Long-Distance Running Routes' Flat Equivalent Distances from Race Results and Elevation Profiles

Dimitri de Smet[1], Michel Verleysen[1], Marc Francaux[2] and Laurent Baijot[3]
[1]*ICTEAM, UCLouvain, Louvain-la-Neuve, Belgium*
[2]*IoNS, UCLouvain, Louvain-la-Neuve, Belgium*
[3]*Formyfit, Enghien, Belgium*

Keywords: Equivalent Distance, Endurance, Race Times, Running, Collaborative Filtering, GPS Track, Elevation, Gradient, Ascent.

Abstract: Running routes' elevation profiles affect their *runnability* and therefore athletes' average speeds: route distances alone are not sufficient to predict or evaluate running times. This is an issue for race preparation, race strategy, performance comparison and runner workload planning anytime the ground is sloping. This paper proposes a methodology to establish route equivalent distances expressed as a function of their elevation profiles. The same expression can be used to compute gradient adjusted speeds either for athlete pacing during races or to analyze their performances afterward. The approach is first based on race results and addresses the problem of attendees' level disparities by evaluating races and athletes at the same time using a race performance model. Subsequently, this paper use polynomial and piecewise linear regressions on the instant slope along the routes to express equivalent distances. They match previous studies with constant slopes and extend to the case of varying slopes.

1 INTRODUCTION

Race distances alone are not sufficient to evaluate athlete race times. For instance, an athlete who runs a marathon distance in 3:30' could be a well prepared casual runner if it happened to be at the Berlin's Marathon (known to be particularly flat) or he could be a world class champion if it was the Pikes Peak Marathon (known to be one of world's toughest marathon). Elevation gradient, weather conditions, altitude, vegetation, uneven ground and ground firmness affect athlete speeds. A flat equivalent distance, that reflects all race characteristics is useful in many ways: athletes can prepare races considering realistic race lengths. It also makes athlete ranking possible even if they did not attend to the same races (this also requires a performance model). This paper describes a methodology that assigns flat equivalent distances to routes based, first, on race times and, then, on their elevation profiles. The flat equivalent distance is defined as the distance that would be run, on average, in the same time if the ground was flat; all other above-mentioned conditions being equal. The methodology is applied on endurance races ranging from 8 to 159km.

Two paths could be taken to achieve the same goal: through metabolic measurements or through statistics on race results. In the first approach, flat equivalent distance can be computed as the distance that would lead to the same energy expenditure on flat ground. The relationship between energy expenditure and gradient of ascent is established by (Minetti et al., 2002) by measuring athletes oxygen uptake on an inclined treadmill. This first approach is perfectly fine to assess workload or to plan weight reduction program but might not be accurate in what concerns race times prediction because it does not target it specifically. The second approach infers a relationship between race average speeds and their elevation profile (Kay, 2012; Scarf, 1998; Scarf, 2007). This paper follows the second approach and addresses two problems that previous works do not take into account.

The first problem is that race results cannot be easily compared because races are not run by the same set of athletes. There could be attendee level differences that depend on the popularity of the race. For instance, some local races may fail to attract world elite runners. On the other hand, some very popular races may attract a crowd of casual runners. This implies that races results (race records or any race statistics) do not necessarily reflect its objective *runnabi-*

Smet, D., Verleysen, M., Francaux, M. and Baijot, L.
Long-Distance Running Routes' Flat Equivalent Distances from Race Results and Elevation Profiles.
DOI: 10.5220/0006937000560062
In *Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support (icSPORTS 2018)*, pages 56-62
ISBN: 978-989-758-325-4

lity. This problem is solved by taking a collaborative filtering approach, inspired by (de Smet et al., 2017), that evaluates both athletes and races at the same time. In this way, equivalent distances approximation can be computed from a collection of race results taking the attendees' level into account.

The second problem is that previous attempts fully describe race elevation profiles by two global metrics that are either the *cumulative elevation gain*[1] (Scarf, 2007) or the average elevation gradient of non-loop races (fell running races) that present a relatively constant gradient (Kay, 2012). This is rarely observed in practice. In the present paper, the full elevation profile is considered; this allows for a more realistic relationship extraction.

To establish the desired relationship between flat equivalent distances and route elevation profiles, one need to possess races' equivalent distances for which the elevation profiles are known. For this purpose, races equivalent distances are, first, established using race results. This step is referred to as *collaborative filtering* (Section 3). Then, taking elevation profile data as inputs, a regression model that reproduces the obtained race equivalent distances is built (Section 4). These two steps are validated by assessing how equivalent distances improve race time prediction compared to actual distances.

In the following, all equations are expressed using SI unit system: speeds in [m/s], times in [s] and elevation gradient in [m/m]. Other units are used in figures for convenience.

2 DATA SOURCES

The two steps methodology requires two kinds of data. In the first step (the collaborative filtering part), race results are used to compute flat equivalent distances for a set of races. In the second step (the flat equivalency modelling part), elevation profiles are used to model flat equivalent distances as a function of the instant elevation gradient along the routes.

2.1 Race Results

A set of 228 031 races times was gathered by parsing official results of 616 Belgian races. They represent a large variety of endurance races that took place in 2014 and 2015. From these results a subset of 179 674 race times (445 races, 7480 athletes) is kept to obtain a data set presenting properties that allow a collaborative filtering approach to operate (as explained in Section 3.3). Race results are used to compute flat equivalent distances that are then put in relation with race elevation data.

2.2 Elevation Data

Race routes data were collected through measurements made by runners during the races using their sports watches. Runners uploaded *tracks* and made them publicly available to the online community. Those tracks contain data such as geographic coordinates, timestamps and altitudes. Consumer grade GPS-based elevations have poor accuracy (Bauer, 2013). In a previous work (de Smet et al., 2017) route elevations were gathered by querying publicly available topography data such as SRTM data (Shuttle Radar Topography Mission) or *Google Maps APIs*. They are both based on radar topography survey made from space. It is observed that, in such databases, the altitudes of the treetops is assigned to route parts that are covered by trees: this causes artificial high elevation gradients on routes that pass under trees.

Fortunately, some high-end sports watches include barometric altimeter that have good relative accuracy: the altitude is known with an additive bias that would need to be calibrated. The relative accuracy is what is of primarily interest, the absolute altitude accuracy being irrelevant to our purpose as our analysis is based on elevation gradient only. Routes recorded with such devices could be found only for 129 races of our 445. Those 129 races are used to model our flat equivalent distance model from elevation data.

Although more than two thirds of the races were not used in the flat equivalency modelling part, they are still useful in the collaborative filtering part because they help to characterize athletes and therefore improve the flat equivalent distance estimation of the 129 races that are used in the flat equivalency modelling section.

2.3 Instant Elevation Gradient

Elevation profiles as they are recorded, even by high-end devices, are noisy signals that need to be filtered; especially because our application requires to take the gradient: the derivative of a noisy signal can take artificially high amplitudes. A simple way to take the gradient and smoothing at the same time is to take the average altitude on a n-meters distance ahead minus the average altitude on the same distance behind. The chosen distance acts then as a smoothing factor. More formally, if the elevation profile $e(x)$ is re-sampled every meter, its gradient $g(x)$ at distance x is given

[1] The cumulative elevation gain is the sum of all positive vertical displacement along the route.

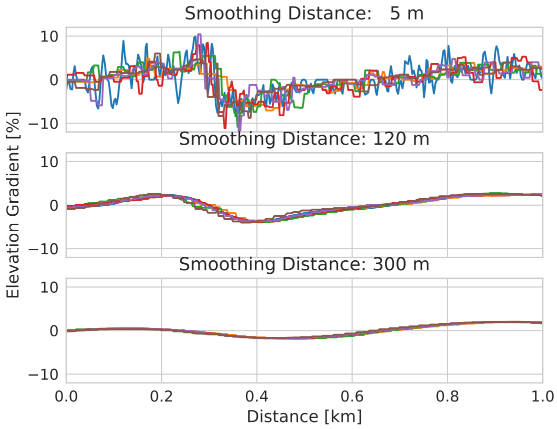

Figure 1: Elevation gradient with different smoothing distances recorded by 6 different sports watches featuring a barometric altimeter. On the first plot, the elevation gradient is under-smoothed: it shows some details that do not correlate among measurements. At the opposite, on the last plot, the elevation gradient is over-smoothed: some details present in the 6 measurements vanish.

by

$$g(x) = \frac{\sum_{i=x+1}^{i=x+n} e(x) - \sum_{i=x-n}^{i=x-1} e(x)}{n^2}, \tag{1}$$

with n being the smoothing distance in meter. Smoothing reduces the measurement noise but it also reduces fast changing details of the real gradient. Therefore, choosing the smoothing distance n results in a trade-off that is solved by analyzing its effect on several track measurements of the same routes recorded by different runner devices. As independence of the measurement can be assumed, if the filtering distance is too small, the random noise is different on each of the measurement making them less correlated. At the opposite, a too large smoothing distance makes disappear some details that are present on all the measurement traces. The smoothing problem is illustrated in Figure 1. Smoothing factor of 120 meters is observed to be a good trade-off: average tracks pairwise correlation reach a plateau and most small scale elevation details remain visible.

3 COLLABORATIVE FILTERING

The idea of using collaborative filtering presented by (de Smet et al., 2017) is to take benefit of a collection of race times of many athletes on many races to learn enough information about every race and every athlete to best explain race results. The basic underlying model is that race times are explained by a linear model of the race paces: the race pace $p_{a,r}$ of athlete a on race r is a sum of products of N race parameters r_i times N athlete parameters a_i:

$$p_{a,r} = \sum_{i=1}^{N} a_i \cdot r_i. \tag{2}$$

Races and athletes parameters are found by minimizing the quadratic reconstruction error, i. e. the sum of the squared differences between observed paces and paces computed with Equation (2).

The simplest case is when $N = 1$ meaning that the race pace is equal to a parameter related to the difficulty of the race times a parameter related to the level of the athlete (low values for high levels). In other words, given all the considered races results, the algorithm outputs race and athlete parameters that allow race time prediction and athlete comparison.

This paper presents a new underlying model that is more physiologically sound and that makes the flat equivalent distance appearing explicitly.

3.1 Race Performances Modelling

Running race times are quite well reflected by a power law (García-Manso et al., 2012): if elevation variations are small enough to not affect the race time, an athlete a is expected to run race r with an average speed s that depends on the distance as

$$log(s) \approx \alpha_1 \cdot log(D_r) + \alpha_0 \tag{3}$$

with D_r being the actual race distance and $\{\alpha_0, \alpha_1\}$ two athlete-specific parameters.

Race time $t = D_r/s$ can be expressed as a function of a fictive flat equivalent distance D_{eq}.

$$log(t) \approx w_1 \cdot log(D_{eq}) + w_0 \tag{4}$$

with $\{w_1, w_0\} = \{1 - \alpha_1, -\alpha_0\}$. Flat equivalent distances are formally defined as the distances that would be run in the same time if the ground was flat. They reflect all parameters that may affect the average race speed.

3.2 Equivalent Distances Estimation

Athlete parameters $(w_{a,0}, w_{a,1})$ and race equivalent distances $D_{eq,r}$ in Equation(4) are unknown but race times must reflect them.

As the number of race results is high, athlete parameters and race equivalent distance can be set so that the average squared deviation to Equation(4) is minimized on observed race results.

Let $log(t_{a,r})$ be the log race time of athlete a on race r and $log(\tilde{t}_{a,r})$ its approximation. Using Equation(4), they can be expressed as

$$log(\tilde{t}_{a,r}) = w_{a,1} \cdot log(D_{eq,r}) + w_{a,0} \approx log(t_{a,r}). \tag{5}$$

The task at hand is now to select race and athlete parameters such that our prediction match observed log race times.

Let Ω be the set of observed (a,r) for which we have race results $t_{a,r}$. Using the least square error criterion, athlete and race parameters stored in vectors **A** and **R** can be selected by solving the optimization problem

$$\underset{\mathbf{A},\mathbf{R}}{\arg\min} \sum_{(a,r)\in\Omega} (log(t_{a,r}) - log(\tilde{t}_{a,r}))^2 \qquad (6)$$

This problem is solved using an *alternate least squares* algorithm that alternates between two least squares problems: optimization of the athlete parameters while holding race parameter constant and optimization of the race parameter while holding athlete parameters constant(Jain et al., 2013). Convergence is observed after several iterations, typically 15 to 20.

Solution to Equation (6) does not necessarily give $D_{eq,r}$ that correspond to races flat equivalent distances: the optimal solution has one degree of freedom because athlete parameters w_1 are allowed to compensate freely race parameters $D_{eq,r}$.

Therefore, an extra constraint is required. Some races present virtually no elevation differences, like for instance, the Berlin Marathon or some races in coastal regions. For races with very flat profile, it can be expressed that their equivalent distance is equal to their actual distance:

$$D_r = D_{eq,r} \qquad (7)$$

For a set Ω_f of selected flat races, the constraint is met, on average, if:

$$\sum_{\Omega_f} D_r = \sum_{\Omega_f} D_{eq,r} \qquad (8)$$

3.3 Data Conditioning

The above mentioned constraint assumes that $a_{a,1}$ and D_{eq_r} compensate each other in the same way for every pair of (a,r). This is not guaranteed unless races have several athletes in common. To ensure that this condition is met. We define a proximity metric between two races as the number of athletes who attend to both of them. By computing this metric for all pair of races, we get an adjacency graph from which community detection algorithm can select clusters of densely connected races. Using the Louvain algorithm presented by (Blondel et al., 2008), only the largest and most densely connected cluster is kept. The later contain 445 races.

In addition, every athlete must have enough race results to be characterized (in other words, to be able to set his w_0 and w_1 parameters) even when one or two race results are kept aside for validation. Therefore, the collaborative filtering part is applied on athletes who have at least 4 race results. This is the case for 7480 of them.

3.4 Validation

The validation of the collaborative filtering part must assess its ability to assign flat equivalent distances to races. Given the proposed definition of the flat equivalent distance, its quality can be evaluated with the accuracy of race time prediction.

For this purpose, 1 percent of the race results are kept aside for validation. All race results but this 1 percent are used to fit the parameters of each athlete in Equation (4) with the equivalent distance provided by the collaborative filtering part. This corresponds to a simple linear regression per athlete. The accuracy is evaluated by computing the error at predicting the 1 percent race times that were not used while solving Equation (6). The process is repeated 100 times to reduce the variance of the error estimate.

4 FLAT EQUIVALENCY MODELLING

Having flat equivalent distance approximations for a set of known races, a regression model can be built to obtain flat equivalent distances based on their elevation profiles.

Instant elevation gradients for each point on the race route are computed from the elevation profile as described in Section 2.3. Let the function $F(g)$ the distance correction that need to be applied to a given sloping distance D presenting an elevation gradient g:

$$D_{eq} = D \cdot F(g). \qquad (9)$$

If the considered distance presents a varying gradient, it can be split in 1-meter sub-section x presenting a gradient $g(x)$. The equivalent distance of the whole route is then the sum of the contributions of each meter :

$$D_{eq,tot} = \sum_{x=0}^{D} F(g_{(x)}), \qquad (10)$$

$F(g)$ can take various forms. The present paper is restricted to piecewise linear and polynomial functions. Section 4.1 presents different models that are found in the literature. Section 4.2 show how model coefficients can be fitted to reproduce flat equivalent distance computed with the technique that is presented in the collaborative filtering part.

4.1 Models

Naismith-like Model

The first rule of thumb, called the Naismith's rule, dates from 1892 and is relayed, among others, by (Scarf, 2007). Naismith's rule formulated in terms of equivalent distances in the sense of our definition can be expressed as

$$\frac{D_{eq}}{D} = F(g) = \begin{cases} 1, & \text{if } (g < 0) \\ 1 + f_N \cdot g, & \text{if } (g > 0) \end{cases} \tag{11}$$

with the f_N constant being evaluated to 7.92 by (Scarf, 2007).

Polynomial Models

Other papers present a 4th or 5th order polynomial expressions that take increased speed for negative gradients into account. (Minetti et al., 2002) gives a relation that express the metabolic energy cost of running by distance unit as a 5th order polynomial of the elevation gradient[2]. This equation will be compared to ours although their definition of equivalent distance in that case would be *the distance that would lead to the same energy expenditure if it was on flat ground.*

(Kay, 2012) gives a 4th order polynomial that can be expressed with a definition of flat equivalent distance that matches ours.

4.2 Model Fitting

As stated earlier model fitting assigns model parameter of the unknown function $F()$ of Equation (10) so that it best reproduce flat equivalent distance that are established solving the optimization problem (6) discussed in the collaborative filtering section.

For the Naismith-like model, fitting Equation (11) requires to set the parameter f_N. This is done by minimizing the least square error of the equation by computing, for each races, cumulative elevation gain.

In previous works, the function F was fitted to route features using global route features either by assuming a constant gradient or by taking the cumulative elevation gain. In our case, races present varying gradient. The general polynomial form can be expressed as

$$\frac{D_{eq}}{D} = F(g) = 1 + \sum_{i=1}^{i=P} f_i \cdot g^i \tag{12}$$

with P, the polynomial order. The independent term is set to 1 because $F(g) = 1$ for $g = 0$: the equivalent distance of a route on flat ground is the distance itself.

The total equivalent distance (10) can then be written as the sum for each meter x along the route as

$$D_{eq,tot} = \sum_{x=0}^{D} [1 + \sum_{i=1}^{i=P} f_i \cdot g^i_{(x)}], \tag{13}$$

which can the be re-arranged as

$$D_{eq,tot} = D + \sum_{i=1}^{i=P} [f_i \sum_{x=0}^{D} g^i_{(x)}], \tag{14}$$

allowing to pre-compute the inner sum for each race route. The P model parameters f_i can then be computed using a simple linear regression.

In the equation above the *runnability* is assumed to only depend on the elevation gradient. This of course not completely true in practice. The underlying assumption is that all other parameters (like weather conditions) are independent of the elevation profile so that they will be averaged out in the regression.

4.3 Validation

Just as for the equivalent distances provided by the collaborative filtering, the quality of the equivalent distance computed using gradient-based formula $F(g)$ is evaluated with the accuracy of race time predictions. Again, for each athlete, parameters $w_{a,0}$ and $w_{a,0}$ in Equation (4) can be set using all race results for which equivalent distance can be computed except 1 percent that is kept aside for validation. The error between validation race results and predicted race results is computed. The process is also repeated 100 times to reduce the variability of the error estimate.

5 RESULTS

This section presents the quality of our flat equivalent distances as predictors for athletes' race times using the Equation (4). As athletes experience high variability in their performances (especially casual runners), race time predictions can not be highly accurate: 4 to 6 percents error are observed on average. Nevertheless, given our definition of equivalent distance, the accuracy of race time prediction is the best way to assess the quality of the computed equivalent distances. The boxplot presented in the Figure 2 shows the relative error at predicting athletes' race times considering different equivalent distances. Models expressed

[2] metabolic energy cost of running was measured by quantity of oxygen uptake

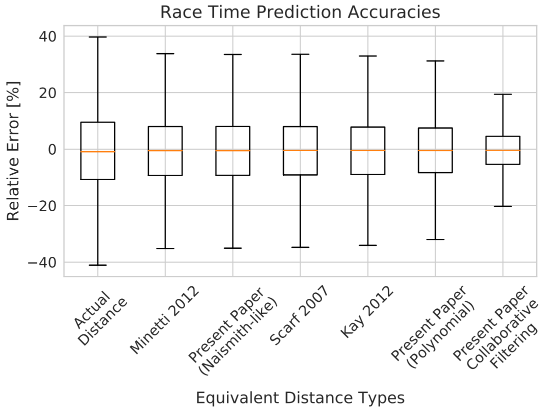

Figure 2: Tukey boxplot representing relative error while predicting race times with different equivalent distances. Computing equivalent distances from race results through collaborative filtering gives the best results.

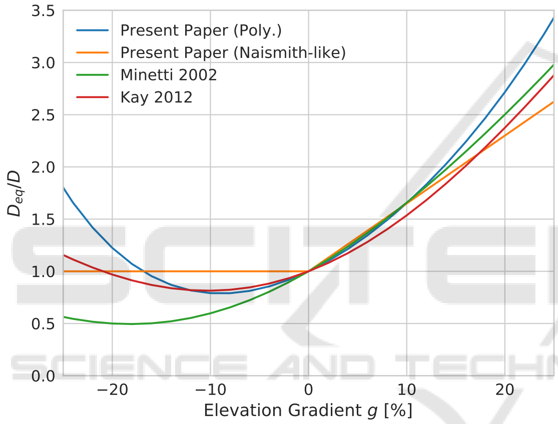

Figure 3: Distance correction based on the elevation gradient.

by Equations (11) and (12) are evaluate; both with literature coefficients and with coefficients fitted on our computed equivalent distances. The Table 1 shows mean relative errors and model parameters.

The Figure 3 shows Naismith's piecewise linear model with original coefficient, two polynomial models with original coefficients and our best polynomial fitting (5th order). We can note their relative similarity in the range that is displayed. As our elevation profiles do not present much gradient outside of the range $[-20\%;20\%]$, this is also the range of validity of our model.

6 DISCUSSION

The methodology presented in this paper shows similar distances correction as previous researches with a novel approach that possess some advantages.

Table 1: Mean relative errors at predicting race times given for different equivalent distances. Actual distances give 7.5%. Collaborative filtering equivalent distances give 4%.

Model	Naismith-like (Scarf, 2007)	5th order polynomial (Minetti et al., 2002)
Equation #	(11)	(12), $P = 5$
Coefficients	$f_N = 7.92$	$\{f_{1..5}\} = \{5.42, 12.9, -12.0, -8.44, 43.2\}$
MRE	6.4 %	6.3 %
Model	Naismith-like (present paper)	4th order polynomial (Kay, 2012)
Equation #	(11)	(12), $P = 4$
Coefficients	$f_N = 6.5$	$\{f_{1..4}\} = \{3.64, 17.8, -3.10, -23.8\}$
MRE	6.3 %	6.1 %
Model		5th polynomial (present paper)
Equation #		(12), $P = 5$
Coefficients		$\{f_{1..5}\} = \{3.90, 220, 23.6, 6.36, -5.34\}$
MRE		5.8 %

We obtained flat equivalent distances by applying a collaborative filtering technique on race results. This technique takes benefit of physiologically sound power law to evaluate races' equivalent distance and athletes' level at the same time. By doing so, we address the problem of athlete level disparity among the different races.

The obtained flat equivalent distances are used to build models that take races elevation profile data as inputs. Unlike previous works, we apply our distance correction model to routes with varying gradient.

Results prove that the computed equivalent distances are more relevant than the actual distances as race time predictors. The best distance correction is the one that is computed on race results because it captures all parameters that affect race times (elevation gradient, weather conditions, ground firmness, etc.). Equivalent distances solely based on the routes' elevation profiles give all similar improvements.

Races considered in this paper are all loop races: they end at the same place as they begin. Therefore, flat equivalency formulas can not be safely generalized to routes for which it is not the case. This is obvious for Naismith-like formulas: in appearance, the model only considers decreased speed in uphill sections; but actually, as it depends on race results, the model coefficient f_N accounts for the fact that there are necessarily downhill sections were the speed is at least a little increased. Indeed, a route with only uphill sections, would be slower than what is predicted by the Naismith-like formulas that are considered here.

The polynomial expression that is obtained for distance adjustment can serve as-is to correct instant speed on a race route. This could be used, for instance, to optimize race management by providing a gradient-adjusted target speed along the route.

REFERENCES

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

de Smet, D., Verleysen, M., and Francaux, M. (2017). Running race times prediction and runner performances comparison using a matrix factorization approach. In *5th International Congress on Sport Sciences Research and Technology Support*.

García-Manso, J., Martín-González, J., Vaamonde, D., and Da Silva-Grigoletto, M. (2012). The limitations of scaling laws in the prediction of performance in endurance events. *Journal of theoretical biology*, 300:324–329.

Jain, P., Netrapalli, P., and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM.

Kay, A. (2012). Pace and critical gradient for hill runners: an analysis of race records. *Journal of Quantitative Analysis in Sports*, 8(4).

Minetti, A. E., Moia, C., Roi, G. S., Susta, D., and Ferretti, G. (2002). Energy cost of walking and running at extreme uphill and downhill slopes. *Journal of applied physiology*, 93(3):1039–1046.

Scarf, P. (1998). An empirical basis for naismith's rule. *Mathematics Today-Bulletin of the Institute of Mathematics and its Applications*, 34(5):149–152.

Scarf, P. (2007). Route choice in mountain navigation, naismith's rule, and the equivalence of distance and climb. *Journal of Sports Sciences*, 25(6):719–726.