Measuring Marathon Courses: An Application of Statistical Calibration Theory

BY

RICHARD L. SMITH and MARK CORBETT

Reprinted from

THE JOURNAL OF THE ROYAL STATISTICAL SOCIETY

SERIES C (APPLIED STATISTICS)

Volume 36, No. 3, 1987

(pp. 283–295)

PRINTED FOR PRIVATE CIRCULATION

1987

Appl. Statist. (1987)
36, *No.* 3, *pp.* 283–295

Measuring Marathon Courses: An Application of Statistical Calibration Theory

By RICHARD L. SMITH†

University of Surrey, UK

and MARK CORBETT

Imperial College, London, UK

[Received October 1985. Revised December 1986]

SUMMARY

Marathon and other road-running courses are frequently measured by means of a calibrated bicycle wheel. The process involves an initial calibration over a standard distance, followed by the course measurement itself. We would like to know how accurate this procedure is. We develop both maximum-likelihood and Bayesian methods of analysis, and apply them to data collected for the 1984 Olympic Games. An innovation is the use of dynamic models to cope with changes in the calibration constants.

Keywords: Calibration; Dynamic models; Marathon course measurement

1. Description of the Problem

The official length of a marathon course is 42, 195 metres (26 miles, 385 yards). With the increased popularity of road running as a sport, greater attention is being given to the accurate measurement of courses. For the tens of thousands running the London or New York marathon, a few metres either way are unlikely to make much difference, though most would probably feel cheated if the course measurement was significantly in error (by 500 metres say). For those chasing major championships and "world records", with their increasingly substantial financial rewards, the importance of accurate measurement is accentuated.

The International Amateur Athletics Federation (IAAF) does not recognise "world records" for the marathon, though its Handbook has recently started to list unofficial "world best performances". Present IAAF rules on course measurement are not very precise, and are not rigidly enforced, and this has led to controversy over some recent "world records" including those at New York 1981 and Rotterdam 1985. We therefore believe that the time is right for a careful statistical assessment of the whole question of course measurement.

The most widely used, and officially recommended, method of measurement is the *bicycle method*, which we now describe. An ordinary bicycle is fitted with a *revolution counter*, which records accurately the number of revolutions of the front wheel. The bicycle is first ridden over a *standard distance*, at least one kilometre of usually straight and flat road, in order to calibrate the counter. The standard distance is measured by surveyor's steel tape or electronic distance meter, and may be assumed exact for statistical purposes. The bicycle is then ridden over the route to be measured, care being taken to follow the shortest legal route for the

† *Address for correspondence*: Dept of Mathematics, University of Surrey, Guildford, Surrey GU2 5XH.

© 1987 Royal Statistical Society 0035–9254/87/36283 $2.00

runners. The counter is checked at the beginning and end of the measurement, and the distance computed by simple arithmetic. If possible, the bicycle is again ridden over the standard distance, so as to check that there has been no significant change in the calibration.

This method was introduced in England by the Road Runners Club around 1960; a reference is Jewell (1961). It has gradually displaced other methods. It is used for all events run on the roads. The principal sources of error are considered to be variation in the calibration conditions due, for example, to temperature variation, and error by the measurer in failing to follow the shortest possible course. (The latter is the main reason why crude methods of measurement, for example using a car milometer, are considered totally unacceptable.) It is widely considered that the method, when correctly applied by an experienced measurer, is accurate to about 1 part in 1000. This implies a tolerance of approximately 50 metres in a marathon measurement, and this figure has been adopted (imprecisely) by the IAAF as a maximum allowable error.

The starting point for our own work is a very detailed report (Brennand *et al.*, 1984) on the course measurement of the 1984 Olympic Marathon in Los Angeles. In contrast with the usual procedure, these authors used 13 measurers each of whom took a total of 12 calibration measurements, spread over eight baselines (standard distances) as follows:

Baseline	*Length* (km)	*Location*
B0	1.000178	near the start
B1	0.955978	start + 6 km
B2	0.379007	start + 10 km
B3	0.601258	start + 14 km
B4	0.768575	start + 20 km
B5	0.974693	start + 26 km
B6	1.000030	start + 34 km
B7	1.000000	near the finish

In addition to the baselines, the main part of the course was divided into 13 intervals, making 25 intervals including the baseline. The data, consisting of counter readings for each interval and each cyclist, are presented in Table 1. This table excludes a few short sections which were not measured by all the riders.

We do not have space to discuss the numerous data plots, error analyses, etc., included in the report, but we do reproduce one plot of particular interest. Fig. 1 shows the estimated number of counts per km, averaged over all riders, for each of the eight baselines. The horizontal axis measures (approximately) the time of day. It can be seen that there is a significant variation, of the order of 1 part in 1000, between the beginning and end of the day. Brennand *et al.* suggested that this was due to temperature variation and tried fitting a linear or cosine regression curve (amongst others) through the points. Our own impression is that there are two distinct groups of baselines, a "morning set" B0–B3 and an "afternoon set" B4–B7. This also corresponds to the fact that B0–B3 were measured in heavy rain, the rest in dry and warm conditions. Therefore, it appears to us more reasonable to split the entire data set into two halves (morning and afternoon), rather than to try to model calibration variation by a smooth curve. We return to this point in Section 5.

The main purpose of our analysis is to give an estimate of the overall error in the method. Such an estimate is possible because of the large number of baselines and the use of 13 independent cyclists. This involves us in the statistical theory of calibration, which has been greatly expanded in recent years. We give a brief review of this theory in Section 2. A model is proposed in Section 3, followed by a derivation of the maximum likelihood solution. For various reasons which we outline as we go along, we prefer a Bayesian approach and this is outlined in Section 4. The main data analysis then given in Sections 5 and 6, followed by a summary of our conclusions.

TABLE 1 (Part 1)

Readings of bicycle counter for cyclists 1–7

	1	2	3	4	5	6	7
B0	9402.5	9499.0	9639.5	9601.0	9601.0	9458.0	9348.5
B0	9404.0	9499.5	9645.0	9599.0	9598.0	9458.0	9349.5
B0	9402.0	9499.0	9642.0	9599.0	9599.0	9458.0	9348.0
B0	9403.0	9498.0	9645.5	9600.0	9601.0	9458.0	9352.0
1	12163.5	12287.0	12466.5	12419.0	12409.0	12239.0	12090.0
2	14981.5	15132.0	15363.5	15298.0	15297.0	15070.0	14896.0
B1	8985.0	9077.0	9214.0	9173.0	9172.0	9036.0	8934.0
3	33584.5	33915.0	34447.0	34301.0	34278.0	33770.0	33396.0
B2	3563.5	3600.0	3654.0	3638.0	3633.0	3585.0	3541.0
4	39787.0	40208.0	40811.5	40624.5	40624.0	40025.0	39558.0
B3	5651.0	5711.0	5792.5	5768.5	5770.0	5686.0	5617.0
5	18017.5	18211.0	18473.0	18392.0	18395.0	18128.0	17915.0
6	23949.0	24225.0	24572.5	24464.0	24458.0	24115.0	23821.0
B4	7215.5	7296.0	7400.0	7368.0	7366.0	7259.0	7172.0
7	40070.5	40530.0	41112.0	40940.0	40916.0	40335.0	39851.0
8	19091.0	19309.0	19585.5	19504.0	19497.0	19216.0	18984.0
B5	9146.0	9253.0	9382.5	9340.0	9337.0	9203.0	9095.0
9	26090.0	26383.0	26762.0	26654.0	26640.0	26257.0	25949.0
10	49814.0	50375.0	51089.5	50874.0	50848.5	50130.0	49527.0
B6	9386.0	9492.0	9626.0	9590.0	9586.5	9449.0	9330.0
11	5736.0	5800.0	5886.0	5859.0	5859.0	5772.0	5703.0
12	5404.0	5466.0	5543.5	5519.0	5514.0	5438.0	5374.0
13	1585.5	1602.0	1624.0	1615.0	1621.0	1591.0	1578.0
B7	9385.0	9494.0	9628.0	9584.0	9582.0	9445.0	9336.0
B7	9385.5	9493.0	9628.0	9585.0	9582.0	9447.0	9335.0

TABLE 1 (Part 2)

Readings of bicycle counter for cyclists 8–13

	8	9	10	11	12	13
B0	9362.0	9398.0	9427.0	9343.0	9293.0	9680.0
B0	9364.0	9401.0	9430.0	9347.0	9295.0	9681.0
B0	9364.0	9400.0	9431.0	9343.0	9293.0	9679.0
B0	9366.0	9402.0	9431.0	9347.0	9295.0	9681.0
1	12125.0	12163.0	12196.0	12091.0	12023.0	12525.5
2	14919.0	15007.0	15032.0	14893.0	14805.0	15420.5
B1	8952.0	8985.5	9014.0	8935.0	8882.0	9248.0
3	33458.0	33582.5	33688.5	33385.0	33208.0	34558.0
B2	3551.0	3562.0	3571.5	3538.0	3523.0	3666.0
4	39642.0	39792.5	39912.5	39529.0	39341.0	40948.0
B3	5631.0	5649.5	5668.5	5616.0	5587.0	5816.0
5	17956.0	18010.5	18075.0	17908.0	17814.0	18543.0
6	23870.0	23955.0	24031.0	23809.0	23682.0	24652.0
B4	7191.0	7215.0	7236.0	7169.0	7133.0	7426.0
7	39934.0	40102.5	40206.0	39830.0	39622.0	41256.0
8	19025.0	19088.5	19151.0	18976.0	18877.0	19655.0
B5	9144.0	9141.0	9173.0	9088.0	9045.0	9413.0
9	25999.0	26084.0	26165.0	25925.0	25795.0	26856.0
10	49626.0	49795.5	49963.0	49485.0	49248.0	51249.5
B6	9353.0	9384.5	9415.0	9327.0	9278.0	9669.5
11	5716.0	5733.0	5748.0	5697.0	5669.0	5904.0
12	5386.0	5402.0	5419.0	5369.0	5344.0	5562.0
13	1575.0	1576.0	1588.5	1568.0	1566.0	1629.5
B7	9353.0	9379.5	9410.5	9327.5	9280.0	9662.5
B7	9352.0	9375.5	9411.0	9321.5	9278.0	9660.0

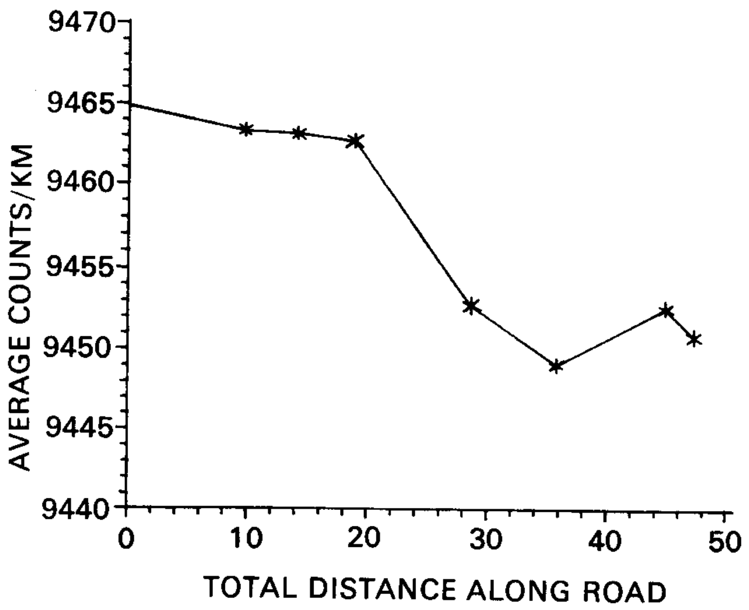

Fig. 1. Variation of calibration constants with time

2. Some Remarks on Statistical Calibration Theory

The classical linear calibration problem is specified by the equation

$$y = \alpha + \beta x + \varepsilon \tag{2.1}$$

between a quantity of interest x and a measurement y. Here α and β are unknown constants and ε is a random error. We are given a calibration sample (x_i, y_i), $1 \leqslant i \leqslant n$, with x_i precisely measured and usually controlled, and then wish to make inference about a future value, $x = x'$ say, based on observed $y = y'$. The classical estimator is $\hat{x}' = (y' - \hat{\alpha})/\hat{\beta}$ where $\hat{\alpha}$, $\hat{\beta}$ are derived from the regression of y on x. Under normality assumptions on ε, this is also the maximum likelihood estimator. Confidence intervals may be derived by Fieller's (1954) method. The use of $\hat{x}'$ was criticised by Krutchkoff (1967) on the grounds that it has infinite variance, and it is known that Fieller's method may lead to a confidence set for x' consisting of the whole real line or even two disjoint semi-infinite lines. Both phenomena arise from the possibility $\beta \approx 0$ and, despite Williams' (1969) rebuttal of Krutchkoff, one is left with the feeling that some form of conditional inference is required. This leads us naturally to a Bayesian formulation of the problem, since in Bayesian statistics problems of conditioning are taken care of automatically. Bayesian solutions were obtained by Hoadley (1970) and Hunter and Lamboy (1981). Brown (1982) considered a multivariate generalisation of (2.1), and generalised both Fieller's technique of interval estimation and Hoadley's Bayesian analysis. It is evident that Brown prefers the Bayesian approch and forceful support was given to this by Lindley, in discussion of Brown's paper. Lindley emphasised that a Bayesian solution of calibration problems required only the laws of probability.

Our own approach has been strongly influenced by the Bayesian arguments of Brown and Lindley, and we develop this in section 4. For comparison, however, it is useful to develop also a maximum likelihood solution, which we do first.

3. Maximum Likelihood Solution

Consider now the road measurement problem, in which x is the true length of the road being measured and y is the number of counts given by the bicycle counter. We may take $\alpha = 0$ in (2.1) but there are 13 different values of β, corresponding to the 13 cyclists. Note that taking $\alpha = 0$ corresponds to assuming there is no "start-up" effect. We could treat this as an instance of Brown's multivariate model, but have found it simpler to start from scratch. We consider the model

$$y_{ij1} = \beta_j z_i + \varepsilon_{ij1},\ \varepsilon_{ij1} \sim N(0, s_{ij1}^{-1}\sigma^2), \quad (1 \leqslant i \leqslant p,\ 1 \leqslant j \leqslant n) \tag{3.1}$$

$$y_{kj2} = \beta_j x_k + \varepsilon_{kj2},\ \varepsilon_{kj2} \sim N(0, s_{kj2}^{-1}\sigma^2), \quad (1 \leqslant k \leqslant q,\ 1 \leqslant j \leqslant n)$$

where y_{ij1}, y_{kj2} are the counts obtained by measurer j on the ith calibration interval, the kth course interval, respectively. We let n denote the number of measurers, p the number of calibraton intervals, q the number of course intervals. The true (known) length of the ith calibration interval is denoted z_i, and the unknown length of the kth course interval is x_k. The ε's are assumed independent with σ^2 unknown, and the values of s_{ij1}, s_{kj2} are initially taken to be known constants of proportionality. A plausible model for these is

$$s_{ij1} = z_i^{-m}, \quad s_{kj2} = \gamma x_k^{-m} \tag{3.2}$$

where $m = 0$ would imply that the variance is independent of distance measured and $m = 1$, 2 would imply that, respectively, the variance and the standard deviation are proportional to distance. Considerations based on the independence of disjoint sections suggest that $m = 1$ is the right value, but we have also tried $m = 0$, $m = 2$ for comparison. The constant γ in (3.2) is intended to reflect the possibility that it may be harder to ride the bicycle accurately on the course then on the calibration intervals. If this were the case, we would have $0 < \gamma < 1$.

We may now develop the maximum likelihood solution, treating $\mathbf{x} = (x_1, \dots, x_q)$ as a vector of parameters of interest and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)$ and σ^2 as nuisance parameters. The likelihood function is given by

$$\left(\prod_{i,\,j} s_{ij1}^{\frac{1}{2}}\right)\left(\prod_{k,\,j} s_{kj2}^{\frac{1}{2}}\right)\sigma^{-n(p+q)} \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{ij}\sum s_{ij1}(y_{ij1} - \beta_j z_i)^2 + \sum_{kj}\sum s_{kj2}(y_{kj2} - \beta_j x_k)^2\right\}\right]. \tag{3.3}$$

Conditionally on $\mathbf{x}$, the maximum likelihood estimates of $\boldsymbol{\beta}$ and σ^2 are given by

$$\hat{\beta}_j = \frac{\sum_i s_{ij1} z_i y_{ij1} + \sum_k s_{kj2} x_k y_{kj2}}{\sum_i s_{ij1} z_i^2 + \sum_k s_{kj2} x_k^2}, \tag{3.4}$$

$$\hat{\sigma}^2 = \frac{R}{n(p+q)},$$

where

$$R = \sum_{ij}\sum s_{ij1}(y_{ij1} - \hat{\beta}_j z_i)^2 + \sum_{kj}\sum s_{kj2}(y_{kj2} - \hat{\beta}_j x_k)^2 \tag{3.5}$$

leading to the profile likelihood

$$l(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) \propto \left(\prod_{i,\,j} s_{ij1}^{\frac{1}{2}}\right)\left(\prod_{k,\,j} s_{kj2}^{\frac{1}{2}}\right) R^{-n(p+q)/2}. \tag{3.6}$$

We have retained constants of proportionality depending on the s_{ij1}, s_{kj2} in order to allow comparison of different models.

Maximum likelihood estimation proceeds by numerical optimisation. Under (3.2), we may also treat γ as an unknown parameter and maximise (3.6) with respect to both $\mathbf{x}$ and γ. Apart from the objections raised in Section 2, this procedure involves a large number of nuisance parameters and therefore raises the queston of whether the asymptotic theory of maximum likelihood estimation may be applied to this case.

4. Bayesian Solution

The Bayesian method is to compute a posterior distribution $p(\mathbf{x} \mid \mathbf{y}, \mathbf{z})$ under the assumption of a prior distribution for $(\mathbf{x}, \boldsymbol{\beta}, \sigma^2)$. We shall simplify things slightly by assuming that the prior distributions of $\mathbf{x}$ and of $(\boldsymbol{\beta}, \sigma^2)$ are independent, and by treating the calibration intervals $\mathbf{z}$ as fixed rather than random. The posterior distribution is then given by the formula

$$p(\mathbf{x} \mid \mathbf{y}) = p(\mathbf{y} \mid \mathbf{x})\pi(\mathbf{x}) / \int p(\mathbf{y} \mid \mathbf{x})\pi(\mathbf{x}) d\mathbf{x} \tag{4.1}$$

where $\pi(\mathbf{x})$ is the prior density of $\mathbf{x}$ and

$$p(\mathbf{y} \mid \mathbf{x}) = \int p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}, \sigma^2) d(\boldsymbol{\beta}, \sigma^2) \tag{4.2}$$

plays the role of a likelihood function for $\mathbf{x}$.

We vary from Hoadley (1970) and Brown (1982) by permitting a full conjugate family for the prior density $\pi(\boldsymbol{\beta}, \sigma^2)$. Suppose that vt^2/σ^2 has a χ^2_v distribution and, conditionally on σ^2, $\beta_1, \ldots, \beta_n$ are independently normal with β_j having mean b_j, variance $a_j^{-1}\sigma^2$. Thus

$$\pi(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-v-n-2} \exp\left[-\frac{1}{2\sigma^2} \left\{ vt^2 + \sum_{j=1}^{n} \mathrm{a}_j(\beta_j - b_j)^2 \right\} \right]. \tag{4.3}$$

Multiplying (4.3) by (3.3) and integrating out first $\boldsymbol{\beta}$ and then σ^2, we obtain

$$p(\mathbf{y} \mid \mathbf{x}) \propto \left(\prod_{i,\, j} s_{ij1}^{\frac{1}{2}} \right) \left(\prod_{k,\, j} s_{kj2}^{\frac{1}{2}} \right) R^{-v/2 - n(p+q)/2}. \tag{4.4}$$
$$\prod_j \left(a_j + \sum_i s_{ij1} z_i^2 + \sum_k s_{kj2} x_k^2 \right)^{-\frac{1}{2}},$$

where

$$R = vt^2 + \sum_j a_j(\tilde{\beta}_j - b_j)^2 + \sum_{ij}\sum s_{ij1}(y_{ij1} - \tilde{\beta}_j z_i)^2 + \sum_{kj}\sum s_{kj2}(y_{kj2} - \tilde{\beta}_j x_k)^2 \tag{4.5}$$

and

$$\tilde{\beta}_j = \frac{a_j b_j + \sum_i s_{ij1} z_i y_{ij1} + \sum_k s_{kj2} x_k y_{kj2}}{a_j + \sum_i s_{ij1} z_i^2 + \sum_k s_{kj2} x_k^2}. \tag{4.6}$$

A limiting case of this analysis is the improper prior obtained by setting $v = 0$, $a_j = 0$. In that case (4.6) agrees with (3.4), (4.5) with (3.5), and the only difference between (4.4) and (3.6) is the final factor in (4.4). This factor arises because we have integrated out the nuisance parameters, instead of maximising them as in (3.6).

In this case, therefore, it appears that the difference between the profile likelihood and posterior density (under the improper prior) is not very great. Professor P. J. Brown has pointed out to us that the profile likelihood (i.e. (4.4) without the final factor) fails to tend to zero as $x_k \to \pm\infty$, so it cannot strictly be renormalised and treated as a posteror density. Brown and Sundberg (1987) have studied the theoretical properties of profile likelihood and Bayesian solutions in the general multivariate calibration model.

In the data analysis which follows, we shall assume (3.2) with m fixed. It is then convenient to treat γ as an unknown parameter, and we may write $p(\mathbf{y} \mid \mathbf{x}, \gamma)$ for the expression in (4.4). Equation (4.1) then becomes

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{\int p(\mathbf{y} \mid \mathbf{x}, \gamma)\pi(\mathbf{x}, \gamma)d\gamma}{\iint p(\mathbf{y} \mid \mathbf{x}, \gamma)\pi(\mathbf{x}, \gamma)d\gamma d\mathbf{x}} \tag{4.7}$$

where we have assumed $(\mathbf{x}, \gamma)$ to be *a priori* independent of $(\boldsymbol{\beta}, \sigma^2)$.

We shall not, however, attempt to implement (4.1) or (4.7) directly. Instead, assuming the prior density in $\mathbf{x}$, π, to be flat in the region of interest, we shall approximate $p(\mathbf{x} \mid \mathbf{y})$ (in the case γ known) or $p(\mathbf{x}, \gamma \mid \mathbf{y})$ (in the case γ unknown) by a multivariate normal distribution whose mean is the point which maximises p and whose covariance matrix is the inverse of the hessian of $-\log p$. This is a well-known approximation which is operationally identical to treating (4.4) as a likelihood function and applying maximum likelihood analysis. The justification, however, lies in the quadratic approximation to $-\log p$, and this may be checked by plots of the likelihood surface.

In the case γ unknown, this means that we maximise (4.4) with respect to $x_1, \ldots, x_q$ and γ to obtain $\hat{x}_1, \ldots, \hat{x}_q$ and $\hat{\gamma}$ say. The inverse of the hessian matrix of $-\log p$ is an approximation to the posterior covariance matrix, and the square roots of the diagonal elements of this matrix are approximate standard deviations of the posterior distribution. With a slight abuse of terminology, we shall refer to these as "standard errors". Finally, an estimate of the total distance

$$x = \sum_{k=1}^{q} x_k$$

is

$$\hat{x} = \sum_{k=1}^{q} \hat{x}_k$$

with "standard error" equal to the square root of

$$\sum_{k=1}^{q} \sum_{l=1}^{q} \operatorname{cov}(\hat{x}_k, \hat{x}_l).$$

These procedures were implemented in Fortran, using the NAG routine E04CGF for derivative-free quasi-Newton optimisation. Starting values for x_k are easily obtained, and we took 1 as the starting value of γ. Convergence was surprisingly rapid considering the dimensionality of the problem; typically about one second CPU on Imperial College's Cyber 855. Numerical derivatives (for the standard errors) were obtained using the NAG routine D04AAF, adapted to handle multivariate functions.

5. Results

The first model we tried was the one defined by (3.1) and (3.2) with $m = 0$, γ treated as unknown. Taking the improper prior $v = 0$, $a_i \equiv 0$, we obtain an estimate $\hat{x} = 30{,}904.1$ metres with standard error (as defined above) of 2.0 metres. We also found $\hat{\gamma} = 0.98$, standard error 0.15. The estimate $\hat{x}$ represents the total length of these 13 sections of the course—the remainder of the marathon course included the calibration intervals, a lap of the track at the end and a few short sections not included in the present discusson.

The estimate depends on a number of model assumptions and we now consider the effect of each of these.

(a) *effect of prior distribution on* $(\boldsymbol{\beta}, \sigma^2)$

The estimate just quoted is based on an improper prior but, since there is a fair amount of past experience with the method, it is also quite practicable to produce a proper prior of the form (4.3). A number of different values for a_i, b_i, v and t were tried; the total variation in $\hat{x}$ was 30,903.9 m. to 30,904.25 m. We conclude that the prior distribution has negligible effect on the estimates.

(b) *Approximation to posterior distribution*

Recall that we are using a normal approximation to the posterior distribution instead of carrying out the integration (4.7). Some indication of the appropriateness of this procedure may be gained from plots of the function $p(\mathbf{y} \mid \mathbf{x}, \gamma)$. In Fig. 2 we plot $p(\mathbf{y} \mid \mathbf{x}, \gamma)$ against $x_1' = \{x_1 - E(x_1)\}/\text{c.s.d.}(x_1)$ where $E(x_1)$ is the posterior mean of x_1 and c.s.d. (x_1) is the conditional standard deviation of x_1 given the other parameters, both estimated via our normal approximaton. In this plot $x_2, x_3, \ldots$ are fixed at $\hat{x}_2, \hat{x}_3, \ldots$. Superimposed on the plot is the standard normal density $(2\pi)^{-\frac{1}{2}} \exp(-x_1'^2/2)$. As may be seen, the two curves are indistinguishable. We conclude that the conditional posterior distribution of x_1, given the other parameters and assuming a flat prior distribution, is exactly the normal distribution derived from our approximation. Similar plots were obtained for x_2 to x_{13} but not for γ, whose conditional posterior distribution was significantly skewed to the right (Fig. 3). This figure suggests, however, the transformation $\gamma \leftarrow \log \gamma$, i.e. we treat $\log \gamma$ as the unknown parameter in place of γ itself. After the transformation the picture is much closer to the normal distribution. We also tried some contour plots of $\log p(\mathbf{y} \mid \mathbf{x}, \gamma)$. Fig. 4 shows a plot for the variables (x_1, x_2), other parameters held fixed. The contours are nearly circular, suggesting good conformity with the bivariate normal approximation to the conditional posterior distribution of (x_1, x_2). Similar plots were obtained for other pairs of x-values. Similar plots involving γ have shown that the posterior joint distribution of x_1 *and* γ, say, is far from normal,

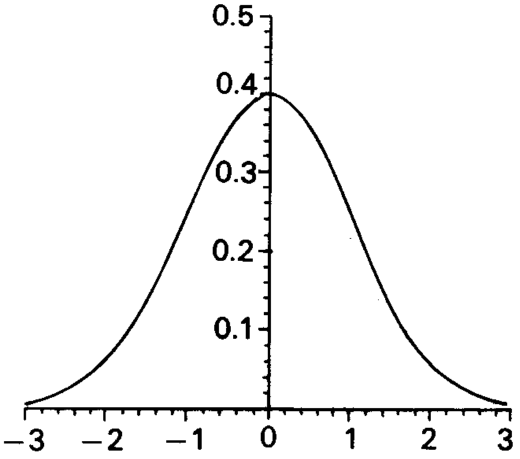

Fig. 2. Conditional posterior density of x_1

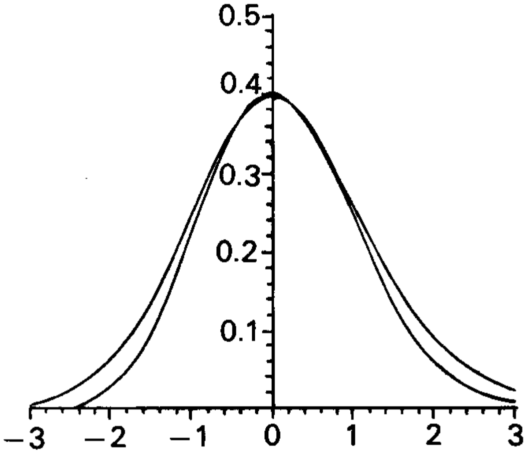

Fig. 3. Conditional posterior density for γ and standard normal curve

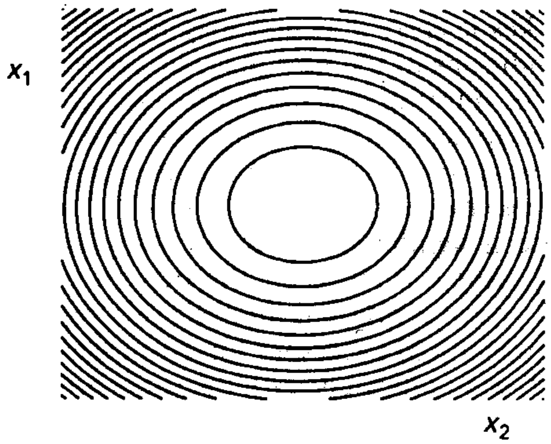

Fig. 4. Contour plot for log posterior density of (x_1, x_2)

but after the transformation to log γ the contours are again nearly circular, suggesting that the bivariate normal approximation is now acceptable.

These plots cannot be taken as conclusive, since they are only one- and two-dimensional projections of a 14-parameter function. Nevertheless they provide an indication that the normal approximation is good, provided we make the transformation from γ to log γ.

(c) *Effect of m*

Three values $m = 0, 1, 2$ were tried in (3.2) and $-\log p(\mathbf{y} \mid \mathbf{x}, \gamma)$ minimised with respect to $\mathbf{x}$ and γ, separately for each m. The minimum values obtained were 1578.3, 1552.5, 1633.8, respectively. This suggests that $m = 1$ is the best fit, which accords with intuition as explained in Section 3. The effect on $\hat{x}$, however, is not great: 30,903.7 m. (s.e. 2.0) with $m = 1$, 30,902.8 m. (s.e. 2.5) with $m = 2$. Presumably, the correct choice of m would be important if the method were applied over a wider range of distances, for example if the whole course were measured at once instead of being broken into sections. The effect on γ was noticeable: $\hat{\gamma} = 2.07$ (s.e. 0.32) under $m = 1$, 1.30 (s.e. 0.20) under $m = 2$. Note that, contrary to expectation, we now have $\hat{\gamma} > 1$ (but see (e) below).

Since data and intuition both suggest $m = 1$, we adopt this value in subsequent analysis.

(d) *Effect of γ*

The estimates of $\hat{\gamma}$ in the cases $m = 1$, 2 suggest that there is evidence against $\gamma = 1$. Nevertheless, when we re-fitted the model fixing $\gamma = 1$, we found no change in the estimates of $\hat{x}$ and only slight changes in standard error (1.8 in the case $m = 1$). We conclude that the practical effect of γ is very small.

(e) *Changes in the calibraton coefficients*

The possibility that the calibration coefficients are changing with time has already been noted (Fig. 1). This effect is potentially far more serious than any of those discussed so far, and is discussed at length in the Los Angeles report (Brennand *et al.*, 1984).

Residuals may be formed from the calibration data by defining

$$r_{ij} = \tilde{\sigma}^{-1} s_{ij1}^{\frac{1}{2}}(y_{ij1} - \tilde{\beta}_j z_i), \quad 1 \leqslant i \leqslant p, \quad 1 \leqslant j \leqslant n$$

with $\tilde{\sigma}$, $\tilde{\beta}_j$ the Bayes or least squares estimators. Neglecting the effect of parameter estimation, these are independent standard normal under the assumed model. For the present data, the row means $n^{-1}\Sigma_j r_{ij}$ are 0.83, 1.04, 0.88, 1.14, 0.72, 0.44, 0.50, -0.77, -1.39, -0.09, -1.09, -1.24. There appears to be a clear separation between the first seven and the last five, confirming the impression created by Fig. 1. No other plots or summary statistics calculated from the residuals produced any clear evidence of departure from assumptions.

The Los Angeles report makes it clear that the early measurements (first 12 rows of data) were collected in cool, wet conditions in the morning, the remainder in warm, dry conditions in the afternoon. This suggests splitting the data into two groups and applying the same method of analysis to each group separately. Doing this with $m = 1$, γ unknown, we obtain $\hat{x} = 12{,}610.4$ (s.e. 0.6) for the five course intervals in the morning session, $\hat{x} = 18{,}299.7$ (s.e. 0.8) for the eight in the afternoon session, a combined total of 30,910.1 (s.e. 0.9). This is some 6 m longer than the earlier estimate, a significant change in comparison with the standard errors. The corresponding values of $\hat{\gamma}$ were 0.20, 0.38 for the two groups.

Taking into account Fig. 1 and the residual analysis just reported, we conclude that this splitting of the data is an important and necessary improvement on the original analysis. The comparisons reported in (b)–(d) were repeated, with the same broad conclusions as previously. In particular, although we now find $\hat{\gamma} < 1$, we could fix $\gamma = 1$ with very little change in the conclusions (total again 30,910.1, s.e. 1.1).

(f) *Comparison between maximum likelihood and Bayesian analyses*

As an alternative to the foregoing analyses, we could work directly with the profile likelihood as shown in Section 3. Repeating the analysis of (e) (split data, $m = 1$) we note some change in $\hat{\gamma}$ (0.23, 0.48 for the two groups) but no change at all in the total $\hat{x}$ (30,910.1 m). The reason for this may be that the final factor in (4.4), whose presence (in the case of a flat prior) is the only difference between the two approaches, is almost independent of the x_k's within the narrow range of uncertainty. It would be of interest to see to what extent this phenomenon is observed in other applications of the theory.

(g) *Normality of error distribution*

This point is important, because it is has been suggested that the principal source of error in practice is the cyclists failure to follow the correct route. If that were correct, then most errors would be the direction of over-measurement, resulting in a right-skewed distribution.

We have considered this by examining the residuals defined in (e), for the split-data model with $m = 1$. The first four sample moments of the 156 residuals were -0.02, 1.08, -0.24, 4.47 which seems reasonably consistent with the theoretical values of 0, 1, 0, 3. A probability plot (Fig. 5) is close to a straight line with the exception of a very small number of outliers. The

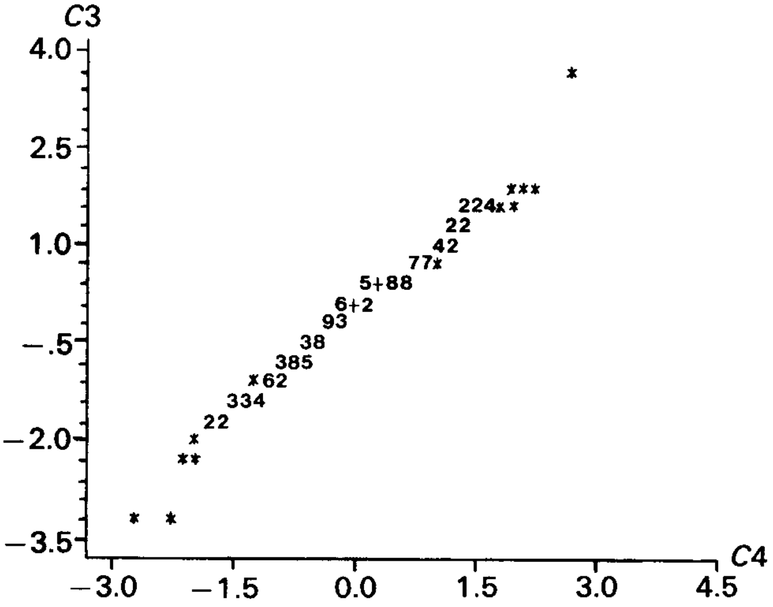

Fig. 5. Normal probability plot for standardised residuals from calibration measurements (split data, $m = 1$)

largest positive outlier (3.68), corresponds to a case in which the cyclist is known to have overshot the baseline.

We conclude that there is no evidence to contradict the normal distribution. This conclusion must be treated cautiously, however, since it is clear that the cyclists in Los Angeles were exceptionally careful to follow the correct procedure. It may well be that, for a less experienced group of measurers, there would be many outliers and a highly skewed error distribution.

6. A Dynamic Model

The Los Angeles report (Brennand *et al.* 1984) devotes much space to the effects of, and ways of dealing with, changes in the calibration constants as a reslt of meteorological or other external influences. Our own analysis has confirmed that this is a major source of concern. It is therefore of interest to explore other ways of dealing with the problem. In this section we fix $m=\gamma=1$.

A broad class of models is expressed by the equation

$$\beta_{j,t}=\beta_{j,t-1}+\varepsilon_t, \quad t=1, 2, \ldots, T, \tag{6.1}$$

where $\beta_{j,t}$ represents the value of β_j at the tth calibration interval ($t=0$ at the start), and ε_t, $t \geqslant 1$, represents a series of disturbances. There are numerous ways of modelling these disturbances. The Los Angeles report considers models which are effectively equivalent to assuming either a parametric regression function or nonparametric but subject to a local smoothing requirement. We consider an alternative *dynamic* model in which the disturbances are treated as random variables. This is very much in the spirit of state-space approaches to time series problems (see e.g. West *et al.* 1985) and also fits in very well with the Bayesian strategy we have adopted in earlier sections.

The prior distribution of $\boldsymbol{\beta}$ and σ^2 is again assumed to follow (4.3) with $v=0$, $a_j \equiv 0$. Conditionally on $\boldsymbol{\beta}$ and σ^2, we take $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T$ to be independently $N(\mu, \sigma^2/\tau^2)$ where μ and τ are two additional parameters. Thus we obtain the prior density

$$\pi(\boldsymbol{\beta}, \boldsymbol{\varepsilon}, \sigma^2 \mid \mu, \tau) \propto \sigma^{-n-T-2}\tau^T \exp\left\{-\frac{\tau^2}{2\sigma^2}\sum_1^T (\varepsilon_t-\mu)^2\right\}. \tag{6.2}$$

The likelihood function is given by (3.3) with β_j at each appearance replaced by the appropriate value of $\beta_{j,t}$. Multiplying prior and likelihood together, we have

$$p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\varepsilon}, \sigma^2 \mid \mathbf{x}, \mu, \tau) \propto c_1 \sigma^{-n(p+q+1)-T-2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\theta}'A\boldsymbol{\theta}-2\boldsymbol{\theta}'\mathbf{b}+c_2)\right\} \tag{6.3}$$

where $\boldsymbol{\theta}'=(\boldsymbol{\beta}'\boldsymbol{\varepsilon}')$, and the matrix A, vector $\mathbf{b}$ and scalars c_1 and c_2 depend on $\mathbf{y}$, $\mathbf{x}$, μ and τ. Completing the square and integrating with respect to first $\boldsymbol{\theta}$ and then σ^2, we find

$$\boldsymbol{\theta}'A\boldsymbol{\theta}-2\boldsymbol{\theta}'\mathbf{b}+c_2=(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'A(\boldsymbol{\theta}-\boldsymbol{\theta}_0)+R,$$

where $\boldsymbol{\theta}_0$, the Bayes estimator of $\boldsymbol{\theta}$ given $\mathbf{x}$, μ, τ, satisfies

$$A\boldsymbol{\theta}_0=\mathbf{b}.$$

Hence

$$p(\mathbf{y} \mid \mathbf{x}, \mu, \tau) \propto c_1 \mid A \mid^{-\frac{1}{2}} R^{-(p+q)(n/2)}. \tag{6.5}$$

In the same way as before, we may think of this expression as a likelihood function of $\mathbf{x}$, μ and τ, maximising to obtain estimators $\hat{x}_1, \ldots, \hat{x}_q$, $\hat{\mu}$, $\hat{\tau}$. An alternative "no drift" model is to fix $\mu=0$, maximising with respect to the other parameters.

Numerical implementaton was based on the NAG routines *F03AEF* and *FO4AGF*, to solve (6.4) and evaluate $|A|$, within the routine *E04CGF* for numerical optimisation. Despite the apparent complexity of this procedure, maximisation required only 6–7 seconds CPU time.

We treated the first four calibration measurements (all of the same baseline, measured consecutively) as corresponding to $t = 0$, then the subsequent baselines as $t = 1, 2, \ldots$ up to $t = T = 7$ for baseline B7 which was measured twice. Under $\mu = 0$ we obtained Bayes estimates of $\varepsilon_1, \ldots, \varepsilon_7$ as -1.39, -0.39, -1.18, -8.72, -3.53, 2.60, -1.38, which follow well the pattern in Fig. 1. The overall $\hat{x}$ was 30908.19 (s.e. 1.69); $-\log p(y \mid \mathrm{x}, \mu = 0, \hat{\tau}) = 1439.0$ ignoring an additive constant. Under the alternative model with μ unknown we estimate $\varepsilon_1, \ldots, \varepsilon_7$ as -1.40, -0.55, -1.45, -8.11, -3.51, 2.23, -1.33 with $\hat{\mu} = -2.02$; $\hat{x} = 30908.08$ (s.e. 1.69); $-\log p(\mathrm{y} \mid \mathrm{x}, \hat{\mu}, \hat{\tau}) = 1437.9$. There is no significant difference in $\hat{x}$ and a "likelihood ratio test" based on the usual chi-squared approximation would accept $\mu = 0$. The two models ($\mu = 0$, μ unknown) would of course differ if used to predict future values, a point which may be relevant if considering the use of such models in a more general calibration context.

7. Summary and Conclusions

We have considered two principal approaches, a "static" approach, developed in Sections 3 and 4 and applied separately to the two halves of the data in Section 5, and a "dynamic" approach developed in Section 6. The latter approach is the more aesthetically pleasing since it avoids the somewhat *ad hoc* splitting of the data. The results, for the 13 sections being measured, are summarised in Table 2, together with the corresponding results from the Los Angeles report. The latter are extracted from a mass of results given in the report. Avoiding technical details, they are the median results for the 13 cyclists and the minimum of all 13, each cyclist's data being processed to obtain a separate estimate for each section.

It can be seen that the first three columns are very similar, the spread of the final estimates being less than 2 m and therefore totally insignificant in practical terms. The final column was however, the one which the Los Angeles authors finally adopted as their "official" measurements. The rationale behind this was to provide some insurance against the risk of a short course ("short course prevention factor"). The main advantage of our approaches over theirs is that we are able to obtain approximate standard errors, or standard deviations of the posterior distribution, based on a combinaton of data from all the cyclists. These have been

TABLE 2
Comparison of results

Section	*Our results*		*Los Angeles Report*	
	Method 1(1)	*Method 2*(2)	*Median*(3)	*Minimum*(4)
1	1293.91	1293.84	1293.45	1292.51
2	1593.96	1593.87	1593.84	1592.80
3	3572.71	3573.03	3573.38	3571.52
4	4232.94	4233.49	4233.71	4230.94
5	1916.82	1917.31	1917.61	1916.35
6	2552.43	2549.53	2551.39	2549.56
7	4270.59	4269.68	4269.63	4267.21
8	2034.42	2033.99	2034.18	2033.32
9	2779.96	2780.41	2779.75	2778.89
10	5306.85	5307.70	5307.11	5303.71
11	611.10	611.03	611.14	610.65
12	575.76	575.69	575.81	575.33
13	168.64	168.62	168.70	168.00
Total	30910.09	30908.19	30909.70	30890.79

(1) As in Section 5: split data, $m = 1$
(2) As in Section 6: dynamic model, $\mu = 0$
(3) Median of 13 cyclists
(4) Minimum of 13 cyclists

very small, around 2 m, suggesting that the Los Angeles measurers were in fact being excessively conservative in their final report. The discrepancy of around 20 m is still very small in comparison with the length of a marathon, but it represents about 4 seconds running time which could be significant if a world record is at stake.

In conclusion, we have been surprised by the small standard errors, which are much smaller than the generally accepted tolerance of 40–50 m. We believe that this results from a combination of three factors, (a) that 13 cyclists measured the course instead of the usual one, (b) there were frequent calibration intervals incorporated into the course, (c) the cyclists were all experienced in the method and were exceptionally careful to eliminate all controllable errors. As an example of statistical calibration theory, the analysis illustrates the practical application of the Bayesian approach, and through the dynamic model suggests a possible general approach to recalibration problems.

Acknowledgements

The bulk of this paper is based on the second author's M.Sc. thesis, written at Imperial College under the direction of the first author. The work was supported by an S.E.R.C. Studentship. We are grateful to Mr John Jewell for informing us about the Los Angeles report, and for numerous discussions on the subject. We thank Mr Robert Letson for permission to quote material from the Los Angeles report. We also thank Professor P. J. Brown for some comments and suggestions, and for sending us a copy of Brown and Sundberg (1987).

References

Brennand, J., Scardera, R., Letson, R. A., Knight, T., Corbitt, T. and Steinfeld, A. (1984) *Final Certification Report on the 1984 Olympic Marathon.* Circulated privately by Robert A. Letson, 4369 Hamilton St. No. 4, San Diego, CA 92104, U.S.A. (225pp.)

Brown, P. J. (1982) Multivariate calibration. *J. R. Statist. Soc.* B, **44**, 287–321.

Brown, P. J. and Sundberg, R. (1987) Confidence and conflict in multivariate calibration. *J. R. Statist. Soc.* B, **49**, 46–57.

Fieller, E. C. (1954) Some problems in interval estimation. *J. R. Statist. Soc.* **B**, **16**, 175–185.

Hoadley, B. (1970) A Bayesian look at inverse linear regression. *J. Amer. Statist. Ass.*, **65**, 356–369.

Hunter, W. G. and Lamboy, W. F. (1981) A Bayesian Analysis of the linear calibration problem. *Technometrics*, **23**, 323–328.

Jewell, J. C. (1961) Report on road course measurement. In *Athletics Weekly*, April 1961.

Krutchkoff, R. G. (1967) Classical and inverse regression methods of calibration. *Technometrics*, **9**, 425–439.

West, M., Harrison, P. J. and Migon, H. S. (1985) Dynamic generalized linear models and Bayesian forecasting. *J. Amer. Statist. Ass.*, **80**, 73–97.

Williams, E. J. (1969) A note on regression methods in calibration. *Technometrics*, **11**, 189–192.

ROYAL STATISTICAL SOCIETY
25 Enford Street, London W1H 2BH

Printed in Great Britain by Henry Ling Ltd., at the Dorset Press, Dorchester, Dorset