A SIMULATION OF THE FINISH-LINE SYSTEM
OF A MAJOR RUNNING RACE

R.C. VAN DEN HONERT
Department of Statistical Sciences
University of Cape Town, Rondebosch 7700
South Africa

and

K. WESTVIG
Decision Dynamics, P O Box 44445
Claremont 7735
South Africa

ABSTRACT

The finish-line system of a major road race is investigated by means of a simulation model, in an attempt to help race management identify conditions that may lead to congestion of runners at this point. Based on the results of the model, race management was presented with a simple tool which could be used *in situ* to analyse a given situation, identify if congestion was likely to occur, and then consider corrective action.

INTRODUCTION

The Two Oceans Marathon held each Easter Saturday around the scenic Cape Peninsula is one of South Africa's most popular road races. Since it's inception in 1970 the number of entries in this 56 km ultra-marathon has grown to over 9000 per annum, and the race is covered live on national television. Race management have acquired a reputation for providing an efficiently-organised race, and in order to maintain this reputation they are committed to constantly investigating ways in which the race organisation can be improved.

One of the main concerns of any race management centres around the finish-line. If the runners arrive at the finish line at a rate significantly greater than the rate at which they can be "processed", a queue will build up in the finishing lanes (and even back up on to the course itself) (TAC [4]). Since every participant in the Two Oceans Marathon receives comprehensive computerised results by time, sex, age, team, time-splits etc (one of the features of this race), the exact finishing time and details of each runner has to be collected as they cross the finish-line. The processing of a runner comprises correlating the runner's finishing time and position with his/her personal identity and other details.

Whilst the race has never yet been marred by serious runner-congestion at the finish-line, Two Oceans management felt that recent races had come close to this situation (although they had no quantitative evidence of this), and hence were keen to have the current finish-line system investigated for efficiency. This investigation would, it was hoped, give them a "handle" on the extent of the problem. It would also indicate what measures should be taken to ensure that congestion is avoided in future races. In general, build-ups of runners would occur if either (1) the number of finishers increased whilst the processing capability remained constant (e.g. increased race entry); or (2) the number of finishers remained constant whilst the processing capability decreased (e.g. some of the finishing-lanes (in which the data is captured) being closed due to hardware failures or other causes). Race management were happy to consider a worst-case analysis, i.e. examination of the worst possible situation that might arise from any scenario. Providing the worst case situation lay within some boundary conditions decided upon by management (in terms of queue-lengths or queueing times), the queues were deemed to be "under control". It was thus decided to construct a decision support system which could be used before the race (once the total race entry was established) and during the race (as and when lane closures might occur) to establish the maximum extent of finish-line congestion. In the longer term, the system could be used to identify to what extent certain changes to the race organisation might help reduce finish-line congestion.

THE FINISH-LINE SYSTEM

In the past few years the finish-line system has been computerised, and the organisation and management of this aspect of the race (hardware and software) is handled by a professional computer consulting company. Each runner's personal information is entered onto a data-base as their entry is received. As each runner approaches the finish of the race on race-day, he/she chooses one of a number of parallel finishing-lanes (which begin about 10 metres after the official finish-line), and moves down the lane to a point where his/her race number (which is worn on his/her running vest) is read by a computer-terminal operator. The operator enters the number into the computer (the operator attempts to record the runner as close to the official finish-line as possible). Further down the lane another race official removes a tear-off strip from the runner's race number. This strip contains the runner's race number in bar-code form. The bar-code is "read" by a further race official using a light-pen, and the strip is filed. The strip can be used to supply a manual back-up of finishing positions in the event of the major electronic (computerised) position/time recording system being rendered non-functional. A software program matches the runner's race number as entered in the computer with the bar-code number entered from the strip, effectively "processing" the runner. The program then matches the race number with the runner's identity and personal particulars from the data-base and stores the runner's elapsed race time and position in the data-base for results processing by the computer consulting company immediately after the race has terminated. The runner then leaves the lane. Figure 1 shows the typical finish-lane layout. The entire process (tearing off of the strip and recording of the runner's time) takes only a few seconds.

A SIMULATION MODEL

Maltz [2] describes a method for preparing results for road races that ensures finishers of an accurate statement of their order and time of finish. However it provides no insight as to how congestion may be avoided. TAC [4] describes various possible finish-line systems, and provides empirical information on the maximum rate of queue build-up for the case where the finishing times are normally distributed. Investigation of the finishing times of runners in recent Two Oceans Marathons have shown that they are far from normally distributed.
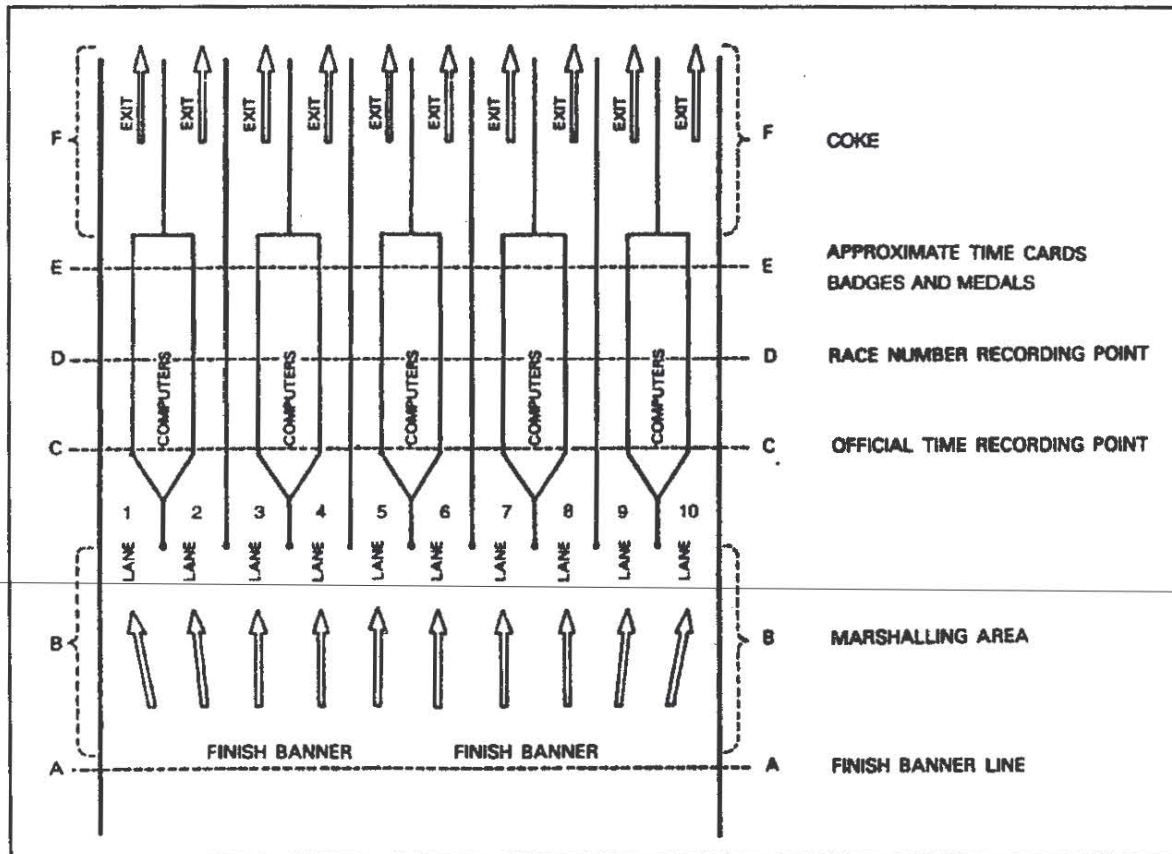
Figure 1 Typical finish-lane layout of the Two Oceans Marathon

The situation at the finish-line of a road race can be viewed merely as a large queueing system. Simulation is a standard technique to investigate discrete event systems in which queueing might take place, and it was decided to construct a simulation model of the race to enable race management to investigate the finish-line situation. An application of the use of simulation to provide input for improved planning of a road race can be found in Farina et al [1], who used a model to experiment with various interval starts and finish-lane configurations in the successful Bolder Boulder 10 km race in the USA.

A typical simulation model for the analysis of the finish-line of a road race is fairly simple. The "runners" are all started off together in a single batch and are "delayed" for a time equal to the time it takes them to complete the race. They then select one of the available lanes after the finish-line, and are processed after waiting their turn in the queue they have selected. Thereafter they leave the system. Inputs that would have an impact on the queue-lengths at the lane fronts

are the distribution of race times for all runners, the distribution of processing-times in the finish-lanes and the distribution of lane usage. Empirical data from the 1992 Two Oceans Marathon will be used to discuss appropriate inputs. Figure 2 shows the frequency distribution of finishers in the 1992 race in 5-minute intervals (the same general trends have held for the past several years). It will be noted that the frequencies do not increase smoothly as time increases: noticeable are peaks just before the 4-hour mark (the cut-off time to earn a coveted silver medal), the 4½-hour mark, and the 5-hour mark, which are obvious goals for runners. The peak just before 4 hours 40 minutes corresponds to an average running speed of 5 minutes per km. Furthermore, in 1992 almost 50% of the field finished in the last 40 minutes of the race (the race has a cut-off time of 6 hours), arriving at an almost constant rate throughout this period and creating the greatest pressure on the finish-lane operators. The finishing-time distribution observed here is quite unlike that described in Farina et al [1], who observed that finishing times in the Bolder Boulder 10 km race were lognormally distributed, or the experience of TAC [4], who state that "the distribution of finishers as a function of time is roughly a normal distribution for most races".
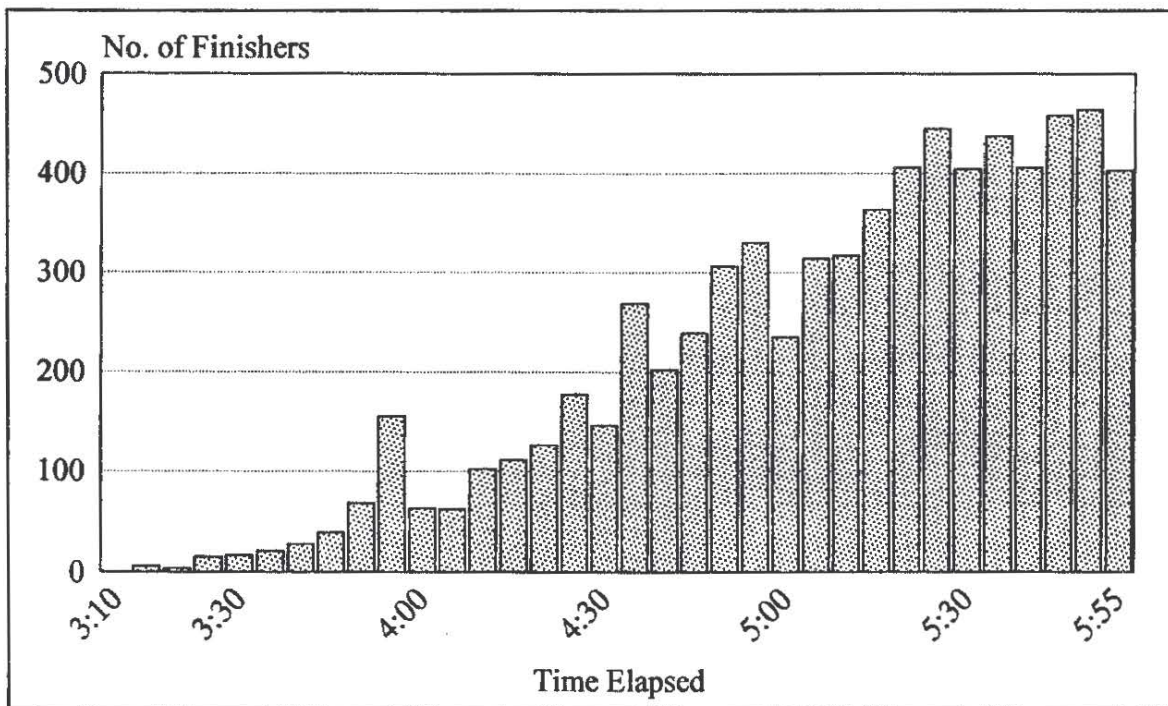

Figure 2 Frequency of finishers in the 1992 Two Oceans Marathon

The process-time distribution was also created from historical data of the 1992 race. A block of 100 successive finishers from each finishing lane was randomly chosen from the last 40 minutes of the race (i.e. when there would be a virtual continuous flow of runners) and examined. This allowed an empirical distribution of processing time to be built up for each lane. Observation showed that the processing-time distributions were almost identical from one lane to another (implying that the operators were all working at approximately the same rate), and thus the data could be lumped together to yield a single process-time distribution applicable over all lanes. Since the assumption of a continuous flow of runners could not be guaranteed, this empirical distribution provided an upper bound on the processing times. After discussion with finish-line officials it was decided that, for simulation purposes, the process-time distribution be compressed downwards, and approximated by a triangular distribution with minimum of 1,5 seconds, mode of 3,0 seconds and maximum of 4,8 seconds. The empirical distribution of lane usage is shown in Figure 3.
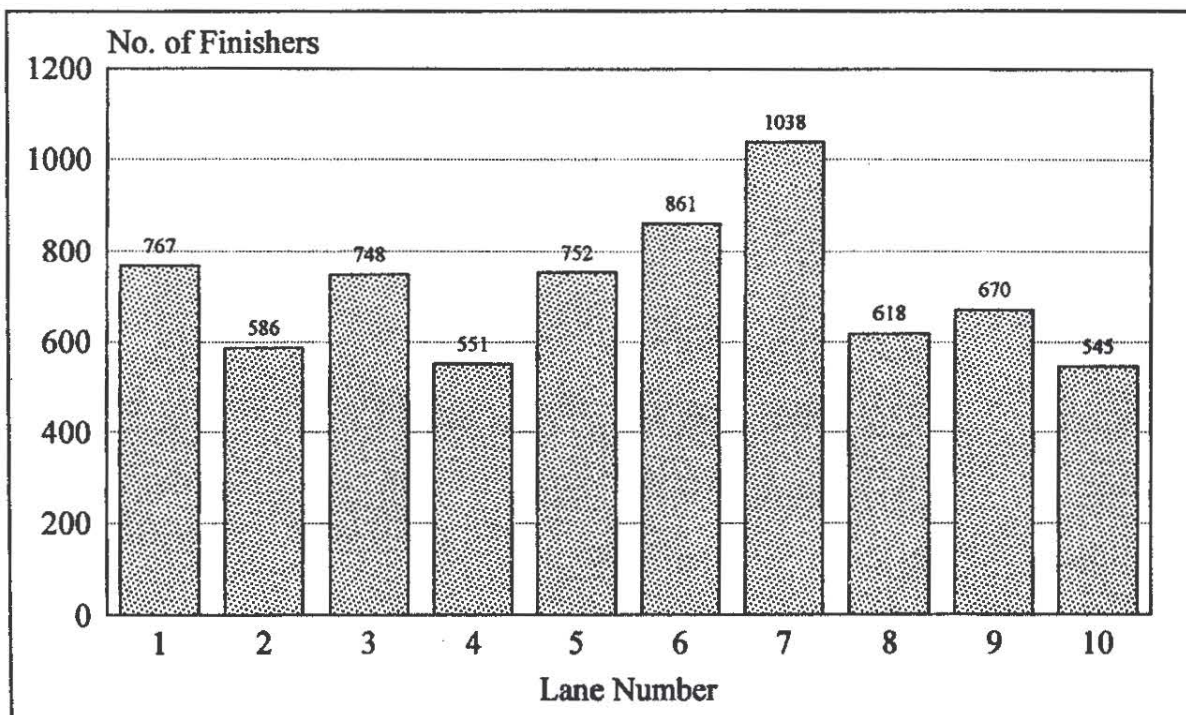
Figure 3 Distribution of lane usage

Lane 7 had the highest usage, possibly since it was midway between the "inside" lane as the runners came around the last bend, and the middle of the track. The

other lanes had appreciably less usage. This could in part be because some of them were closed temporarily, mainly due to hardware failures. Note that lane 1 (the extreme outer lane) proved to be relatively popular. This was because all prize-winners, collapsed runners and permanent-number achievers (10 races run) are guided through this lane.

Since race management were happy with a worst-case analysis, it was decided to develop a single-queue, single-server model in place of a multiple-queue, multiple-server model. This approach does away with the need to use a distribution of lane usage, or to attempt to model complex queue-picking rules. Note that this approach will tend to provide overly-conservative results. Since lane 7 processed the most runners in 1992 (1038 runners), it can be inferred that if there was to be congestion in any lane, this was the most likely candidate lane. Thus in the model 1038 runners "ran" the race and they all finished down a single lane. The simulation model was written in SIMAN IV, a specialised simulation language (see Pegden et al [3]), and all runs were performed on a 386 PC. It is worth noting here that SIMAN can handle multiple-queue, multiple-server configurations, such as in Figure 1, and allow for breakdowns. There is, however, no empirical information available regarding time between breakdowns or repair times, nor is anything known of the queue-picking behaviour of runners. The objective of the whole study, though, is the identification of the worst possible case, and this can be adequately modelled by means of a single-queue, single-server model. In fact, since the single-queue, single-server model does not consider breakdowns, the model may indeed not be as overly-conservative a representation of reality after all. Twenty replications were performed and the outcomes were examined together with race management to ensure that the model was indeed modelling the situation observed in 1992 and 1993. Statistics examined included the maximum queue-length and the maximum time spent in the queue, as well as the dynamics of these time-dependent variables. Testimony of race officials present at the finish-line allowed a check as to whether or not the model was accurately reflecting reality. The simulated results (averaged over all 20 replications) show that this tunnel reached an average *maximum* length of 6,15 runners with a standard deviation of 1,18 runners, and the *maximum* time queued was as long as 18,2 seconds on

average with a standard deviation of 3,6 seconds. Since race management were only considering the worst possible case, they could be 95% sure that the maximum queue-length would not exceed 6,61 runners and that the maximum time spent in the queue would be less than 19,6 seconds.

Figure 4 shows how the queue-length changes through time for a single replication with 1038 finishers. For well over 95% of the time the queue-length is 3 or less and only occasionally rises above 5, reaching a maximum of 8 in the last minutes of the race.
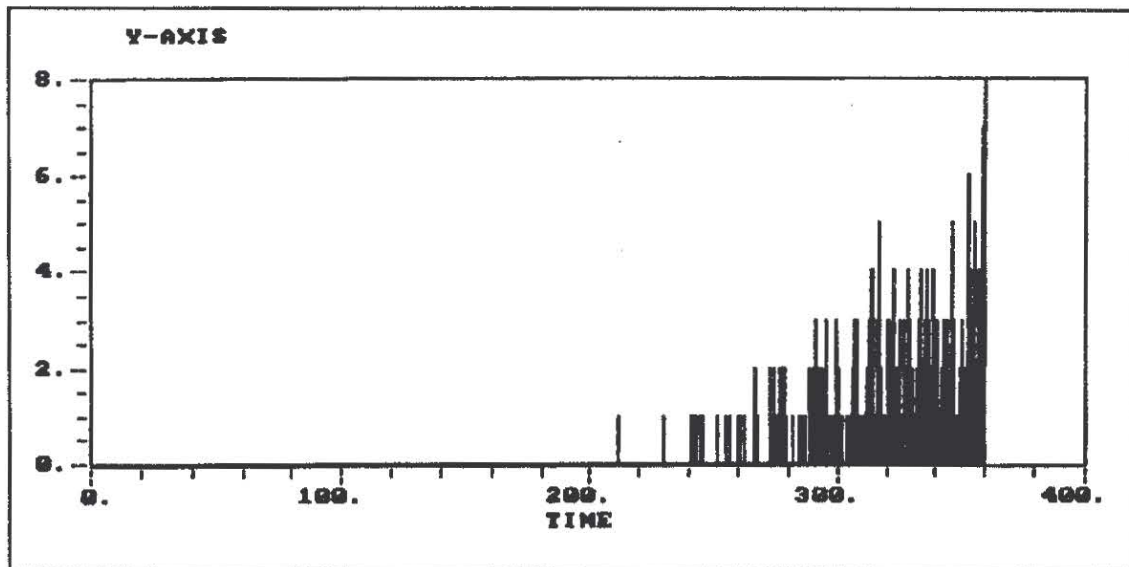

Figure 4 Maximum queue-length through time (1038 finishers)

The dynamics of the maximum time spent in the queue show similar patterns: in the single replication considered above, for the bulk of the race the time spent in the queue is acceptable (below 10 seconds = 0,167 minutes), but for short periods it increases significantly, peaking at almost 30 seconds at race closure. Race management deemed this queue-behaviour to be representative of reality. Furthermore they considered the system to be operating "under control"; a queue-length of around 10 and a waiting-time of 30 seconds were considered to be appropriate maxima.
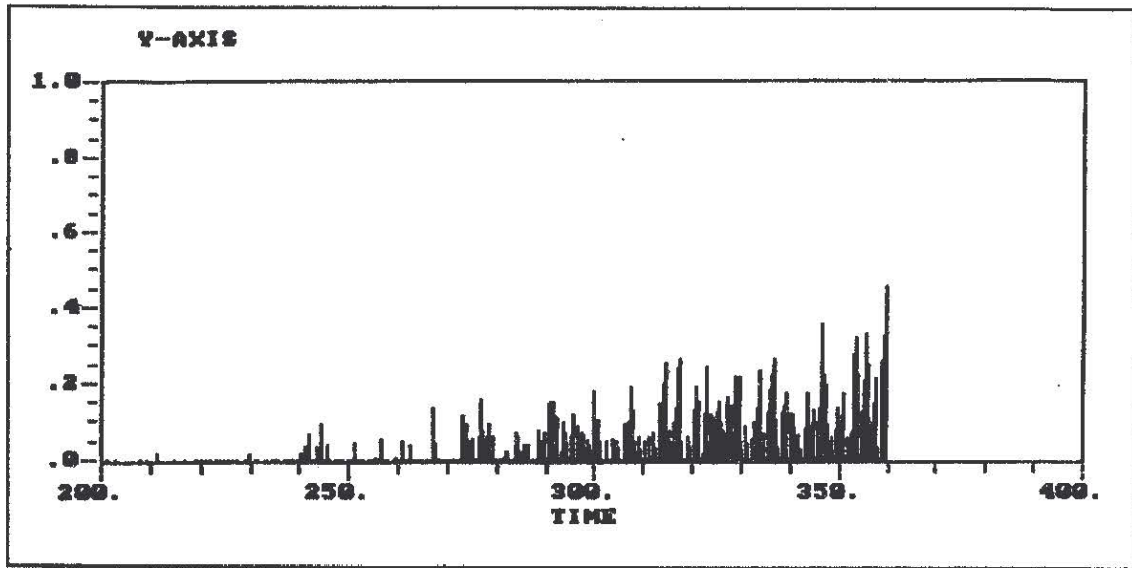
Figure 5   Maximum time spent in the queue (1038 finishers)

SIMULATION RUNS

Once the simulation model had been validated on recent runnings of the race, various alternative scenarios for future races could be formulated and tested using the simulation model. Figure 6 shows the number of finishers for the years 1977 to 1993.
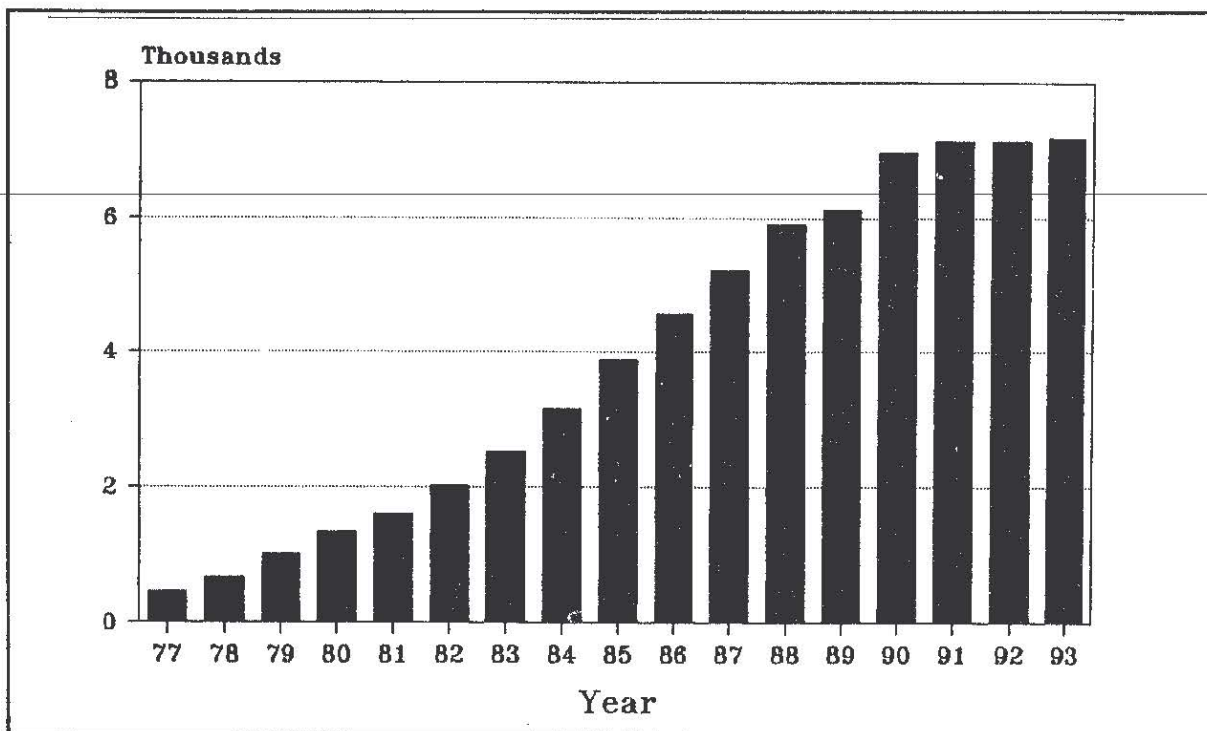

Figure 6   Number of finishers, 1977 - 1993

In the late 1970's and early 1980's in particular the race showed spectacular annual growth. Overall there has been a compounded growth in finishers of 18,8% per annum over the last sixteen years. This includes a recent period of virtually no growth in finishers from 1989 to 1993, but this was ascribed to the economic recession in the country at the time which would have discouraged many up-country runners (usually making up over 75% of the field) from making the journey to the Cape. However with South Africa's readmittance to world sport, the Two Oceans has been widely advertised overseas and race management believed that there will be a steady, if not spectacular, annual increase of 3% to 5% in entries in the near future. Furthermore, computer hardware breakdowns have become a regular feature in the finishing-lanes. Single lane breakdowns have been observed, but have caused only limited and temporary congestion until repairs could be effected. Race management were becoming increasingly worried about the prospect of multiple lane breakdowns compounded by an increased field size.

An increase in the number of finishers of, say, Y% might be expected to increase the number of finishers in each lane by Y%. It was assumed that all "new" finishers would have the same finishing-time distribution as the current (observed) finishers, and thus the relative frequencies observed in Figure 2 would remain unaltered. If a lane is closed, the runners that "would have" finished in that lane must now be redirected to another lane. In practice a marshall merely stands in front of the closed lane (thus blocking it) and the runners themselves decide on an alternative finishing lane. Race officials have observed that in this event runners approaching a closed lane are most likely to redirect themselves to a lane adjacent to the closed one, or, less likely, to a lane two away from the closed one. Race management decided (based on their observations) that an adjacent lane is used about 70% of the time and a lane two away the remainder of the time. (Naturally this logic breaks down at the edges (lanes 1, 2, 9 and 10): they decided that if lane 1 or 10 was closed, 70% of runners would choose the adjacent lane (lane 2 or 9 respectively) and the balance would choose the next lane (lane 3 or 8 respectively). If lane 2 or 9 were closed 50% would choose the extreme (outer or inner) lane, 35% would choose the other adjacent lane and the balance would choose the next lane towards the middle of the track). The effect of the

redirection of finishers to alternative lanes would be to increase the number of finishers in those lanes. The distribution of arrivals through time in those lanes would, of course, depend on exactly when the lane closure occurred, but the worst possible scenario would be if the lane was closed from the start.

With reference to Figure 3, the empirical distribution of lane usage in 1992 was affected by lane closures (lane 4 was closed for a significant period). Furthermore, in 1993 lane 6 (as opposed to lane 7 in 1992) had the greatest through-flow of finishers. It was thus decided that this distribution be smoothed to reflect more accurately the situation that may exist if no lane closures occurred. The simulation model was rerun using the smoothed distribution of lane usage for various increases in the number of finishers in the lane. The mean maximum queue-lengths and waiting times for processing (in seconds), as well as upper 95% confidence limits for these amounts are given in Table 1 and shown graphically in Figures 7 and 8 respectively.

	No. of runners	Maximum queue length		Maximum waiting time	
		Mean	Upper level	Mean	Upper level
Status quo	1035	6,2	6,6	18,2	19,6
10% increase	1140	7,6	8,2	21,9	23,5
20% increase	1245	8,7	9,3	25,4	27,5
30% increase	1350	10,2	11,1	30,3	33,1
40% increase	1450	11,8	12,6	34,5	37,4
50% increase	1550	17,5	18,8	51,7	55,0
60% increase	1650	31,8	33,7	87,5	92,5

Table 1   Maximum queue-lengths and waiting times for $n \geq 1035$ runners

The upper 95% confidence limits do not deviate appreciably from the mean queue-lengths and waiting times since the standard deviation of these values over all replications of the simulation is small in all cases. From Figures 7 and 8 it will be noted that, contrary to race management's beliefs, the current situation is far from being out of control: the current finish-line system could apparently tolerate a 30% increase in the number of runners before exceding race management's current boundary conditions on queue-length and waiting times. Beyond that, however,

the situation deteriorates extremely rapidly. This is based on the smoothed distribution of lane usage, which assumed no lane closures. The incidence of lane closures will negatively affect queue-lengths and waiting times.
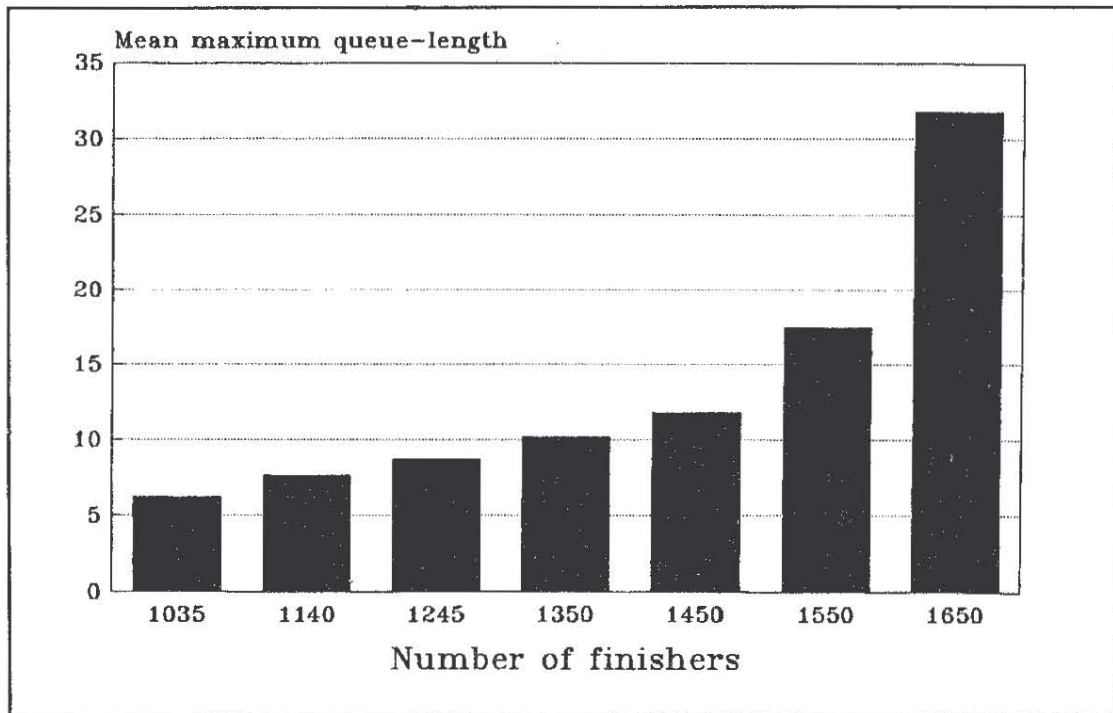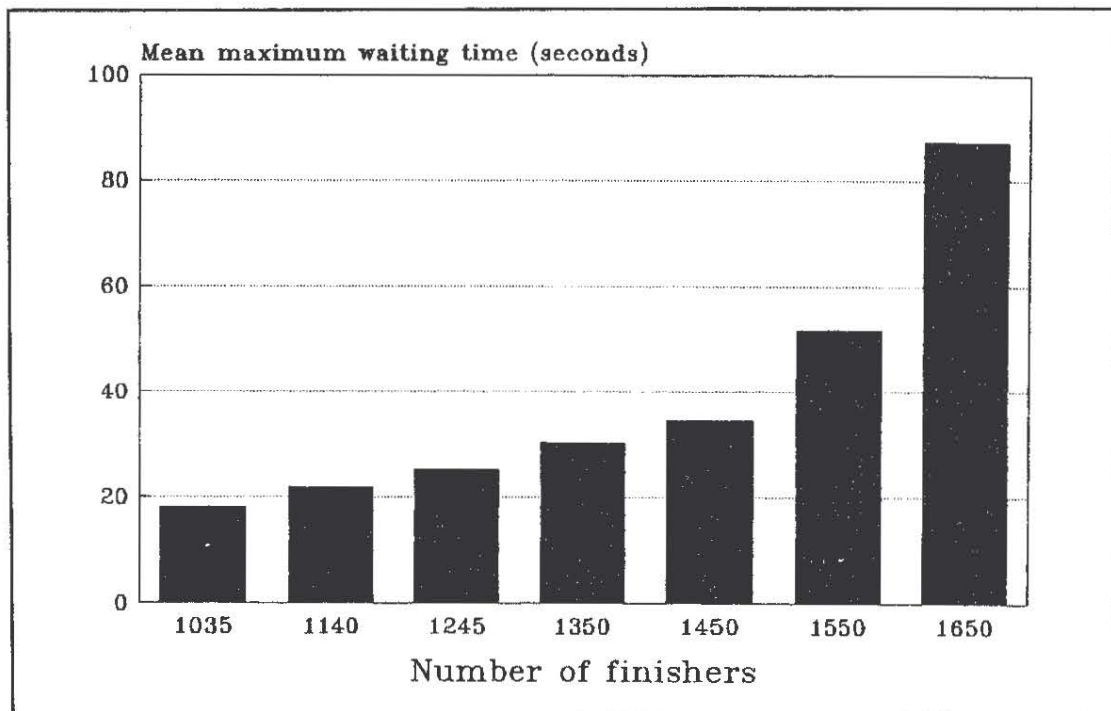

Figure 7 Maximum queue-lengths for n ≥ 1035 runners

Figure 8 Maximum waiting-times for n ≥ 1035 runners

Before the race, but once the total number of entries (or the total number of expected finishers) is known, race management may estimate the number of runners expected to finish in each lane using the smoothed distibution of lane usage. If the maximum expected lane usage is greater than about 1000, Table 1 affords them an indication of the maximum queue-lengths and waiting times they might expect. If lanes are closed at any stage during the race, the worst possible congestion effect can be guaged by determining the new total number of runners expected in each lane (using the expected number of finishers in each lane and the redirection logic described above) and comparing the maximum expected number to Table 1. For example, if 7090 runners are expected to finish the race, and if lanes 6 and 8 were both to be closed from early in the race and not reopened, the total number of runners expected to be processed in lane 7 might be as great as 1630, leading to a maximum expected queue-length of around 27 runners, who might have to wait for as long as 70 seconds to be processed (clearly out of control). This process is easily automated and speeded up by programming the logic into a PC spreadsheet program, which can be used by race management while the race is under way. The user merely enters the number of the lane(s) to be closed, and the program alerts him/her if the queueing process is likely to get out of control.

FUTURE OPTIONS FOR RACE MANAGEMENT

Management now had a simple, but satisfactory, decision-support tool which could be used *in situ* to quantify the congestion effect of increased field size and hardware breakdowns. In the light of their experience with the system they could now consider several possible alternative courses of action to ensure minimal queue build-up at the finish-line in future races. Those that could possibly be considered are:

(1)   Increase the number of lanes at the finish-line (10 were in use in 1992 and 1993);

(2)   Modify the processing procedure to reduce the time it takes to process finishers;

(3)   Modify the start procedure so that runners would arrive at the finish-line more spread out, i.e. release the runners in "blocks" of, say, 1000 runners

at constant intervals of time.

(4)   Modify the layout of the finish-line area, i.e. force certain groups of runners (for example, ladies and veterans) to finish in specified tunnels (in possibly in a different section of the finishing area).

(5)   Limit the number of entrants to the race.

Whilst race management were reticent to make any radical changes in the short-term which might mitigate against the character of the race, the simulation model can easily be modified to handle any of the above alternatives. The most likely options for future investigation appear to be options (1) and (2). At present there are 10 parallel finishing lanes used, a decision taken based on the size of the canopy used to cover and protect the finishing area (!), and which is large enough for only 10 lanes. It is possible that this could be modified or enlarged to handle more lanes if necessary. The computer consultants are furthermore keen to evaluate the effect of utilising a more efficient (and expensive) hi-tech processing procedure which has been successfully used in large road races in Britain and the USA recently. This involves the embedding of a transponder in each runner's race number (expensive) which is automatically read by an electronic reader positioned at the finish-line (efficient), thus doing away with the present semi-manual processing system in the finishing-lanes. This would thus drastically reduce the manpower requirement at the finish-line and be more reliable. It is expected that these investigations will be on-going over the next few years, and the simulation model of the finish-line could play a pivotal role in guaging the effect on the lane queues.

CONCLUSIONS

This case-study is successful proof of the versatility of OR/MS techniques, this time in the field of sport and recreation. Race management had no a priori idea of the extent of runner congestion caused by increased race entry and lane closures, and soon realised that this model may offer some quantification of this vexing problem. They thus were happy to be involved in all stages of the project, from supplying data to reviewing results and making suggestions as to future options, especially as they started to see the benefit of the models. This, together with

frequent report-backs, were a feature in the successful completion of the project and acceptance of the results (Benvenuti [5]). The model thus met with total acceptance, to the extent that the computer consultants are eager for a similar analysis to be carried out on the finish-line system at the (even larger) Comrades Marathon which has upwards of 13 000 finishers, and which they also manage. The model will provide decision support to the race organisers to help manage this tricky aspect of the race in future years.

ACKNOWLEDGEMENTS

REFERENCES

[1]   R. FARINA, G.A. KOCHENBERGER and T. OBREMSKI, The computer runs the Bolder Boulder: a simulation of a major running race, *Interfaces*, 19, 48-55 (1989).

[2]   M.D. MALTZ, Management Science in the long run, *Interfaces*, 13, 40-45 (1983).

[3]   D. PEGDEN, R.E. SHANNON and R.P. SADOWSKI, *Introduction to Simulation using SIMAN*, McGraw-Hill Inc, New Jersey (1990).

[4]   THE ATHLETICS CONGRESS OF THE USA (TAC), *Road Race and Finish Line Management*, TAC, Indianapolis (1985).

[5]   R. BENVENUTI, Critical success factors in real-world OR, Paper presented at the 1992 annual conference of the Operations Research Society of South Africa (1992).