Prognozowanie możliwości sportowców w lekkoatletycznych dyscyplinach biegowych - streszczenie pracy

Witold Chodor

Maj 2015

Wiele osób zadaje sobie pytanie jakie są granice możliwości sportowców oraz jak długo będą oni jeszcze w stanie bić kolejne rekordy. Ja również bardzo często stawiałem sobie podobne pytanie. Dlatego też, kiedy natrafiłem na artykuł D.C. Blesta "lower bounds for athletic performance" z 1996 roku postanowiłem zbadać bardziej szczegółowo ten temat w oparciu o obszerniejsze dane.

W mojej pracy starałem się znaleźć predykcje granic możliwości lekkoatletów biegających na 8 dystansach: 100m, 200m, 400m, 800m, 1500m, 5000m, 10000m i 42195m (tzw. maraton). Jako obserwacje przyjąłem rekordowe czasy na wymienionych dystansach w latach olimpijskich, począwszy od 1912 roku, a skończywszy na 2012 roku.

Tabela 1: Rekordy świata w latach olimpijskich

Rok	Rekordowe czasy (s) na dystansie:							
	100m	200m	400m	800m	1500m	5000m	10000m	42195m
1912	10.6	21.3	47.8[c]	111.9[d]	235.8	876.6	1858.8	9634.2
1916[a]	10.6	21.2[b]	47.4[c]	111.9[d]	235.8	876.6	1858.8	9366.6
1920	10.6	21.2[b]	47.4[c]	111.9[d]	234.7	876.6	1858.8	9155.8
1924	10.4	21.2[b]	47.4[c]	111.9[d]	232.6	868.2	1806.2	9155.8
1928	10.4	21.2[b]	47.0	110.6	231.0	868.2	1806.2	8941.8
1932	10.3	21.1	46.2	109.8	229.2	857.0	1806.2	8941.8
1936	10.2	20.7	46.1	109.7	227.8	857.0	1806.2	8802.0
1940[a]	10.2	20.7	46.0	106.6	227.8	848.8	1792.6	8802.0
1944[a]	10.2	20.7	46.0	106.6	223.0	838.2	1775.4	8802.0
1948	10.2	20.7	45.9	106.6	223.0	838.2	1775.4	8739.0
1952	10.2	20.6	45.8	106.6	223.0	838.2	1742.6	8442.2
1956	10.1	20.6	45.2	105.7	220.6	816.8	1722.8	8259.4
1960	10.0	20.5	44.9	105.7	215.6	815.0	1698.8	8116.2
1964	10.0	20.2[b]	44.9	104.3	215.6	815.0	1695.6	7931.2
1968	9.95	19.83	43.86	104.3	213.1	796.4	1659.4	7776.4
1972	9.95	19.83	43.86	104.3	213.1	793.0	1658.4	7713.6
1976	9.95	19.83	43.86	103.5	212.2	793.0	1650.8	7713.6
1980	9.95	19.72	43.86	102.33	211.36	788.4	1642.4	7713.6
1984	9.93	19.72	43.86	101.73	210.77	780.41	1633.81	7685.00
1988	9.92	19.72	43.29	101.73	209.46	778.39	1633.81	7610.00
1992	9.86	19.72	43.29	101.73	208.86	778.39	1628.23	7610.00
1996	9.84	19.32	43.29	101.73	207.37	764.39	1598.08	7610.00
2000	9.79	19.32	43.18	101.11	206.00	759.36	1582.75	7542.00
2004	9.79	19.32	43.18	101.11	206.00	757.35	1580.31	7495.00
2008	9.69	19.30	43.18	101.11	206.00	757.35	1577.53	7439.00
2012	9.58	19.19	43.18	101.01	206.00	757.35	1577.53	7382.00

[a] W 1916, 1940 i 1944 Olimpiady nie zostały rozegrane z powodu I i II wojny światowej.
[b] Biegi odbywały się na nieoficjalnym dystansie 220 jardów (201.17 m).
[c] Biegi odbywały się na nieoficjalnym dystansie 440 jardów (402.34 m).
[d] Biegi odbywały się na nieoficjalnym dystansie 880 jardów (804.68 m).

Modelowanie występów lekkoatletów

Niech i-ty wiersz tabeli 1 oznacza dany rok olimpijski ($i \in \{1, ..., n\}$, n – liczba olimpiad). Z kolei j-ta kolumna (pomijając pierwszą) dotyczy kolejnych dystansów ($j \in \{1, ..., m\}$, m – liczba dystansów).

Wstępna analiza danych nasunęła mi pomysł opisania obserwacji przy pomocy następującego modelu potęgowego dla i-tego roku olimpijskiego

$$t_{ij} = e^{\alpha_i} d_j^{\beta_i} \eta_{ij}, \tag{1}$$

gdzie:

t_{ij} – rekordowy czas uzyskany w i-tym roku olimpijskim podczas biegu na j-tym dystansie,
d_j – j-ty dystans,
α_i, β_i – współczyniki modelu potęgowego dla i-tego roku olimpijskiego,
η_{ij} – składnik losowy o rozkładzie $\eta_{ij} \backsim \ln\mathcal{N}(0, \sigma^2)$, dla ustalonego i η_{ij} są niezależne,

W tym momencie podstawowym zadaniem stało się znalezienie ocen współczynników α_i oraz β_i. Okazało się to dużo łatwiejsze po zlogarytmowaniu stronami równania (1). W efekcie, dla każdego i -tego roku olimpijskiego otrzymałem model liniowy

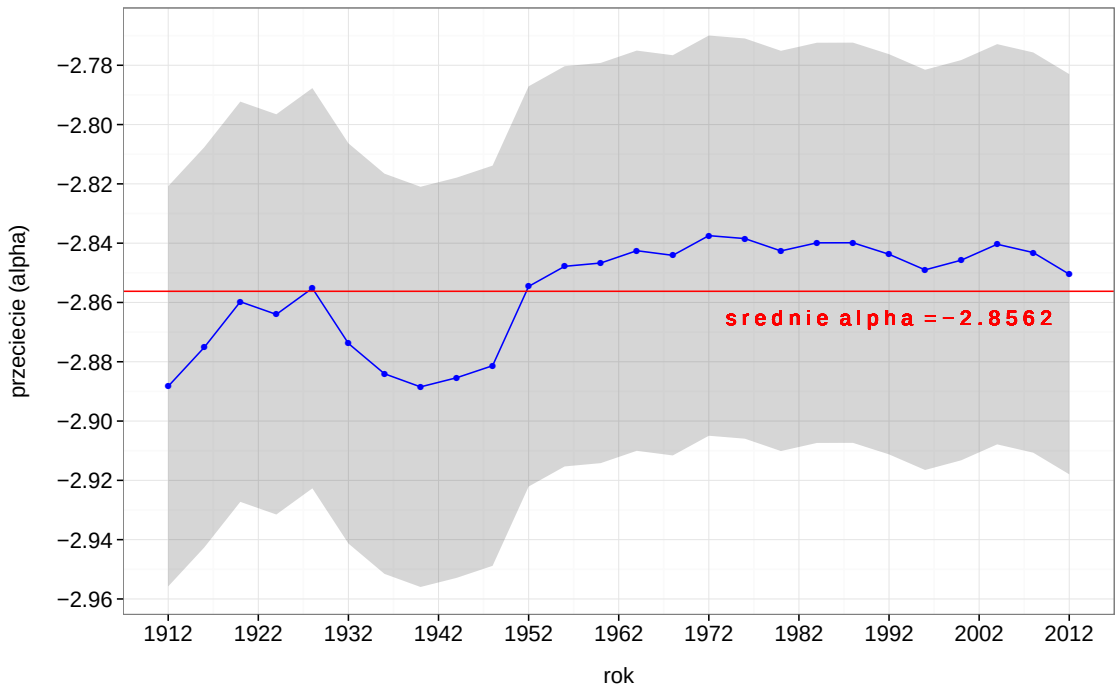
$$T_{ij} = \alpha_i + \beta_i D_j + \xi_{ij}, \tag{2}$$

gdzie:

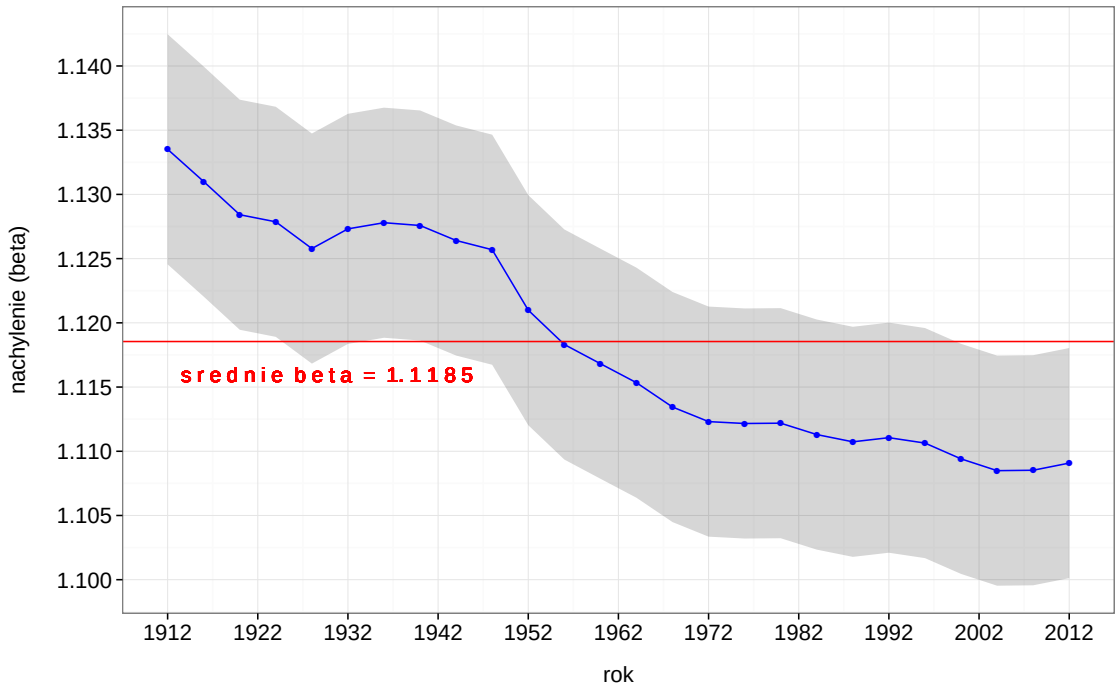
$$T_{ij} := \ln(t_{ij}), D_j := \ln(d_j), \ln(\eta_{ij}) := \xi_{ij}.$$

Przy pomocy oprogramowania statystycznego R udało mi się znaleźć oceny parametrów α oraz β. Na rysunkach (1a) oraz (1b) widzimy jak zmieniają się wartosci ocen tych współczynników w kolejnych latach olimpijskich. Warto zwrócić uwagę na fakt, że średnia wartość oceny parametru przecięcia mieści się całkowicie we wszystkich przedziałach $[\widehat{\alpha_i} - \sigma_{\widehat{\alpha_i}}, \widehat{\alpha_i} + \sigma_{\widehat{\alpha_i}}]$.

(a)



(b)



Rysunek 1: Wykresy przedstawiające oceny parametrów α oraz β dla kolejnych podmodeli regresji tzn. dla kolejnych lat olimpijskich: (1a) zależność pomiędzy rokiem a przecięciem α, (1b) zależność pomiędzy rokiem a nachyleniem β. Czerwona linia oznacza średnią wartość ocen danego parametru. Na szaro zaznaczone zostały przedziały: $[\widehat{\alpha}_i - \sigma_{\widehat{\alpha}_i}, \widehat{\alpha}_i + \sigma_{\widehat{\alpha}_i}]$, $[\widehat{\beta}_i - \sigma_{\widehat{\beta}_i}, \widehat{\beta}_i + \sigma_{\widehat{\beta}_i}]$. Opracowanie własne.

Przypomnę, że zależało mi na znalezieniu wartości granicznych $T_{\infty,j}$ dla każdego z 8 dystansów. Byłoby to łatwiejsze gdybym miał ocenić wartość tylko jednego ze współczynników modelu. Przyjąłem zatem, że dla każdego $i \in \{1, ..., n\}$

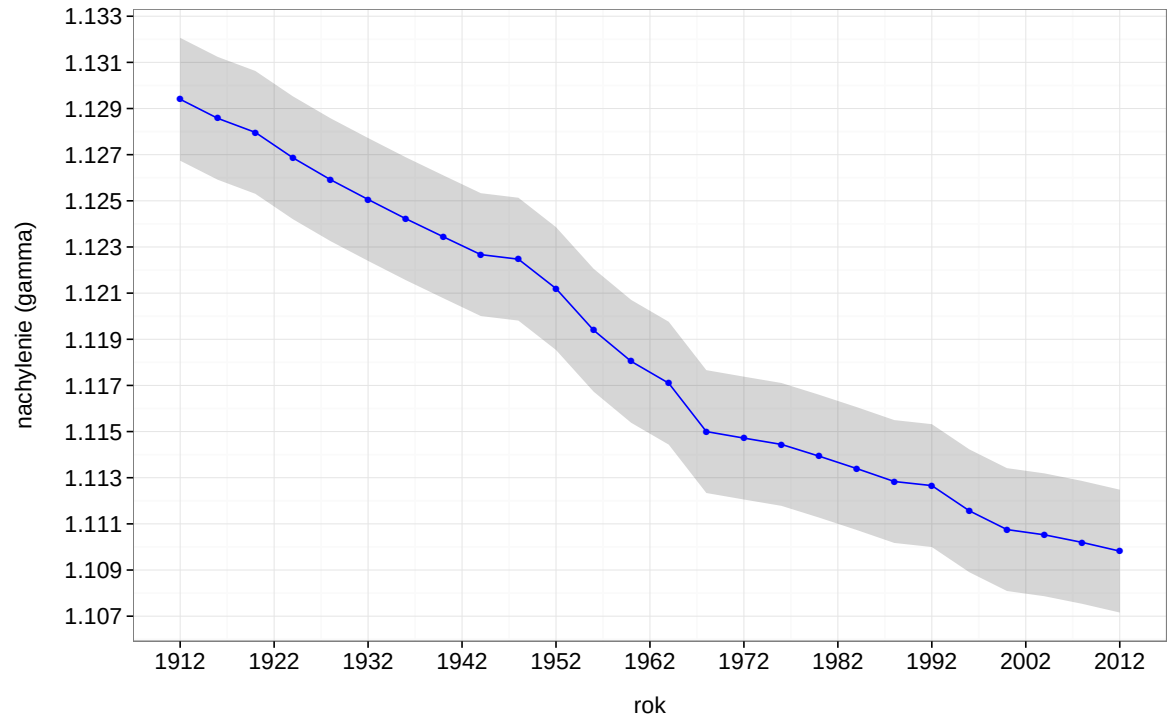
$$\alpha_i := \frac{1}{n}\sum_{i=1}^{n} \widehat{\alpha_i} = -2.8562 = A,$$
$$\beta_i := \gamma_i.$$

Wówczas model (2) przyjął następującą alternatywną postać

$$T_{ij} = A + \gamma_i D_j + \xi_{ij}. \tag{3}$$

Z formalnego punktu widzenia model (3) jest modelem zagnieżdżonym w modelu (2). Test ilorazu wiarogodności dla tych dwóch modeli pozwolił mi stwierdzić, że model alternatywny nie jest istotnie gorszy od modelu (2). Dlatego też w dalszej analizie wykorzystałem model alternatywny.

Podobnie jak wcześniej użyłem programu R do znalezienia ocen parametru γ. Na rysunku 2 widać wyraźnie, że kolejne wartość parametru γ tworzą ciąg ściśle malejący. Oczywiście muszą one również być większe od zera albowiem T_{ij} oraz D_j są dodatnie (nawet od jeden - trudno bowiem spodziewać się, żeby sportowcy zaczęli biegać dłuższe dystanse ze średnią prędkością większą niż na krótkich dystansach). Te dwie informacje pozwalają nam stwierdzić, że musi istnieć granica γ_∞.



Rysunek 2: Wykres przedstawiający oceny parametru γ dla kolejnych podmodeli regresji tzn. dla kolejnych lat olimpijskich. Na szaro zaznaczone zostały przedziały $[\widehat{\gamma_i} - \sigma_{\widehat{\gamma_i}}, \widehat{\gamma_i} + \sigma_{\widehat{\gamma_i}}]$ dla kolejnych ocen parametru γ. Opracowanie własne.

Predykcja granic możliwości lekkoatletów

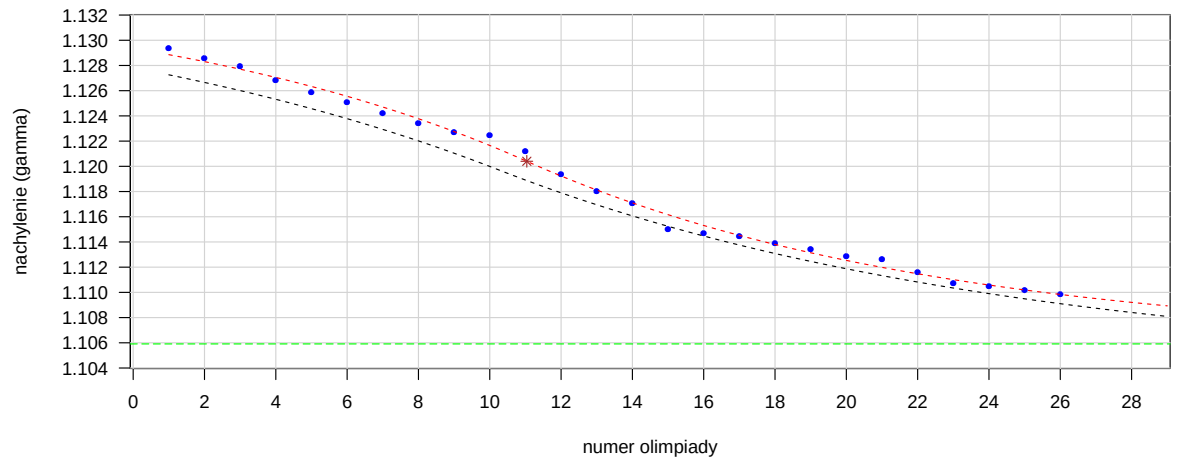
W tej części pracy podstawowym zadaniem było znalezienie krzywej nieliniowej jak najdokładniej opisującej wartości ocen parametru γ. Mając taką krzywą byłem w stanie wyznaczyć wartość graniczną γ_∞. Wprowadziłem następujące oznaczenia:

x_i – zmienna objaśniająca, czyli numer i-tej olimpiady, x_i są niezależne,
y_i – zmienna objaśniana, czyli wartość oceny parametru γ_i dla i-tej olimpiady,
$\boldsymbol{\varphi}$ – wektor nieznanych parametrów $\boldsymbol{\varphi} = (\varphi_1, ..., \varphi_p)$,
f – funkcja nieliniowa ze względu na co najmniej jedej parametr φ_i,
ε_i – składnik losowy o rozkładzie $\varepsilon_i \backsim \mathcal{N}(0, \sigma^2)$, ε_i są niezależne.

Wówczas otrzymałem model nieliniowy nastepującej postaci

$$y_i = f(x_i, \boldsymbol{\varphi}) + \varepsilon_i.$$

Wziąłem pod uwagę 7 różnych modeli nieliniowch, z których, jak się okazało, najdokładniej obserwacje γ_i opisuje model wykładniczy antysymetryczny (rysunek 3).



Rysunek 3: Oceny parametru γ dla kolejnych olimpiad wraz z krzywymi modelu wykładniczego antysymetrycznego
$y = \varphi_4 + \varphi_1 \exp[-\varphi_2(x - \varphi_3)]$ dla $x \geqslant \varphi_3$,
$y = \varphi_4 + \varphi_1\{2 - \exp[\varphi_2(x - \varphi_3)]\}$ dla $x < \varphi_3$. Linia czarna przerywana oznacza model z parametrami początkowymi $\varphi_1 = 0.016$, $\varphi_2 = 0.069$, $\varphi_3 = 10.267$, $\varphi_4 = 1.1037$. Linia czerwona przerywana to model najlepiej dopasowany do danych. Linią zieloną przerywaną oznaczona została γ_∞. Brązowa gwiazdka oznacza punkt przegięcia optymalnej krzywej. Opracowanie własne.

Gdybym w tym momencie zakończył analizę to okazałoby się, że przewidywane granice możliwości sportowców na niektórych dystansach są większe niż rekordy z 2012 roku, co oczywiście jest bez sensu. Dlatego też warto przyjrzeć się resztom $\widehat{r}_{i,j} = T_{i,j} - \widehat{T}_{i,j}$ w modelu (3). Analiza rysunku 4 pokazuje, że dla dystansów: 200m, 400m, 10000m oraz maratonu reszty są ujemne przez większość olimpiad, co uzasadnia dlaczego $T_{i,j} < \widehat{T}_{i,j}$ dla j=2, 3, 7, 8. Przyjąłem zatem, że

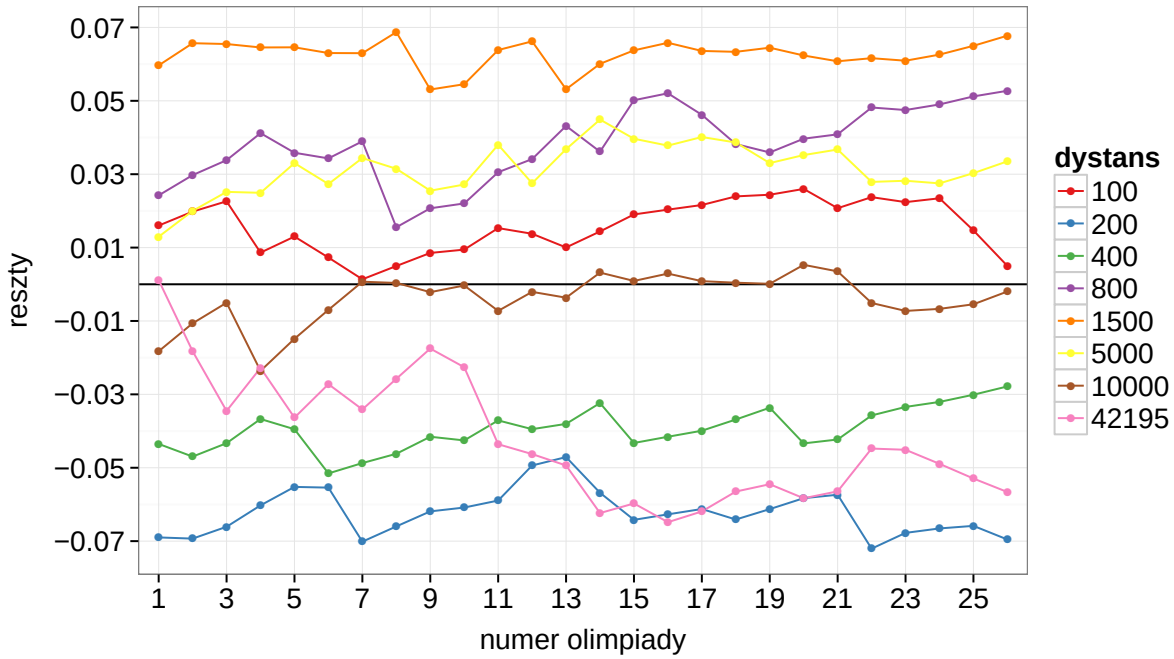
$$\widetilde{\gamma}_\infty = \widehat{\gamma}_\infty - (\max_{i\in\{1,\ldots,26\}} |\widehat{r}_{i,j}|)/D_j, \tag{4}$$

dzięki czemu otrzymamy następujący wzór na obliczenie czasów granicznych

$$\widehat{t}_{\infty,j} = \exp(\widehat{T}_{\infty,j}),$$

gdzie

$$\widehat{T}_{\infty,j} = \begin{cases} A + \widetilde{\gamma}_\infty D_j = A + \widehat{\gamma}_\infty D_j \ - \max\limits_{i\in\{1,\ldots,26\}} |\widehat{r}_{i,j}| & \text{dla } j = 2,3,7,8 \\ A + \widehat{\gamma}_\infty D_j & \text{dla } j = 1,4,5,6 \end{cases}.$$



Rysunek 4: Wartości reszt w zagnieżdżonym modelu liniowym (3) dla poszczególnych dystansów w kolejnych latach olimpijskich. Opracowanie własne.

Po uwzględnieniu korekty (równanie (4)) możemy przedstawić granice możliwośći lekkoatletów na badanych 8 dystansach. Diagnostyka, najlepiej (według malejącej wartości RSS) opisującego oceny parametru γ, modelu wykładniczego antysymetrycznego wykazała, że nie spełnia on wszystkich wymaganych założeń. Dlatego też uznałem, że warto pokazać pewien przedział granic możliwości lekoatletów. Stąd też w tabeli 2 umieściłem również predykcje uzyskane przy pomocy dwóch innych modeli, zajmujących odpowiednio drugie i trzecie miejsce (według malejącej wartości RSS).

Dystans (m)	Aktualny rekord (s)	Przewidywane czasy (s) dla modelu:			
		Antysymetryczny wykładniczy (2012)	Granice możliwości		
			Antysymetryczny wykładniczy	4-parametrowy Gompertza	Logistyczny
100	9.58	9.53	9.36	9.42	9.47
200	19.19	20.57	18.75	18.88	19.01
400	43.18	44.4	41.19	41.5	41.84
800	101.01	95.83	93.35	94.14	94.98
1500	206	192.53	187.08	188.81	190.65
5000	757.35	732.49	708.41	716.04	724.19
10000	1577.53	1580.88	1489.19	1506.54	1525.09
42195	7382	7813.32	7023.04	7117.72	7219.15

Tabela 2: Przewidywane granice możliwości lekkoatletów. W kolejnych kolumnach dane są: badany dystans; rekordy świata z 2012 roku na wybranych dystansach; przewidywane czasy na 2012 rok z gammą uzyskaną przy pomocy modelu antysymetrycznego wykładniczego; przewidywane granice możliwości lekkoatletów ze skorygowaną gammą uzyskaną kolejno przy pomocy modeli: antysymetrycznego wykładniczego, 4-parametrowego Gompertza i Logistycznego.