top right in Figure 6.16b) fluctuates more around the original data than the Gaussian model. It first over-estimates the bump in the slope from 3 min/km and the under-estimates the change in the slope up to 4 min/km. After the 4 min/km the slope is over-estimated but matches the line well after 4.5 min/km.

Figure 6.17 shows the density-, distribution curves and the diagnostic plots for the distribution on pace from the SCMs. The Gaussian model misses the changes in the slope of the original line and shows great departure from the straight line in the qq-plot. The Gauss-log model also fails to pick up the initial bump in the original data and then over-estimates the line after 4 min/km. The amplitude of the curvature in the residual plot is reduced compared to the Gaussian model with the qq-plot still exhibiting bi-modality of



(a) Gaussian model



(b) Gauss-log transformed model

Figure 6.16: The distribution curves for pace from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 2.

the error terms with a left tail. These models are therefore not considered any further.

From Table 6.3 the Gaussian GAM for pace scores the lowest AIC at -14.65, but has fitted two negative values. The Gauss-log transform from the SCM has the second lowest AIC score with no negative values and is monotonic. The GAM with the Gauss-log transform is not monotonic and has the third lowest AIC score. Despite the negative fitted values, the Gaussian GAM is the selected model to estimate the overall form for running pace for this athlete. This model picked up both the changes in the slope of the original data, which is an important feature of the fitted line to show improvement or sustained running form. Further developments of an algorithm of the fitted line may remove any negative values or re-set them to 0%.

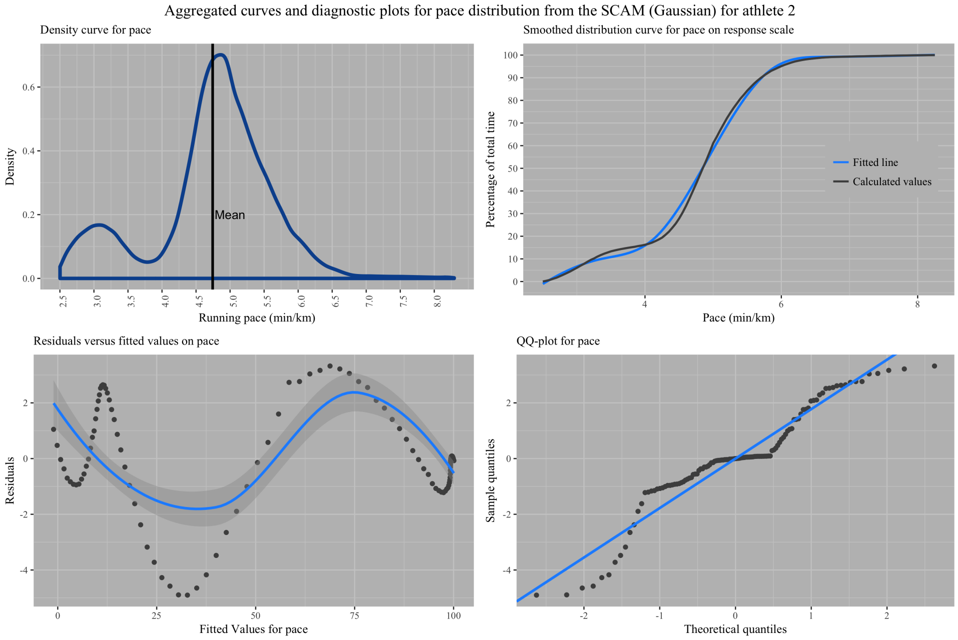
Figure 6.18 shows how the selected model is broken down into the running activities. The bi-modality of the pace is apparent in the density plot (left graph in Figure 6.18). Track running is bi-modal with the first peak higher than the second. Road race peaks before road running, indicating that the runner generally runs faster during a race than during training runs on the road. Both road running and road racing present with a small first peak at 3.25 min/km, perhaps indicative of the runner's ability to push a final kick at the end of the race or training road run. The bump in the distribution curve (right graph in Figure 6.18) for pace corresponds to the first peak from the density plot. The athlete spends more than half of his time running faster than 4 min/km during track training. The curve for road racing is shifted to the left of road running and remains as such. This implies that the athlete is capable to run at faster paces for longer during races than during roads runs. The shift of this curve away from road running and towards track training may show that the training work on the track is indeed improving his racing pace.

6.4 Case study D: the heart-rate runner

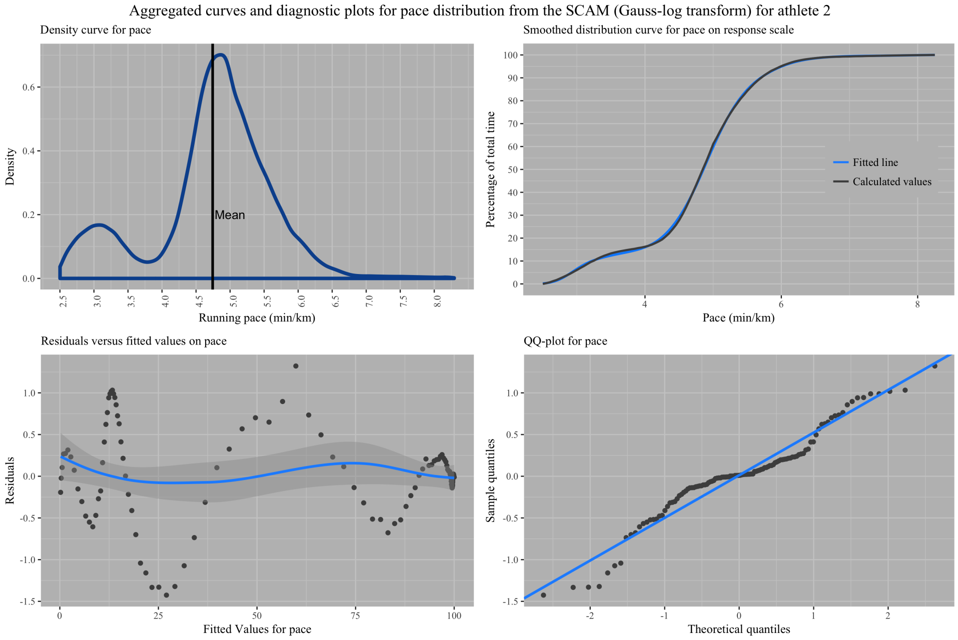
Table 6.4 contains the summary of the model adequacy measures for the distribution regression models on cadence and pace.

6.4.1 Analyses of the GAM and SCM on cadence

Figure 6.19 shows the histogram, the fitted distribution curve and the diagnostic plots for both the Gaussian model family and the Gauss-log transformed model on the distribution for cadence from the GAM. The histogram for cadence is slightly skewed to the right but there is no bi-modality present. There is a slight slow down of the descent of the distribution curve's original data between a cadence of 92 and 93, corresponding with the almost even bar heights at these cadences in the histogram. The fitted line for the



(a) Gaussian model



(b) Gauss-log transformed model

Figure 6.17: The distribution curves for pace from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 2.

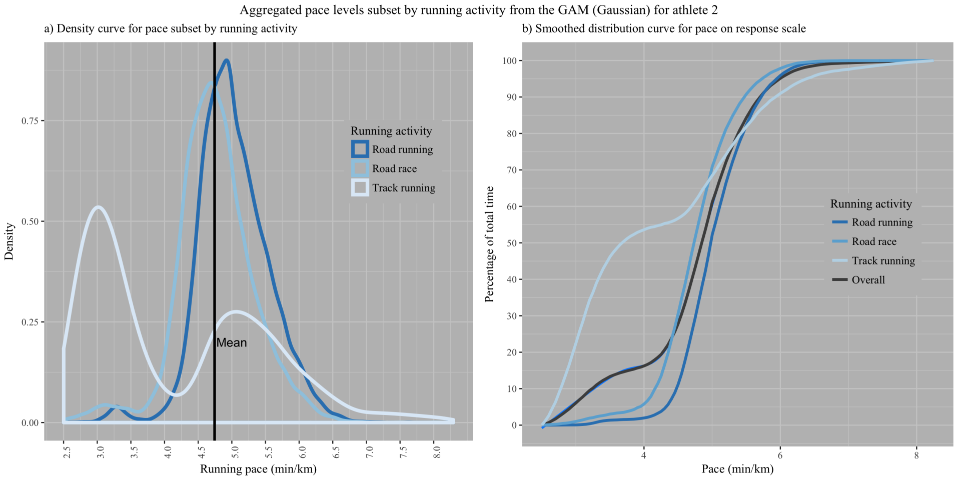


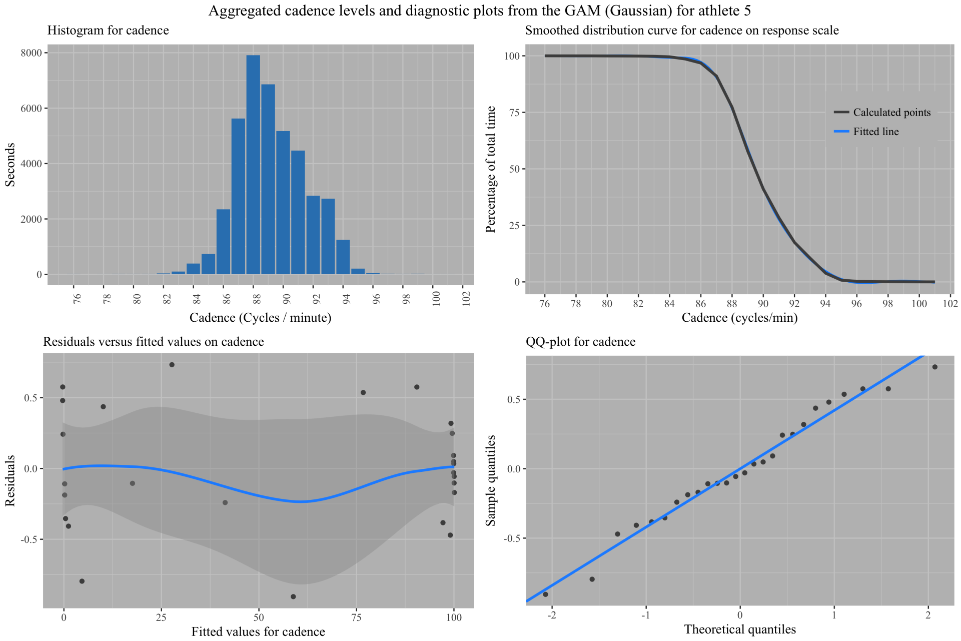
Figure 6.18: The distribution of pace subset by running activity for athlete 2

Table 6.4: Model adequacy measures for the distribution regressions for athlete 5

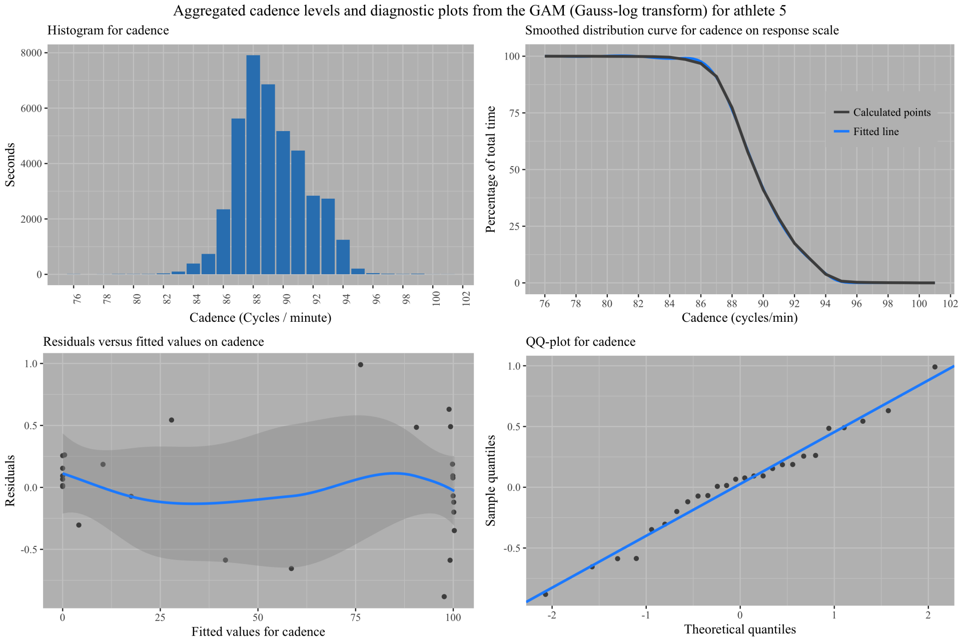
Variable	AIC score	R^2	Negative values	Monotonicity score	Technique	Distribution
Cadence	49.60	0.9999	3	5	gam	gaussian
Cadence	48.43	0.9999	0	4	gam	gauslog
Pace	-53.07	1.0000	4	5	gam	gaussian
Pace	-207.69	1.0000	0	0	gam	gauslog
Cadence	98.95	0.9989	5	0	scm	gaussian
Cadence	100.60	0.9989	0	0	scm	gauslog
Pace	197.82	0.9998	18	0	scm	gaussian
Pace	-115.43	1.0000	0	0	scm	gauslog

Gaussian model (top right in Figure 6.19a) does well to pick up this slight change, although it underestimates it by a small margin. The fitted line does go below the 0% line, which is undesirable. The Gauss-log transform (top right in Figure 6.19b) does a slightly better job at picking up the decrease in the slope between the cadences of 92 and 93 and did not fit any negative values. However, it seems to be unstable at the higher percentages and is not be monotonic, which is undesirable. Neither of these models' error distributions are normal with the errors near the 100% fitted values in the residual plots almost forming a vertical line .

Figure 6.20 shows the density, distribution and diagnostic plots for cadence from the SCM for both the Gaussian and the log-transformed models. The Gauss-log transform model (top right in Figure 6.20b) completely misses the deceleration of the slope in the critical region and is therefore not considered any further. The Gaussian model (top right in Figure 6.20a) does well to match the change in the slope in the critical region between a cadence of 92 and 93, but over-estimates the data for the cadences up to 85, after which it over-estimates the sharp negative increase in the slope between 86 and 88. Further it performs relatively well except for the negative fitted values at the upper range of cadence. The error terms show great departure from normality, both in the residual



(a) Gaussian model



(b) Gauss-log transformed model

Figure 6.19: The distribution curves for cadence from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 5.

and the qq-plots.

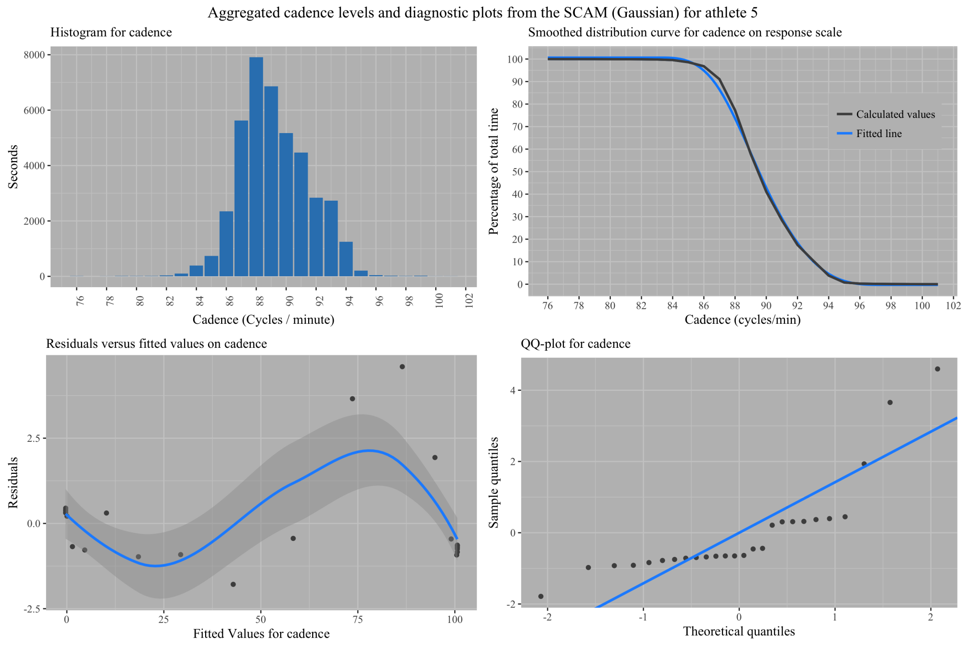
The selected model is the Gauss-log GAM with it's breakdown into the running activities shown in Figure 6.21. Despite it not being monotonic throughout the course of the line, it still presented with the lowest AIC score and picked up the changes in the slope in the critical regions. Road racing is responsible for the right-skewness seen in the histogram and is also skewed to the right itself (left graph in Figure 6.21). Trail running has a long tail to the right and peaks at the same cadence level for the overall data. The skewness is reflected in the distribution curves (right graph in Figure 6.21). Trail running overtakes road running percentages at cadence at the 88 mark. Road racing stays at a higher cadence for longer than both road and trail running. This pattern is different to those seen in both athletes 3 and 4, where the trial running line remains below road running. This might imply that the athlete's capability to run at higher cadences is better during trail running than during road running, albeit the smaller proportion of total time is spent on trail running.

6.4.2 Analyses of the GAM and SCM for the distribution on pace

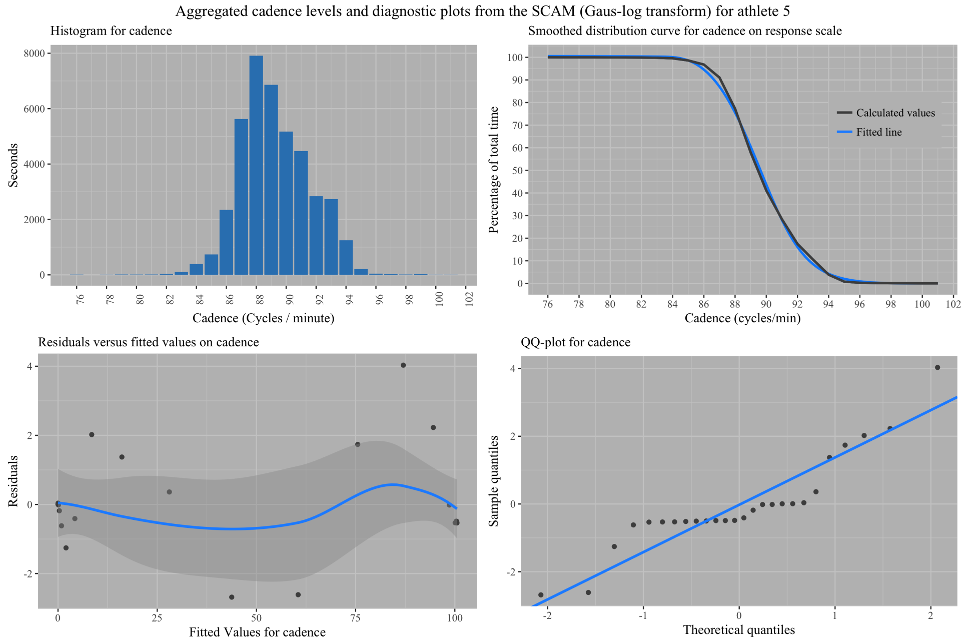
Figure 6.22 shows the density- and distribution curves, residual plots and the qq-plots for pace from the GAMs for both the Gaussian family and the Gauss-log transform. The density curve for pace is slightly skew to the right with a small bump observed at the intersection with the mean. This bump is virtually negligible in the distribution curve. The fitted lines from both the models are a near perfect fit to the original data, however, the Gauss-log model is a better fit.

Figure 6.23 shows the density-, distribution curves and the diagnostic plots for the distribution on pace from the SCMs. The curvi-linear, cyclical pattern in the Gaussian model's residual plot echos the fluctuation of the fitted line around the original data (top right in Figure 6.23a). The Gauss-log line fits the original line well with almost no visible fluctuation around the original data (top right in Figure 6.23b).

From Table 6.4 the Gauss-log transform from the GAM has the lowest AIC score, is perfectly monotonic and fitted no negative values. Neither the Gaussian models from the GAM nor the SCM are suitable choices. The Gauss-log model from the SCM is the next best model to fit the data, based on the AIC alone. The Gauss-log GAM is the selected model for this athlete's distribution of pace, based on its lowest AIC score and its capability to fit the slight change in the slope of the curve. Figure 6.24 shows how the selected model is broken down into the running activities. The same type of picture emerges from the density curve as with the cadence histogram. Trail running presents



(a) Gaussian model



(b) Gauss-log transformed model

Figure 6.20: The distribution curves for cadence from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 5.

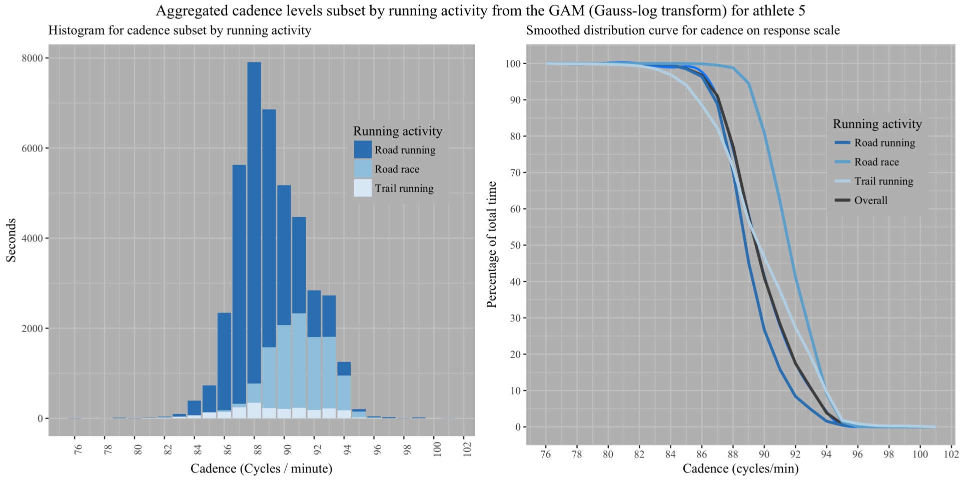
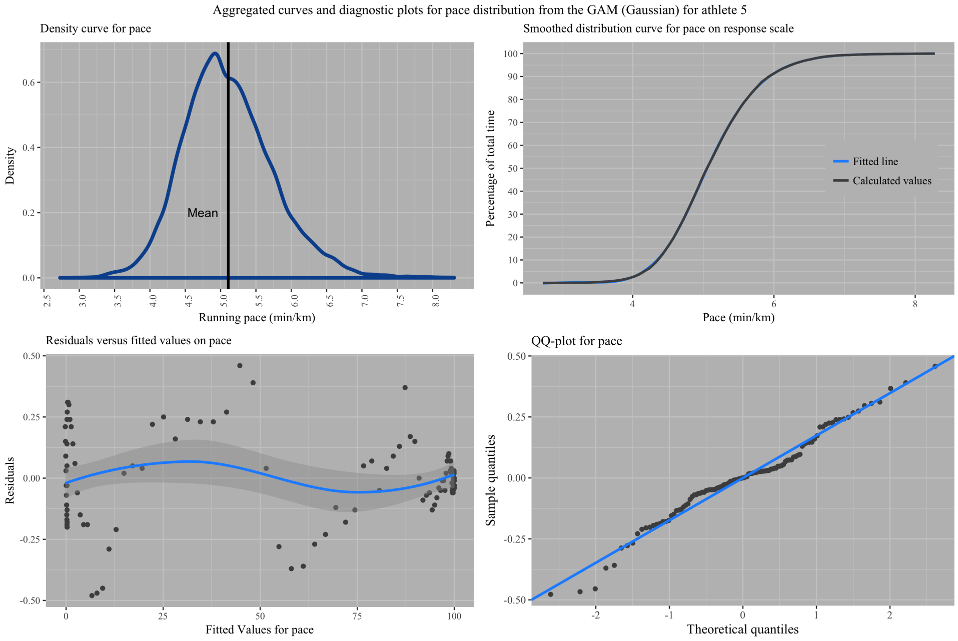


Figure 6.21: The distribution of cadence subset by running activity for athlete 5.

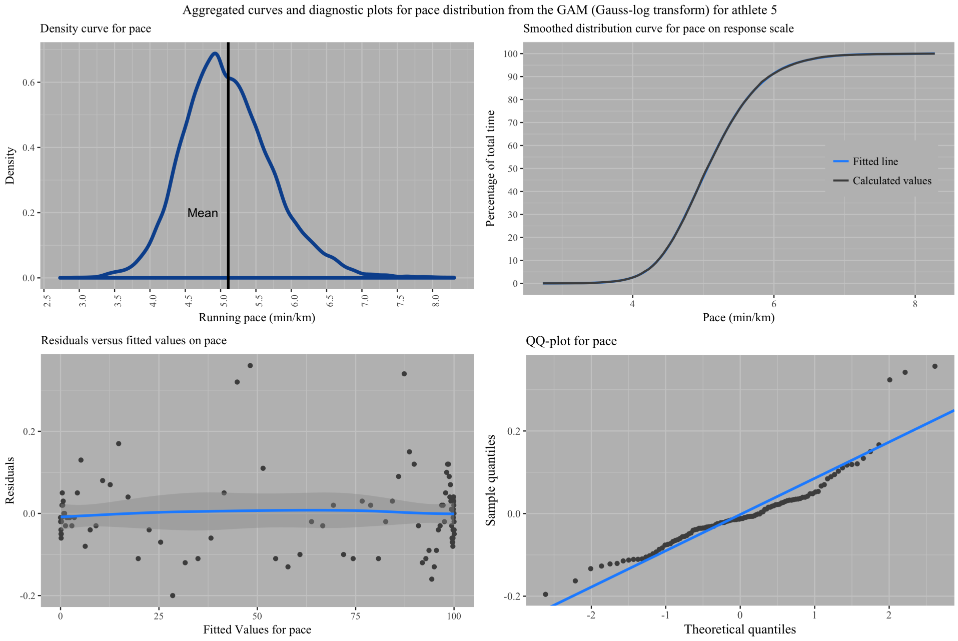
with a long tail to the right and is spread wider than the other two running activities (left graph in Figure 6.24). The peak is also to the left of the other density curves. The shape of the trail running density is echoed in the distribution curve (right graph in Figure 6.24). The descent of the distribution curve towards the faster paces is faster than the descent of the other two activities. The curves cross from the 5 min/km mark. The athlete is capable to spend in the region of 30% percentage of his time at running faster than roughly 4.5 min/km during trail running than during road running and road racing. The crossing of the curves may indicate that the athlete is better at maintaining faster paces during a trail run than during road running.

6.5 Conclusion

The distribution curves are useful to analyse the runner's overall performance and is an effective visualiser of running form, for both cadence and pace. They provide a bird's eye view of the athlete's running form when all of the data is considered and subset into running activities. The estimation of the probability that an athlete will achieve a certain cadence or running pace and maintain it for longer periods of time is satisfactory accurate. The movements of these curves are important to monitor, as it is the shift of the curve in either direction as a whole that is indicative of a runner's improved capabilities or perhaps declining capabilities. Small changes in the slope of the curves are also important, as it indicates improved or declined performances at certain levels. These changes in the slope are pointers to changes in the distribution of the variable, as the underlying data might become bi-modal or even skew. The changes in the underlying data's distribution provide insight into how the athlete is performing on different running surfaces. Skewness of the data may be indicative of outliers that have been included in the data set and need to be

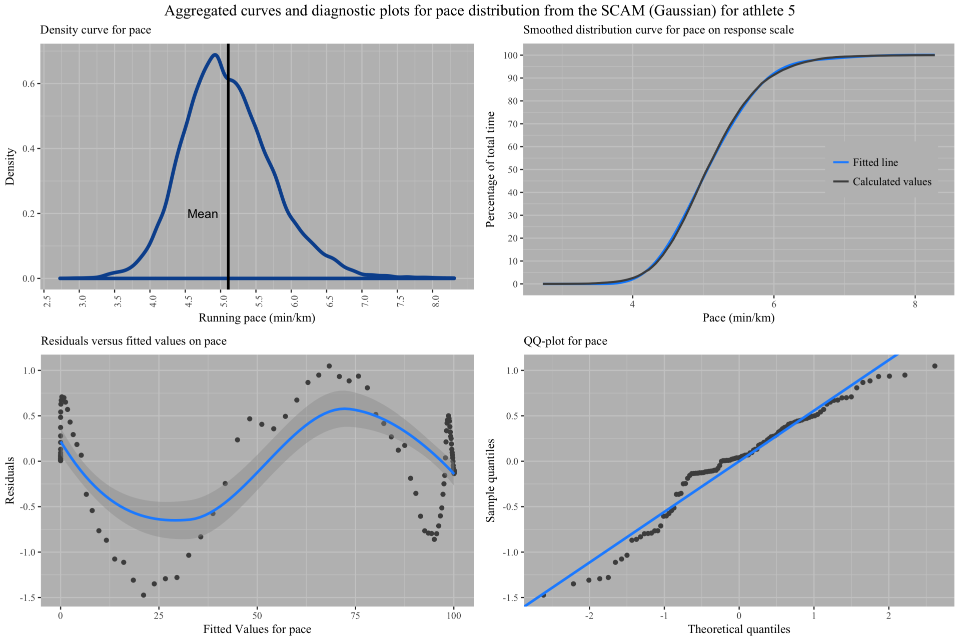


(a) Gaussian model

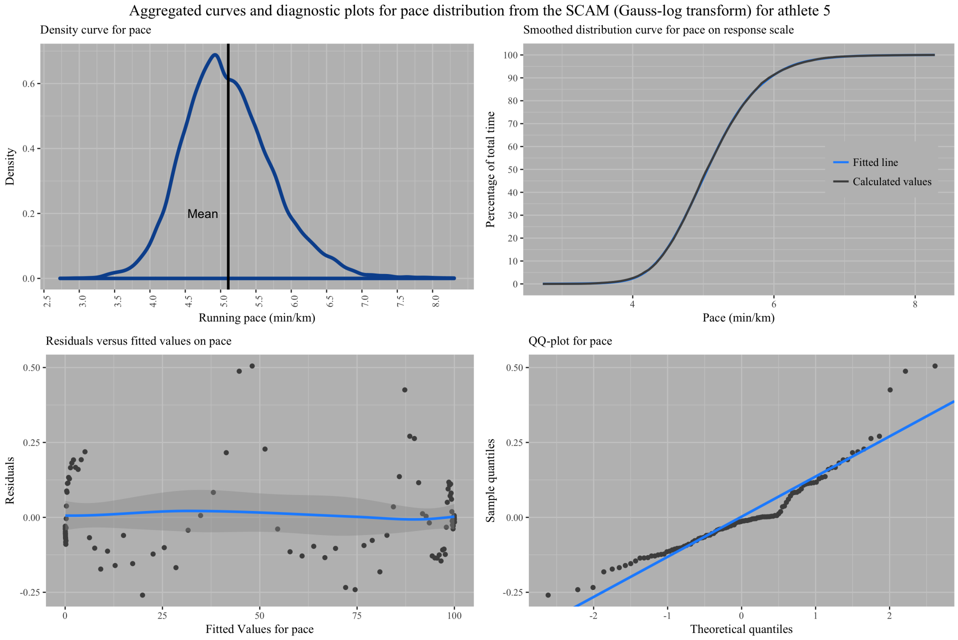


(b) Gauss-log transformed model

Figure 6.22: The distribution curves for pace from the GAM for the Gaussian family and the Gauss-log transformed data for athlete 5.

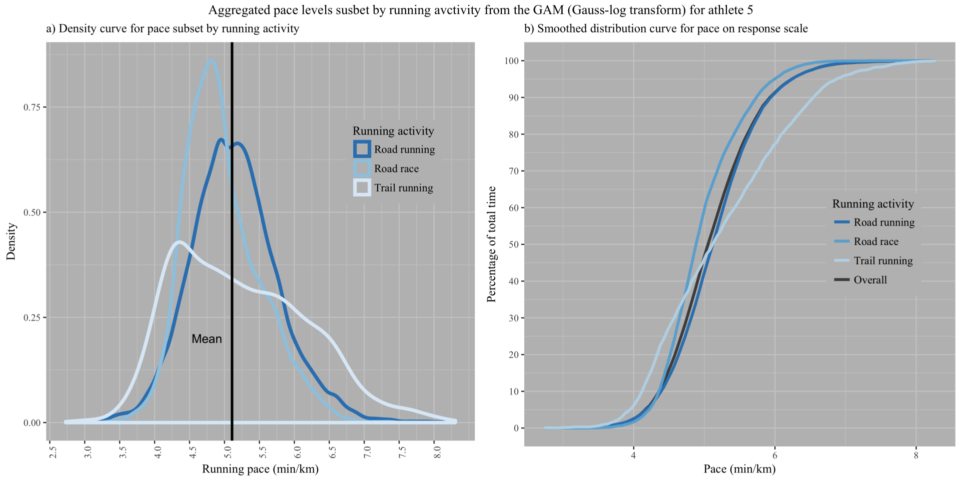


(a) Gaussian model



(b) Gauss-log transformed model

Figure 6.23: The distribution curves for pace from the SCM for the Gaussian family and the Gauss-log transformed data for athlete 5.


Aggregated pace levels susbet by running avctivity from the GAM (Gauss-log transform) for athlete 5

a) Density curve for pace subset by running activity

b) Smoothed distribution curve for pace on response scale

Figure 6.24: The distribution of pace subset by running activity for athlete 5

removed, or it is a sign of momentarily or unexpected changes in performance.

The GAMs outperformed almost all of the SCMs with regards to the AIC scores. The SCMs did worse in the critical regions of the curves and proved to be the least sensitive to pick up the declining rate of change. This insensitivity may probably attributed to the penalty imposed by the model in order to maintain a perfectly monotonic curve. The GAM curves were not all perfectly monotonic and even some of the selected models failed in this adequacy check. The Gauss-log transformed GAMs for cadences were selected for three out of the fours case studies, with only athlete 4 (case study B) having selected the Gaussian GAM for the distribution on cadence. The Gauss-log transformed GAMs for pace were selected for three of the cases studies, with only athlete 2 (case study C) having selected the Gaussian version. All the models' error behaviours were captured in the residual- and qq-plots. The residual plots showed curvi-linear behaviour and the errors continuously departed from the straight line in the qq-plots. The pattern of departure from the straight lines in the qq-plots showed tails as well as a swirling type of behaviour, where the error terms remained on one side of the straight line for a continuous section of the theoretical quantiles and then move across to the opposite side of the line. This type of behaviour is evident of bi-modality in the error terms. The errors for the models are not normally distributed. Most of the residual plots showed an increase of the error term margin in the middle range and lower error terms on the ends of the fitted percentages.

These models serve as estimators for a runner's sustained and overall running form. Cadence is measured as stride cycles per minute, and not steps per minute. To get to steps per minute, cadence must be multiplied by two, which implies that odd step numbers will be excluded from the total count. The continuous nature of the fitted curve is able to fill the gaps for the odd step numbers. The distribution curves may thereby make it possible for an athlete to see step-wise improvement and not necessarily improvements consisting of entire cadences.

The subsetting of the density and the distribution curves into running activities extended the variability of an athlete's running form. It clearly demonstrated that the athletes capabilities vary across different surfaces and under different conditions. A single figure such as an average pace or cadence can misleading when evaluating overall running form as it is not capable to show the athlete small gradual improvements or point out dissimilar performances during running. As this was done for both cadence and pace it is considered as a step towards the development of a model that can monitor running form in the field and progress towards functional regression models where the running form distribution profiles may serve as response or explanatory variables. Future work on the selection of the number of knots is suggested, as well as easing the penalty on the SCM to be continuously monotonic. This will still force a monotonic curve but one that might be sensitive enough to pick up the changes in the slopes of the distribution curves. The combination of the two pace and cadence distribution curves into a tool where "cut-off"-points for the range of pace per cadence can be established may aid in the analysis of running form in Chapter 4.

Chapter 7

Final remarks

The data available from a GPS running watch lend itself to the analyses of the associations and relationships of multiple variables to monitor both internal and external physical loads during running. The unmitigated volume and velocity of the data that are available for analysis can become overwhelming and complex to truly understand. The discipline of data science provides analysts with rigorous and scientific techniques to organise and transform the raw data into mathematical and statistical models that can be used for innovative approaches to training and tactics. Some alternative and different approaches in the analysis and visualisation of the data from the runners watches have been presented in this study. However, the main drive and deliverable for this project are not the analytical models themselves, but instead the usability of the explored data sets with the future purpose of developing meaningful models and data presentation techniques with an acceptable range of accuracy. Despite the study's limitation by its observational nature and dependence on how the athlete uses his device, the overall results and the image of the data that emerged look promising for future work to continue.

The multiple linear regression analysis of cadence, running activity and pace in Chapter 4 produced the interaction effects between cadence and the running activity and their combined effect on pace. The spread of pace per cadence as visualised in the scatter plots and boxplots is extensive and skew, both towards the left and right of the modes. The boxplots for pace per cadence were able to capture the pattern in the data, more so than the scatter plots. The clean-up of the data resulted in more concentrated boxplots per cadence, however the spread mostly remained large and skew, with the exceptions of the higher end cadences. Most of the interaction terms were significant after the data had been cleaned from the outliers. Although this analysis showed that the data from the watch has potential to be used to assess athletes responses to different surfaces and under different conditions, more work is needed to validate the combinations of cadence and pace.

The visual analysis of graded running in Chapter 5 proved to be challenging. The 2D visualisations displayed unexpected symmetry with regards to pace during UR and DR. The cause for the symmetry has been discussed in the chapter, but there is no definitive answer to the observed symmetry. The interaction effects from the log-transformed linear model supported the visual indications from the 2D models that pace becomes faster nearer the 0% grade line, despite the fact that most athletes' fastest mean pace is achieved during DR. The 3D output from the TPS model may be a tool to help a runner understand how he/she runs on slopes, but is subject to uncertainty as the surface of the plate presents the expected pace for each combination of grade and cadence. Some athletes showed weak correlation between cadence and grade, which might give rise to some multicollinearity in the polynomial multiple regression model.

Aggregating the pace and cadence data in Chapter 6 present the athlete with a holistic view on their running form. The GAMs and SCMs proved to be useful and fitted the calculated percentages very well, however the SCM did not pick up the changes in slope in the critical regions. The aggregation may be seen as an extension of the running form analysis in Chapter 4. The subset of the data into the running activities clearly distinguished their running form with the reference to road run training. Monitoring the shifts of the distribution curves and the changes in slopes may be a useful tool to assess overall running form and improvements.

The lifestyle athlete, the recreational runner and professional elite runners all have a common goal pertaining to running: to improve fitness and race outcomes and prevent injuries. An athlete with a good understanding of his/her own physical capabilities allows for realistic race goals, strategies and preparation. The data visualisations that form part of this approach's outcome may give an athlete a graphical tool to quickly determine the state of their running form and whether they are improving. It may also help them to identify areas of regression into poorer running form, which may serve as an early warning system. These visualisations may form the basis for new data presentation practices on the online fitness applications. Athletes may be able to independently monitor their running form and self-manage training methods to improve or sustain a good running form.

The outcome from this project adds to the body of existing knowledge and techniques that leverage the large data sets to obtain insight and make wise decisions pertaining to general running heath, training regimens and race tactics. The study will open up more discussion points and opportunities for the utility of big data and data science application in running and sport in general. Whereas this study focused on running form as a function of cadence, pace and gradients of slope, further work to include data on ground contact time and vertical oscillation is encouraged to create an even better analytical model on running form. Should these analysis approaches be endorsed as accurate and scientific medical methods it may also become tools in the hands of coaches during training and sport therapists for effective rehabilitation of bio-mechanical related injuries from running. As Sands et al. (2017) rightly states: the monitoring tools in sport science bear more advantages for the athlete than prediction functions. To assist athletes in planning and executing their training under changing conditions is a better option than long-term prediction of performance.

Alternatives to evaluating running form are scientific tests in laboratories or on-field assessments that are not always accessible nor practical, especially for the lifestyle or recreational runner. With the development and availability of fitness technology on runners' wrists, more and more runners now have access to important metrics that provide information on running form and physiological status during activities. The tracking technology is opening up research avenues in sport science and physical activity that have not been accessible or even comprehensible before. Nonetheless, in the midst of all the excitement and decision support that the terabytes of numbers provide to researchers, hardware and software developers, practitioners and athletes, we must not loose sight of one thing: sport is there for humans to enjoy, be it as participants or spectators. Sport fulfills the human body's inherent design to move at will, and not to be a marionette who plays by numbers.

Bibliography

Adamakis, M. and Zounhia, K. (2016). Validity of wearable activity monitors and smartphone application for step counting in adolescents. In *FIEP European Congress*, Banja Luka.

Adams, D., Pozzi, F., Carrol, A., Rombach, A., and Zeni, J. (2016). Validity and reliability of a commercial fitness watch for measuring running dynamics. *Journal of Orthopaedic and Sports Physical Therapy*, 46:1–19.

Ahmad, Z., Jamaludin, N., and Hafidz Omar, A. (2018). Monitoring and prediction of exhaustion threshold during aerobic exercise based on physiological system using artificial neural network. *Journal of Physical Fitness, Medicine and Treatment in Sports*, 3:1–3.

Ahmed Memon, M., Soomro, S., Khan Jumani, A., and Kartio, M. (2017). Big data analytics and its applications. *Annals of Emerging Technologies in Computing*, 1.

Balaban, O. and Tuncer, B. (2016). Visualizing Urban Sports Movement. In *34th eCAADe Conference*, volume 2, Oulu, Finland. eCAADe. Project: FCL - Big Data Informed Urban Design - Evidence Informed Design and Planning Processes.

Balaban, O. and Tuncer, B. (2017). Visualizing and analyzing urban leisure runs by using sports tracking data. *City modelling tools*, 1:533–535.

Barrio, I., Arostegui, I., and Quintana, J. M. (2013). Use of generalised additive models to categorise continuous variables in clinical prediction. *BMC medical research methodology*, 13:83.

Belongie, S. (2018). *Thin Plate Spline*. `http://mathworld.wolfram.com/ThinPlateSpline.html`, [Online], Accessed 12 April 2018.

Bourdon, P., Cardinale, M., Murray, A., Gastin, P., Kellmann, M., Varley, M., Gabbett, T., Coutts, A., Burgess, D., Gregson, W., and Cable, N. (2017). Monitoring athlete training loads: Consensus statement. *International Journal of Sports Physiology and Performance*, 12:161.

Brukner, P. and Khan, K. (2006). *Clinical Sports Medicine*. McGraw-Hill Companies.

Burden, R. L., Faires, J. D., and Burden, A. (2016). *Numerical Analysis*. MA: Cengage Learning, 10 edition.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, 2 edition.

Chen, S., Lach, J., Lo, B., and Yang, G.-Z. (2016). Toward pervasive gait analysis with wearable sensors: A systematic review. *IEEE Journal of Biomedical and Health Informatics*, PP:1537.

Chen, T., Nosaka, K., and Wu, C. (2008). Effects of a 30-min running performed daily after downhill running on recovery of muscle function and running economy. *Journal of science and medicine in sport*, 11:271–9.

Chen, Y. and Samworth, R. J. (2016). Generalized additive and index models with shape constraints. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):729–754.

Comrades (2017). *Results History*. Comrades. `http://results.ultimate.dk/comrades/resultshistory/front/index.php?results=true`, [Online]. Accessed 24 January 2018.

Comrades (2018). *Comrades 2018 Entry cap increased to 21500*. `http://news.comrades.com/index.php/media-releases/737-comrades2018-entry-cap-increased-to-21-500`, [Online]. Accessed 24 January 2018.

Cortes, R., Bonnaire, X., Marin, O., and Sens, P. (2014). Sport trackers and big data: Studying user traces to identify opportunities and challenges. *Procedia Computer Science*, 52:1004 – 1009.

Denny, M. (2017). The fallacy of the average: on th eubiquity, utility and continuing novelty of jensen's inequality. *Journal of experimental biology*, 220:139–146.

Dominici, F., McDermott, A., Zeger, S. L., and Samet, J. M. (2002). On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health. *American Journal of Epidemiology*, 156(3):193–203.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(3):89 – 121.

Garmin (2018). *Activity tracking accuracy.* Garmin. `https://www.garmin.com/en-US/legal/atdisclaimer`, [Online]. Accessed 29 January 2018.

Gijon-Nogueron, G. and Fernandez-Villarejo, M. (2015). Risk factors and protective factors for lower-extremity running injuries: a systematic review. *Journal of the American Podiatric Medical Association.*

Gupta, S. and Schneider, M. (2018). *Protecting Customers' Privacy Requires More than Anonymizing Their Data.* Harvard Business Review. `ProtectingCustomers\OT1\textquoterightPrivacyRequiresMorethanAnonymizingTheirData`, [Online]. Accessed 4 August 2018.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–310.

Heiderscheit, B. C., Chumanov, E. S., Michalski, M. P., Willie, C. M., and Ryan, M. B. (2011). Effects of step rate manipulation on joint mechanics during running. *Medicine and Science in Sport and Execrise*, 43:296 – 302.

Hendricks, C. and Philips, J. (2013). Prevalence and incidence rate of injuries in runners at a local athletic club in Cape Town. *South African Journal of Physiotherapy*, 69(3).

Hochmair, H. H., Bardin, E., and Ahmouda, A. (2017). Estimating bicycle trip volume for miami-dade county from strava tracking data. In *The National Academy of Sciences, Engineering and Medicine: Transportation Research Board*, Washington, D.C. Transport Research Board.

Hofner, B., Kneib, T., and Hothorn, T. (2014). A unified framework for contrained regression. Technical report, Cornell University Library.

IAAF (2016). *RIO 2016 Womens 100m.* International Association of Athletics Federations. `https://www.iaaf.org/news/report/rio-2016-womens-100m1`, [Online]. Accessed 25 November 2017.

IAAF (2017a). *Keitany breaks women's-only world record at London marathon.* International Association of Athletics Federations. `https://www.iaaf.org/news/report/london-marathon-2017-keitany-world-record`, [Online]. Accessed 18 January 2018.

IAAF (2017b). *Results.* International Association of Athletics Federations. `https://www.iaaf.org/results/iaaf-world-championships-in-athletics/2017/iaaf-world-championships-london-2017-5151/women/100-metres/`, [Online]. Accessed 25 November 2017.

IDC (2017). *Worldwide wearables market grows 7.3 percent in Q3 2017.* International Data Corporation. `https://www.idc.com/getdoc.jsp?containerId=prUS43260217`, [Online]. Accessed 24 January 2018.

Kaushik, A. (2016). *A Great Analyst's Best Friends: Skepticism and Wisdom.* `https://www.kaushik.net/avinash/great-analyst-skills-skepticism-wisdom/`, [Online]. Accessed 24 January 2018.

Kerren, A., Plaisant, C., and Stasko, J. T. (2011). Information visualization: State of the field and new research directions. *Information visualization*, 10(4).

Kosmidis, I. and Passfield, L. (2015). Linking the performance of endurance runners to training and physiological effects via multi-resolution elastic net. Technical Report arXiv:1506.01388, Cornel University Library.

Kruger, P. S. and Yadavalli, V. S. S. (2016). Probability management and the flaw of averages. *South African Journal of Industrial Engineering*, 27(4):1 – 17.

Mahmoudi, N., Reza Ahmadi, M., Babanezhad, M., and Seyfabadi, J. (2014). Environmental variables and their interaction effects on chlorophyll-a in coastal waters of the southern caspian sea: Assessment by multiple regression grey models. *Aquatic Ecology*, 48:351–365.

Mašić, A., Srinivasan, S., Billeter, J., Bonvin, D., and Villez, K. (2016). On the use of shape-constrained splines for biokinetic process modeling. *IFAC-PapersOnLine*, 49:1145–1150.

Montgomery, D. C., Peck, E. A., and Vining, G. (2006). *Introduction to linear regression analysis.* John Wiley and Sons, Inc.

Montgomery, D. C. and Runger, G. C. (2011). *Applied Statisitics and Probability for Engineers.* John Wiley and Sons, Inc.

Monti, D. (2017). Mary keitany breaks all-women's world record at London marathon. `http://running.competitor.com/2017/04/news/mary-keitany-breaks-womens-world-record-london-marathon_164039`, [Online]. Accessed on 18 January 2018.

Napier, C., Esculier, J. F., and Hunt, M. A. (2017). Gait re-training: out of the lab and into the streets with the benefit of wearables. *British Journal of Sports Medicine.*

Neptune, R. R. and Sasaki, K. (2005). Ankle plantar flexor force production is an important determinant of the preferred walk to run transition speed. *Journal of Experimental Biology*, 208(5):799–808.

Neter, J., Wasserman, W., and Whitmore, G. A. (1988). *Applied Statistics.* Allyn and Bacon, Inc.

OMTOM (2018a). *Old Mutual Two Oceans Marathon History Since 1970.* Old Mutual Two Oceans Marathon. `http://www.twooceansmarathon.org.za/history`, [Online]. Accessed 19 January 2018.

OMTOM (2018b). *Old Mutual Two Oceans Marathon Tackles Cape Water Crisis.* Old Mutual Two Oceans Marathon. `http://www.twooceansmarathon.org.za/news/old-mutual-two-oceans-marathon-tackles-cape-water-crisis`, [Online]. Accessed 6 February 2018.

Ortega, F. B., Ruiz, J. R., Hurtig-Wennlöf, A., Vicente-Rodríguez, G., Rizzo, N. S., Castillo, M. J., and Sjöström, M. (2010). Cardiovascular fitness modifies the associations between physical activity and abdominal adiposity in children and adolescents: the european youth heart study. *British Journal of Sports Medicine*, 44(4):256–262.

Otto, S. (2012). *General additive models and their application in modelling zooplankton lifecycle dynamics.* Stockholm University. `http://www.su.se/ostersjocentrum/english/beam/outreach/films/filmed-lectures-modeling-as-a-tool-to-study-the-baltic-ecosystem-1.170066`, [Online]. Accessed 29 January 2018.

Padulo, J., Powell, D., Milia, R., and Ardigo, L. (2013). A paradigm of uphill running. *PloS one*, 8:e69006.

Passfield, L. and Hopker, J. G. (2016). A Mine of Information: Can Sports Analytics Provide Wisdom From Your Data? *International journal of sports physiology and performance*, pages 1–7.

Patrikalakis, N. M., Maekawa, T., and Cho, W. (2009). *Shape interrogation for Computer aided design and manufacturing.* MIT. Avaiable as a hyperbook edition from `http://web.mit.edu/hyperbook/Patrikalakis-Maekawa-Cho/mathe.html`. Accessed 18 April 2018.

Peterson, C. (2016). Visualization for all: The importance of creating data representations patients can use. In *2016 Workshop on Visual Analytics in Healthcare*, Chicago.

Pileggi, H., Stolper, C., Boyle, J., and Stasko, J. (2012). SnapShot: Visualization to Propel Ice Hockey Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2819 – 2828.

Pya, N. and Wood, S. N. (2015). Shape constrained additive models. *Statistics and Computing*, 25:543 – 559.

Rein, R. and Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for Sport Science. *SpringerOpen*, 5.

Rodríguez-Álvarez, M., Cadarso-Suárez, C., and González, F. (2012). Analysing visual receptive fields through generalised additive models with interactions (invited article with discussion: MarÁa l. durbán and thomas kneib). *SORT. Statistics and Operations Research Transactions*, 1.

RunningUSA (2016). *2016 State of the Sport: U.S. Road Race Trends.* Running USA. `http://www.runningusa.org/state-of-sport-us-trends-2015`, [Online]. Accessed 24 January 2018.

Sands, W. A., Kavanaugh, A. A., Murray, S. R., McNeal, J. R., and Jemni, M. (2017). Modern Techniques and Technologies Applied to Training and Performance Monitoring. *International Journal of Sports Physiology and Performance.*

Savage, S. (2000). *The Flaw of averages.* Stanford management science and engineering. `https://web.stanford.edu/~savage/faculty/savage/FOA%20Index.htm`, [Online]. Accessed 18 January 2018.

Schubart, A. G., Kempf, J., and Heiderscheit, B. C. (2013). Influence of stride frequency and length on running mechanics: a systematic review. *Sports Health*, 6.

Shiach, J. (2015). *B-Splines.* School of Computing Mathematics and Digital Technology, Manchester Metropolitan University. Online lecture at `https://ru-clip.net/video/qhQrRCJ-mVg/b-splines.html`, accessed 18 April 2018.

Smith, N. J. (2015). *Spline regression.* Patsy. `http://patsy.readthedocs.io/en/latest/spline-regression.html`, [Online], accessed 19 April 2018.

Snyder, K. and Farley, C. T. (2011). Energetically optimal stride frequency in running: the effects of incline and decline. *The Journal of Experimental Biology*, 214:2089–95.

Ullrich-French, S. and Smith, A. (2009). Social and motivational predictors of continued youth sport participation. *Psychology of Sport and Exercise*, 10:87–95.

Vaidyanathan, R. (2012). *Piecewise regression with R: plotting the segments.* `https://stackoverflow.com/questions/8758646/piecewise-regression-with-r-plotting-the-segments/`, [Online],Accessed 29 January 2018.

Vernillo, G., Giandolini, M., Edwards, W. B., Morin, J.-B., Samozino, P., Horvais, N., and Millet, G. (2016). Biomechanics and physiology of uphill and downhill running. *Sports Medicine.*

Wilkerson, G. B., Gupta, A., Allen, J. R., Keith, C. M., and Colston, M. A. (2016). Utilisation of Pratice Session Average Inertial Load to Quantify College Football Injury Risk. *Journal of strength and conditioning research*, 30.

Wisbey, B., G Montgomery, P., Pyne, D., and Rattray, B. (2009). Quantifying movement demands of afl football using gps tracking. *Journal of Science and Medicine in Sport*, 13:531–6.

Wood, S. N. (2006). *Generalized Additive Models: A Introduction with R.* Chapman and Hall CRC.

Appendix A

Profile (Euodia Vermeulen)

A.1 Education

- Bachelors in Physiotherapy (BPhysT) at the University of Pretoria, 2010
- Highest Engineering degree: BEng Honors (Industrial) at the University of Pretoria, 2016

A.2 Research interests

- Health, sport and medical related research and development
- Data science, data analytics and data visualisation
- Simulation modeling

A.3 Publications

The following article has been accepted for oral presentation and publishing in the 2018 IEOM African Conference on Industrial Engineering and Operations Management, Pretoria. Papers are indexed in Scopus. Vermeulen, E and Yadavalli, V.S.S., Big data in sport: application and risks.

The following article has been communicated to the IEEE Transactions on Big Data: Vermeulen, E and Yadavalli, V.S.S., Harnessing big data from fitness trackers to visually analyse overground running: a case study.

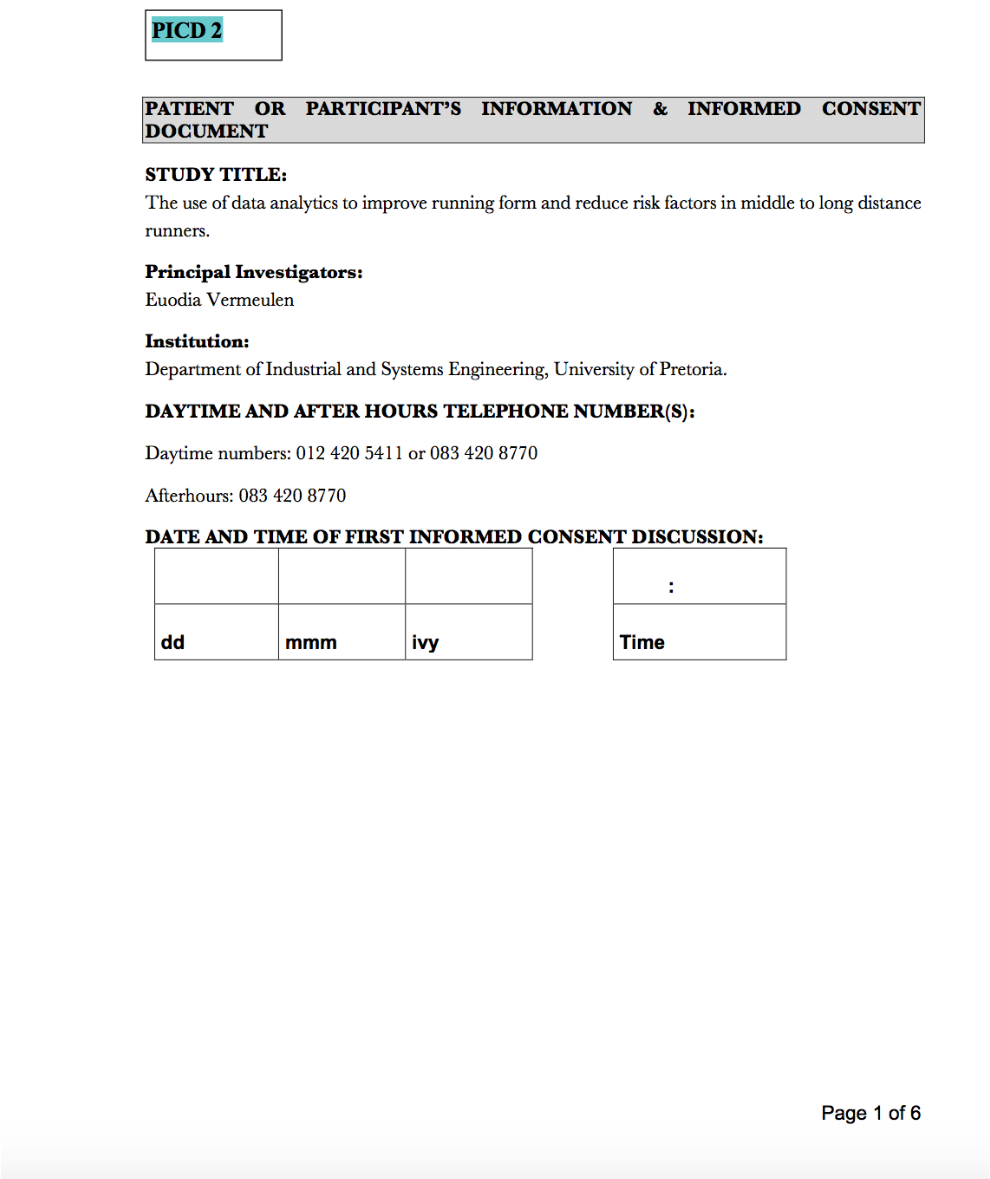
A.4 Awards

- South African Institute of Industrial Engineering awards for best final year student for Industrial Engineering at the University of Pretoria (2015)
- First National Bank award for most consistent academic achievement over four years of study towards BEng: Industrial (2015).

Appendix B

Informed consent form

Athletes completed the informed consent form before the start of data extraction.



PICD 2

PATIENT OR PARTICIPANT'S INFORMATION & INFORMED CONSENT DOCUMENT

STUDY TITLE:

The use of data analytics to improve running form and reduce risk factors in middle to long distance runners.

Principal Investigators:

Euodia Vermeulen

Institution:

Department of Industrial and Systems Engineering, University of Pretoria.

DAYTIME AND AFTER HOURS TELEPHONE NUMBER(S):

Daytime numbers: 012 420 5411 or 083 420 8770

Afterhours: 083 420 8770

DATE AND TIME OF FIRST INFORMED CONSENT DISCUSSION:

dd	**mmm**	**ivy**

:
Time

Page 1 of 6

Dear Athlete,

Dear Mr. / Mrs / Miss / Ms. …………………………….

date of consent procedure …….. /………. /…......

1) INTRODUCTION

You are invited to volunteer for a research study. This information leaflet is to help you to decide if you would like to participate. Before you agree to take part in this study you should fully understand what is involved. If you have any questions, which are not fully explained in this leaflet, do not hesitate to ask Euodia. You should not agree to take part unless you are completely happy about all the procedures involved. In the best interests of your health and fitness, it is strongly recommended that you discuss with or inform your personal doctor, physiotherapist, coach and/or other physical trainers of your possible participation in this study, wherever possible.

2) THE NATURE AND PURPOSE OF THIS STUDY

You are invited to take part in a research study. This study is executed by Euodia Vermeulen and forms part of the requirements to obtain a Masters Degree in Engineering at the University of Pretoria. The aim of this study is to use the data that is generated by your fitness tracking device (the running watch that you wear during your runs) to develop a new analysis technique on running form. Here running form refers to how long you run at a certain cadence (number of steps per minute) during a run and the speeds that you achieve during those times.

This analysis technique will provide your with a better understanding of your inherent running style, the running techniques you use and identify areas for physical improvement.

Your participation in this study is completely voluntary. You are at no time obligated to partake or provide any personal information.

3) EXPLANATION OF PROCEDURES TO BE FOLLOWED

The procedure is very basic. You record your runs (training as well as races) onto your fitness tracker or watch and synchronise the data to your online profile and application on your phone as you normally do. This is all that is required from your side. The researcher extracts the data from your online fitness profile and stores it on a database. From there the data is read into a software program for the purpose of analyses and further work.

The researcher will from time to time observe your track training sessions. This is purely for the purpose to have a mental picture of your running style and techniques when doing the analyses on the data.

4) RISK AND DISCOMFORT INVOLVED.

You bear no physical or mental risk during your participation in this project, as there will be no intervention on training methods or race tactics from the researcher's side. You are under no obligation to run, train or race. When you are unable to run (due to injury, illness or whatever other reason) you are advised not to participate in running activities.

You continue with your normal training program, race schedule and race tactics as you and/or your coach sees fit. The researcher merely requires the data from those training sessions and races from your online fitness profile as well as occasional observations of your training sessions in order to better understand the patterns in the data and the information drawn from it. There will be no video recordings or images captured during these observation sessions. Your data will not be shared with any of the other participants.

Your online profile details will be kept safe and the frequency of access will be kept to a minimum.

5) POSSIBLE BENEFITS OF THIS STUDY.

The study aims to develop a new approach in the analysis of the data from you running sessions. This approach will provide you with a holistic view on your running form and how you generate speed during a run. You will be able to better understand your own capabilities and find or consider new ways to improve your running form and fitness levels. The graph that will be developed as part of this technique will provide you with a better visual tool to quickly determine the state of your running form and whether you are improving.

The analysis technique to be developed will enable you to better prevent injuries by pointing out risk factors and bring to light information that is not apparent form the current way in which you see your data on the online fitness profile or application on your phone.

You should know that the developed technique may not be always be 100% accurate. This is because there are factors that are outside of the technique's control such as your mental state, injuries or illnesses developed on training or race days, the weather, possible malfunctioning of the device and other unidentified factors.

Although the researcher is forever grateful to you if you choose to participate, you will not receive compensation of any kind for you participation.

8) HAS THE STUDY RECEIVED ETHICAL APPROVAL?

This Protocol was submitted to both the Faculty of Engineering and Information Technology (and the Faculty of Health Sciences Research Ethics Committee, University of Pretoria, telephone numbers 012 356 3084 / 012 356 3085 and written approval has been granted by those committees. The study has been structured in accordance with the Declaration of Helsinki (last update: October 2013), which deals with the recommendations guiding researchers on how to do research involving human participants. A copy of the Declaration may be obtained from the investigator should you wish to review it.

9) INFORMATION

If you have any questions concerning this study you should contact:

Euodia Vermeulen cell: 083 420 8770 or email euodiav@gmail.com.

10) CONFIDENTIALITY

All records obtained in this study will be regarded as strictly confidential. The raw data and final results will be kept safe in encrypted (password protected) files. Neither the raw data nor the final results will contain any information which may identify you. You have access to all your data as well as the results pertaining to you. The computer which houses the database is protected by an anti-virus which is updated regularly. The computer is also password protected.

It is required by the University that the results be published in a final report in the form of a thesis as well as an article in a journal. The results will be published or presented in such a fashion that you remain unidentifiable.

11) CONSENT TO PARTICIPATE IN THIS STUDY.

Informed consent / assent

11.1 I, ____hereby voluntarily **grant or deny** (encircle which is applicable) my permission to be a participant in the project as explained to me by Euodia Vermeulen.

11.2 I may withdraw from the study at any time and am under no obligation to finish the proposed study.

11.3 The nature, objective, possible safety and health implications have been explained to me and I understand them.

11.4 I understand my right to choose whether to participate in the project is completely voluntary and that the information obtained will be handled and stored confidentially. I am aware that the results of the investigation will be used for the purposes of publication. The information will only be used for research purposes.

11.5 I grant the researcher access to my online fitness profile for the sake of data extraction and validation thereof.

11.6 I give the researcher permission to observe my training sessions from time to time.

11.7 I understand that the analysis of my running form from the developed technique by this study may not be 100% accurate as explained above.

11.8 I understand that I have access to all my data as well as results pertaining to me.

11.9 I understand that if I do not want to participate in this study I am free to continue running as I have been doing without any restrictions or prejudice from the researcher.

I have read or had read to me in a language that I understand the above information before signing this consent form. The content and meaning of this information have been explained to me. I have been given opportunity to ask questions and am satisfied that they have been answered satisfactorily. I understand that if I do not participate it will not alter my running activities in any way. I hereby volunteer to take part in this study.

I have received a signed copy of this informed consent agreement.

....
Athlete name Date

....
Athlete signature Date

....
Investigator's name Date

....
Investigator's signature Date

....
Witness name and signature Date