Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

SHOULD WE MODEL DEPENDENCE AND NONSTATIONARITY, AND IF SO HOW?

Shane G. Henderson

School of Operations Research and Industrial Engineering
230 Rhodes Hall
Cornell University
Ithaca, NY 14853, U.S.A.

ABSTRACT

It can be difficult to develop, fit to data, and generate from, models that capture dependence and/or nonstationarity in simulation inputs. Should we bother? How should we go about it? I will discuss these issues, focusing on three examples: call center arrivals, ambulance travel speeds and wind modeling for America's Cup races.

1 INTRODUCTION

Suppose we are building a simulation model that describes the visitor flows through an amusement park during a summer. These flows depend on the weather, and weather is well known to exhibit dependence over time. In addition, we can expect the weather to deteriorate towards the end of the summer so that it is also nonstationary. Should we explicitly capture this dependence and/or nonstationarity in our simulation model? If so, how can we go about doing so? The goal of this paper is to provide some help in answering these kinds of questions.

More explicitly, the goal here is to impart a sense of when it is worthwhile going to the trouble of capturing dependence and nonstationarity in an input model, and to suggest how we might do so in a high-level sense. There is no attempt to survey models of dependence and/or nonstationarity. For such a survey for dependence modeling, see Biller and Ghosh (2004), or Biller and Ghosh (2006) for more details.

In trying to offer some very general guidelines, there is a danger that the guidelines become so general as to be useless. I will try to keep things grounded by discussing a number of real examples. These include call center staffing, ambulance deployment and yacht match racing.

In our context, the term "nonstationary" describes a situation where the distribution of a collection of random variables is changing over time. The term "dependent" means that there are relationships within the random variables. We can capture both terms within a single framework by referring, instead of "dependence and/or nonstationarity", to the *joint distribution* of a collection of random variables. The joint distribution fully specifies the statistical behaviour of the random variables in question, and so henceforth, we use that term.

The question of how accurately to model the joint distribution is really one of model complexity. Should the model be as similar as possible to reality, or is it acceptable to skip a few details? This question is addressed in great length in verification, validation and accreditation discussions; see, e.g., Law and Kelton (2000), Chapter 5). An explicit theme there (Law and Kelton 2000, p. 265) is that a model should be an accurate representation of the system "*...for the particular objectives of the study.*"

In this paper we take the position that the simulation study is performed to assist in making some decision. In that case our simulation model should be as simple as possible, but include the aspects of the underlying process that have an appreciable bearing on the decision. This, then, is our answer to the first question posed in the title of the paper. We will see this principle applied in several examples below. (There is nothing new in this answer – it is the standard answer to questions of model complexity.)

In Section 2 we give a brief "modeling philosophy". Then, in Sections 3 (call center staffing), 4 (ambulance deployment) and 5 (yacht match racing) we present our main examples, explaining what the key decisions are, where the complexity of the joint distribution arises, and how we apply our modeling philosophy. Finally, we reflect on the key observations in this paper in Section 6.

2 A MODELING PHILOSOPHY

We first need to identify the aspects of the stochastic process under study that need to be captured with regard to the goals of the study, i.e., answer the question "what *needs* to be modeled?" The ability to answer this question is a skill that is learnt with model-building practice, although there are (at least) 2 concrete approaches. Users of an existing

system usually have a very good sense of what factors play an important role, and such knowledge should be exploited. Also, one can use *simple* models to help determine what to include, as in the ambulance example below.

Once we have determined the aspects of the stochastic process that need to be captured, we can proceed with identifying an appropriate model. The suggested approach for this phase is as follows.

1. Try to *capture the physics.* The complex dependencies of a real system are due to some causal relationships that may be known, or at least partially known. A model that emulates those relationships then emulates the complexities of the dependence structure. The call center example below is a good example of this approach.
2. Sometimes the physical relationships that create complex dependence are not readily apparent, or are too complicated to implement. In that case, *use some construction whose capabilities and failings are well understood*. For example, Gaussian and Poisson processes fall into this category. The ambulance and yacht match racing examples below are good examples of this approach.
   In some situations one will have access to a plethora of clean data. In this case, resampling is a good option. For example, Pritsker (1998) described a simulation model of organ allocation that generates patient arrivals using a nonhomogeneous Poisson process. The characteristics of patients are independently generated by using the characteristics of a randomly selected patient out of all patients that were registered over a 5 year period. This bootstrapping procedure is often easily implemented, but care needs to be taken that the data is clean and representative.
3. Sometimes it is necessary to move beyond standard constructions like Gaussian random vectors, and little is known about the system under study. This often happens, for example, in generating test problems for algorithms, as in Hill and Reilly (2000). In such settings, there are a number of useful models that are based on *partially-specified distributions*. For example, Biller and Ghosh (2004) describe a number of methods that match marginal distributions and correlations.

Once one has obtained a specification of a model, it is then necessary to calibrate it to data, if any. It is often a good idea to try to simplify the model before doing so, because fitting complex models to data can be difficult. One approach to simplification is to run simulations using extreme values of parameters to see whether they make any appreciable difference to the simulation output. If not, then such parameters can usually be dropped from the model, with a resulting simplification in calibration.

We have presented the above steps as a linear process, but in general it is much more of an iterative process. In the next few sections we give examples of the application of this philosophy that will hopefully make this point clear.

3 CALL CENTERS

The staffing problem in call centers (or, more generally, service systems) is a problem that has been studied for a long time; see, e.g., Edie (1954) for early work. In an inbound call center customers arrive over time and receive service from a limited number of agents. The decision in question is how to schedule agents. More precisely, we wish to identify a set of agent shifts of minimum cost that still ensure that customers receive "satisfactory" service. There is a dual problem to this one that seeks to schedule an available set of employees to maximize the level of service to customers. The problems are essentially equivalent, and either statement is fine for our purposes.

It is well recognized that the call center staffing problem exhibits nonstationarity in several ways. First, the arrival rate of calls can vary with the time of day. Second, the arrival rate can exhibit seasonality, in the sense that traffic volumes can vary over a time scale of months. Third, service times can vary in length depending on the time of day; see Brown et al. (2005). We focus on modeling the arrival process.

The *Palm-Khintchine* theorem plays an important role here. See Cinlar (1972) for a very good account of this and related results. In approximate terms, the Palm-Khintchine theorem states that the arrival process that arises from a large number of independent sources, where no source contributes too much to the arrivals, is approximately a Poisson process. (Theorem 3.10 in Cinlar (1972) is a very general version of this result that applies not just to the real line, and covers nonhomogeneous limits as well as the more standard homogeneous result.) In the call center context we have many potential callers, each with a very small probability of calling at any given instant. The result then allows us to conclude that the arrival process is very well approximated by a Poisson process.

We cannot ignore the fact that the arrival rate is time-dependent. The ratio of the maximum arrival rate to the minimum arrival rate in a single day can be very large. Using a flat arrival rate over the day for selecting staff levels would imply a flat staffing level, which is completely unacceptable from the perspective of customer service; customers calling during peak periods would encounter very large delays and abandon without reaching an agent.

The extension of the Palm-Khintchine theorem stated in Cinlar (1972) suggests that an appropriate model of arrivals is a nonhomogeneous Poisson process (NHPP).

We then need to model the arrival rate function. Perhaps the simplest form for this function is a piecewise constant function with known breakpoints. This form is described in Law and Kelton (2000), pp. 390–392) and Henderson (2003).

Let us consider how to fit this function to data. Typically, call center arrival data is aggregated into periods that are on the order of 15 minutes or half an hour long. So we assume that our data consists of the number of customer arrivals per period, with data accumulated over some length of time that is on the order of months or years.

Suppose we assume (as is the usual case in practice) that the arrival rate function follows a weekly cycle. Then we wish to fit a constant arrival rate to each of the periods over the course of a week. If we have n weeks of arrival data, then for each period we have n observations of the number of arrivals in that period. The number of customer arrivals in non-overlapping time periods in a NHPP are independent Poisson random variables, so we fit an arrival rate to each period ignoring observations in other periods. If N_{ij} is the number of arrivals in period j of week i, then we estimate the arrival rate λ_j in period j using

$$\hat{\lambda}_j = \frac{\sum_{i=1}^{n} N_{ij}}{n},$$

where the units of the arrival rate are arrivals per period. (In a sense we have chosen our time units so that periods are of length 1.)

Assuming that our NHPP model is correct, the collection $(N_{ij} : i = 1, \ldots, n)$ should be a collection of i.i.d. Poisson random variables. Perhaps surprisingly, call center data often refutes this; see Jongbloed and Koole (2001), Chen and Henderson (2001) and particularly Avramidis, Deslauriers, and L'Ecuyer (2004) and Brown et al. (2005). The number of arrivals is typically "over-dispersed," that is, the estimated variance of N_{ij} far exceeds the estimated mean of N_{ij}, when under the NHPP assumption these values should be approximately equal. Furthermore, under the NHPP assumption, the number of arrivals in consecutive periods should be independent but this appears to not be the case; empirical correlations can be extremely high.

So now we come to a key modeling question. Does this matter? Or should we just model the arrival process as a NHPP anyway, ignoring the additional complexity that is apparent from the data? To answer this question, we need to think about the purpose of the simulation model. The purpose of the model is to compute the service level as a step towards determining staff shifts. If the arrival process is "more variable" than a standard Poisson process, then queueing theory suggests that the service level will be lower than it would otherwise. This belief is backed up by results in Chen and Henderson (2001) and Steckley, Henderson, and Mehrotra (2005) that prove a deterioration in performance under some conditions. If that deterioration is substantial, then we need to capture the over-dispersion in our model. It is hard to know the extent of the deterioration in service level without implementing a more complex model to begin with, but the results in Chen and Henderson (2001) and Steckley, Henderson, and Mehrotra (2005) suggest that the error can be large in some settings. So conservatively, we lean towards developing a more complex model of the arrival process than a NHPP. The dependence between periods is less of a concern with regard to pre-setting staffing levels, unless we plan to exploit that dependence, perhaps by adjusting staffing levels partway through the day to account for a different-than-expected customer volume.

Continuing with our modeling framework, our next goal is to try to capture the physical process that gives rise to the observed data. The Palm-Khintchine theorem and its extensions are so compelling that it is hard to discard the idea that the arrival process is, on any given day, a NHPP. There are at least three possible explanations for the observed data that are consistent with this view:

1. The arrival process could be nonstationary on the time scale of weeks, during which data was collected. For example, suppose that on any given day the arrival process is indeed a NHPP, but that the overall arrival rate is increasing from week to week. Then the N_{ij}s are all Poisson distributed, but the mean is increasing with i. In the estimation process above we envisaged them having a constant mean. This would lead to over-dispersion. It would also lead to observed correlation in the number of arrivals in nearby periods.
2. The arrival process could be a doubly-stochastic Poisson process. In other words, it is possible that on any given day the arrival rate function for that day is selected randomly and then, conditional on the selected arrival rate function, the day's arrivals occur according to a NHPP with that arrival rate function.
3. Some combination of the above models is possible and plausible. (We ignore this complex possibility below.)

Should we adopt and fit one of these models? The answer to that question depends on what is known about the true system, and what the simulation model is designed to achieve. Discussions with call center staff and managers, as well as analysis of the data, can help to determine whether there are trends in the call volumes or not, thereby clarifying the correct situation. If that does not suggest a strong candidate, then life is more complicated. We assume that the simulation model is used to determine minimum

staffing levels that we then attempt to cover by selecting staff shifts. The shifts will remain in place for k weeks, so that staffing levels remain constant for k weeks.

If k is large, then Model 2 may be a good one to use, even if Model 1 is the correct one. To see why, suppose that the expected number of arrivals per week varies deterministically over the year, and that we have one year's worth of data. If we fit Model 2, then it will appear that there is randomness in the arrival rate distribution. The fitted distribution of the arrival rate will (hopefully) approximately match a histogram of the weekly true expected arrival rate. We will then choose a set of shifts, expecting that performance will vary from good to poor in a random fashion from week to week. In fact, it will vary from good to poor in a deterministic fashion, but over the k weeks will likely average out to the predicted overall performance.

If k is small, then it is more important which model we select. If the arrival rate is changing in a deterministic fashion, then fitting Model 1 is correct, while fitting Model 2 would lead us to believe that the arrival rate varies randomly, and we would most likely overstaff. If, instead, Model 2 is correct but Model 1 is fitted, then we would likely understaff.

The fitting process itself is not the focus of this article. For some approaches, see Whitt (1999), Chen and Henderson (2001) and Brown et al. (2005).

If one doesn't believe that the above models are appropriate, or if after fitting such models to the available data there is still an important discrepancy between the fitted model and the data, then one might abandon the "capture the physics" approach and instead turn to a family of distributions for which the joint distribution is well understood. This is the approach taken in Avramidis, Deslauriers, and L'Ecuyer (2004), where a family of arrival processes is obtained that matches the dependence between arrival counts in different periods better than the models suggested above. It has the advantage that the fit to the data is better, but the disadvantage that it is harder to explain the models to decision makers.

As a side note, the extremely high correlations between arrival counts seen in different periods of the day in Avramidis, Deslauriers, and L'Ecuyer (2004) show that the customer volume seen early in the day is an extremely good predictor of customer volumes later in the day. It seems important to attempt to identify why this is the case. For example, is it possible that special marketing promotions are held from time to time that dramatically influence the customer volumes seen at the call center, but the call center managers are unaware of those promotions? The point is that it may be possible to improve the prediction of call center volumes and thereby reduce the variability that we are trying to staff for, simply by improving the communication between different groups within an organization.

4 AMBULANCE DEPLOYMENT

The ambulance deployment problem is the problem of determining how many ambulances are needed, where to place them, and at what times, in order to ensure satisfactory response time performance. In this section we will focus on the modeling of travel times, but first we offer some (very) brief comments on the arrival process.

The discussion in the previous section on arrival processes for call centers is relevant here. In particular, the results in Cinlar (1972) suggest that the stochastic process describing the time *and location* of arriving calls is well modeled as a marked Poisson process, or equivalently, as a Poisson process in 3 dimensional space (1 dimension for time, and 2 for space). Resnick (1992) gives an accessible introduction to Poisson processes in more general spaces than the real line.

Let us return to the modeling of travel times. Throughout this section we assume for clarity that there are no delays before an ambulance is dispatched, such as turnout time (the time for crews to get their ambulance on the road).

Any simulation model of ambulance operations includes some level of modeling of travel times. This can be as simple as using an "average speed over straight-line distance" calculation, or as complicated as a road network with time-dependent travel speeds on the arcs and/or nodes. Approaches have been developed that fit between these extremes as well. With some exceptions, all models use deterministic travel speeds, even though we know that travel speeds are random, depending, for example, on weather conditions, traffic volumes and traffic-light phasing. Recent work explores the impact of random travel times (ODT Ltd. 2002, Budge 2004, Budge, Ingolfsson, and Erkut 2003). Does this make a tangible difference to the predictions? In other words, is it *worth* explicitly modeling random travel speeds? (This question is, of course, discussed in the work mentioned above. We merely wish to shed more light on the answer.)

Consider a stylized model where a single ambulance answers all calls immediately from its base, so that queueing effects are ignored. Call locations lie inside a circle of radius R kilometers surrounding the base. Let $F(x)$ be the distribution function of the straight-line distance from the base to a call. For example, if the calls are uniformly distributed in the circle, then $F(x) = x^2/R^2$. The ambulance travels in a straight line from the base to the calls. Our performance measure is the fraction of calls that the ambulance reaches in r minutes or less. This long-run measure is equal to the probability p that the ambulance reaches a new call within r minutes under very general ergodicity assumptions.

We compute p for both deterministic travel speeds (p_{det}) and random travel speeds (p_{ran}). Suppose that the ambulance travel speed is deterministic. To keep things simple we let this deterministic speed be 1 kilometer per minute.

So then the ambulance can travel exactly r kilometers in r minutes, and $p_{\text{det}} = F(r)$.

Now suppose that the ambulance travel speed is random and independent of the distance to the call. If X is the distance to the call and S is the travel speed (assumed positive with probability 1), then

$$\begin{aligned} p_{\text{ran}} &= P(X/S \le r) \\ &= EP(X \le rS|S) \\ &= EF(rS). \end{aligned}$$

Notice that if $S = 1$ with probability 1 then $p_{\text{ran}} = p_{\text{det}}$, as expected.

Suppose that $ES = 1$ so that the mean travel speed coincides with the deterministic travel speed, and that F is convex on the range of rS. Then Jensen's inequality shows that $p_{\text{ran}} \ge p_{\text{det}}$, i.e., random travel times *improve* the fraction of calls reached on time relative to deterministic travel times! To understand this perhaps-counterintuitive result suppose that F is continuously differentiable, so that F being convex is equivalent to its derivative f being increasing. The value $f(x)$ gives the density of call locations that are on a circle of radius x, which will typically be increasing in x due to the increasing circumference of circles of radius x. So random travel times "help" in great generality. But *how much* do they help? The answer to this question depends on the distribution of calls around the circle of radius r and the range of the random variable rS.

To get some sense of the difference, suppose that calls are uniformly distributed inside the circle of radius R, and that even at the maximum possible speed, the ambulance travels a distance that is at most R in r minutes. So S lies in the interval $[0, R/r]$ with probability 1. This ensures that there is no "truncation error" in the following calculation. Then

$$\begin{aligned} p_{\text{ran}} &= EF(rS) \\ &= E[r^2S^2/R^2] \\ &= \frac{r^2(\text{Var } S + 1)}{R^2}, \end{aligned}$$

where the last equality follows since $ES = 1$. So the difference in performance between using deterministic travel speeds and using random travel speeds increases with the variability of the travel times. The maximum variance of S that is consistent with the requirements that S lie in the interval $[0, R/r]$ and have mean 1 occurs when S is discrete and is equal to either 0 or R/r, in which case its variance is $R/r - 1$, and then $p_{\text{ran}} = r/R$ and $p_{\text{det}} = r^2/R^2$. If r^2/R^2 is on the order of 80%, so that the ambulance can reach 80% of calls if it travels at the deterministic rate, then r/R is approximately 89%, so that "true" on-time performance could be 89%. This is an enormous difference in performance.

This simple model shows that random travel times can have a significant effect on performance estimates. The model relied on 2 main assumptions. First, travel speed and travel distance are independent. However, it is generally believed that the farther you travel, the faster your average speed, due to the use of arterial roads where travel speeds are higher than in smaller city streets. It is hard to predict the effect of such dependence on the performance values above, and we do not consider that further here. Second, there are no queueing effects. This assumption is hard to brush aside, since ambulances can be heavily utilized. For example, Henderson and Mason (2004) state that in Auckland, New Zealand, ambulances can be utilized more than 50% of the time during peak periods. What is the combined effect of queueing and random travel speeds?

With queueing there is interaction between calls, so the dependence structure of travel times may be important. Some of the effects mentioned earlier impact all trips on a given day (weather, congestion), while others are trip-specific (traffic light phasing). To get some sense of the relative impacts of these different types of dependence we look at some simple queueing models.

Our first model is again of a single ambulance. We assume that calls arrive according to a Poisson process at constant rate λ. Service times for calls include travel to the scene, on-scene time, transport to hospital (possibly), hospital admittance time and time to return to base. Of these service-time components, some are not related to travel and others are. Accordingly, we model the service time as consisting of a non-travel component X, and a travel component Y.

We consider two cases for the travel-time component Y. In Case 1, $Y = D(Y_1 + Y_2)$. Here Y_1 is the time required to travel to the call when the ambulance travels at a "typical" speed. Similarly, Y_2 is the travel-time component *after* the ambulance reaches the scene. The multiplier D is common to all calls received on a given day and has mean 1. It represents the random effects that influence all calls received on a given day. In Case 2, $Y = C_1Y_1 + C_2Y_2$. Here, Y_1 and Y_2 have the same interpretations as before. There is no daily effect D in this model, but now the travel times are perturbed by C_1 and C_2, which are independently chosen for every call, are independent of one another, and have mean 1. Case 2 represents the extreme where there is only a call-specific effect and no effect that is common to all calls received on a day. We compute performance for these extreme cases, as well as the base case where there is no additional travel-time randomness beyond that embodied in Y_1 and Y_2, thereby getting some sense of the relative contributions of the two effects.

We model this system as an $M/G/1$ queue, where the service times S are given by $X + Y$. For Case 1

we approximate the long-run performance by computing the steady-state performance of the queue on a given day conditional on D, and then averaging the results over D. The Pollaczek-Khintchine formula (e.g., Wolff (1989), p. 386)),

$$\tilde{F}(s) = \frac{(1-\rho)s}{s - \lambda + \lambda\tilde{G}(s)},$$

gives the Laplace transform $\tilde{F}(\cdot)$ of the steady-state delay in queue (exclusive of service) in terms of the Laplace transform $\tilde{G}(\cdot)$ of the service-time distribution. The steady-state time required to reach a call is the sum of this delay and DY_1. So we obtain the Laplace transform of the steady-state response time (conditional on D) by multiplying $\tilde{F}(s)$ by the Laplace transform of DY_1, treating D as a constant. To obtain the probability that the response time is t minutes or less, we numerically invert the Laplace transform using an algorithm described in Abate, Choudhury, and Whitt (1999). (To obtain the cumulative distribution function, as opposed to the density, we first divide the Laplace transform by s.) We then uncondition by computing the expected value of the probabilities obtained by numerically inverting Laplace transforms.

For Case 2, the calculations are similar, except that there is no need to do any conditioning, since the service times are i.i.d.

The specific distributions we use are given in Table 1.

Table 1: Distributions for the $M/G/1$ Calculations

Variable	Distribution
X	gamma, mean 45, variance 225
Y_1	gamma, mean 7, variance 9
Y_2	gamma, mean 20, variance 100
D, C_1 and C_2	Discrete

We chose a common distribution for D, C_1 and C_2 for simplicity. It is a 3-point distribution where $P(D = 1) = p$, and $P(D = 1 \pm \Delta) = (1-p)/2$, for $\Delta \in (0, 1)$ and $p \in (0, 1]$. Notice that when $p = 1$, we obtain a model where $D = 1$ a.s., so that we recover results for deterministic travel speeds. The arrival rate was $1/400$ and we took $t = 12$, so we computed the probability that the response time was 12 minutes or less. The results are given in Table 2. The error due to the numerical inversion of the Laplace transforms was on the order of 10^{-8} or less and so can be ignored.

Table 2: The $M/G/1$ Results using a Gamma Distribution (3 sig. figs.)

Parameters	Case 1	Case 2
$p = 1, \Delta = 0$	0.832	0.832
$p = 0, \Delta = 0.25$	0.828	0.828
$p = 0, \Delta = 0.5$	0.816	0.817
$p = 0, \Delta = 0.75$	0.795	0.795
$p = 0, \Delta = 0.99$	0.767	0.766

First notice that our choice of arrival rate ensures that if we ignore the travel-speed variability ($p = 1, \Delta = 0$), then the ambulance reaches 83% of calls on time. This reasonable performance comes at the expense of low ambulance utilization of only 18%. Notice also that random travel speeds have almost no effect until $\Delta = 0.75$. This represents a huge variability in travel speeds, and is far from practical. It appears that randomness (in either case) has almost no practical effect!

To understand these results in relation to the previous ones, we reason as follows. The ambulance is only lightly utilized, so it is often idle when dispatched to a call. The main determinant of reaching a call on time is therefore Y_1. The gamma distribution used here has a density that is low near the cutoff point of 12 minutes, so the sensitivity to travel time randomness (and dependence) is low. If we instead use a triangular distribution for Y_1 as in our previous model, with $R = 10$ we obtain the results given in Table 3. The numerical errors are on the order of 10^{-4} or less.

Table 3: The $M/G/1$ Results using a Triangular Distribution (3 sig. figs.)

Parameters	Case 1	Case 2
$p = 1, \Delta = 0$	0.823	0.823
$p = 0, \Delta = 0.25$	0.800	0.799
$p = 0, \Delta = 0.5$	0.687	0.682
$p = 0, \Delta = 0.75$	0.622	0.612
$p = 0, \Delta = 0.99$	0.584	0.567

We now see a strong dependence on additional randomness, as expected. However, since the results for Cases 1 and 2 are so similar, it seems that there is no need to explicitly model the day-wide effect, and that i.i.d. random travel times are sufficient.

Let us check that this hypothesis still holds when we consider multiple ambulances, say c of them. The $M/G/c$ queue is difficult to analyze, so we instead use the $M/G/\infty$ model. This model was originally used in the emergency response context by Bell and Allen (1969). When an $M/G/c$ queue is lightly to moderately loaded, all c servers are busy only a small fraction of the time, and so the $M/G/\infty$ is an excellent approximation. For more heavily loaded systems its predictions are less accurate but still useful. The $M/G/\infty$ queue can be used to obtain approximations for the probability that a call is answered immediately in an $M/G/c$ queue by computing the steady-state probability that there are $c-1$ or fewer jobs present in the system.

The steady-state distribution of the number of jobs present in an $M/G/\infty$ queue is Poisson with mean λ/μ, where λ and μ are the arrival and service rates respectively. This distribution depends on the service-time distribution only through its mean. Hence, if we have a model where we only change the *variability* of the service time distribution but not the mean, then there will be no effect on our

approximation. So changes along the lines of Case 2 will have no effect. Of course, in the true system with a finite number of ambulances, there will still be sensitivity to the service time variability due to the effect that was described in our earlier model, namely having an appreciable fraction of calls concentrated at a distance from an ambulance base where speed variability plays a role.

So we restrict our use of the $M/G/\infty$ queue to Case 1, where there is a daily multiplier D that is the same for all calls. As with the $M/G/1$ analysis above, we condition on D, compute steady-state performance, and then uncondition to get our performance measure. The conditional mean of a service time given D is $EX + D(EY_1 + EY_2)$. We used the means for X, Y_1 and Y_2 given in Table 1, increased the arrival rate to 1, and used 80 ambulances. We used a normal distribution for D with mean 1 and different values for the variance, truncating the distribution to 3 standard deviations either side of the mean. We numerically computed the integral associated with unconditioning. For the underlying $M/G/c$ queue to be stable on all days with high probability, we need to ensure that

$$\lambda(EX + D(EY_1 + EY_2)) < c.$$

This gives an upper bound on the variance of D (here and henceforth we refer to the variance of the untruncated D, although the calculations are based on the truncated D), because the probability of instability increases with the variance of D.

For $\lambda = 1$ and $c = 80$, the fraction of time that a call is answered immediately ranges from 0.81 when Var $D = 0$ to 0.74 when Var $D = 0.1$. The effect is less pronounced when we take $\lambda = 0.22$ and $c = 20$. In this case the fractions range from 0.81 when Var $D = 0$ to 0.77 when Var $D = 0.22$. As c gets smaller (with a corresponding decrease in arrival rate) the effect becomes less and less pronounced. For example, with $\lambda = 0.04$ and $c = 5$, the range is from 0.84 to 0.82.

So with this model we see that the daily effect *can* be significant when the number of ambulances is large, but it becomes weaker as the number of ambulances gets smaller. To understand this result, note that for a given arrival rate we chose the number of ambulances so that in the deterministic travel speed case, the performance was reasonable. This corresponds with the usual method for planning, where randomness in travel speeds is ignored. The risk-pooling effect of many-server systems (Whitt 1992) ensures that for larger numbers of ambulances, one can utilize them more heavily and still have reasonable response times. When we then add travel-speed variability, this changes the daily *mean* of the service times, with a consequent change in the traffic intensity. When the number of ambulances is small and the servers are lightly-loaded, this change has only a very small impact on performance. But when there are larger numbers of ambulances that are therefore fairly heavily loaded, the fluctuation in the mean service time leads to larger daily fluctuations in performance.

So to summarize the main points in this extended ambulance example,

- Random travel speeds can have a big impact when a significant number of calls are close to the boundary between reachable and unreachable calls.
- Although our non-queueing example suggested that random travel speeds can improve performance, the reverse was true when we incorporated queueing effects. The reason is that, in general, variability exacerbates queueing effects. Furthermore, daily effects that are common to all calls essentially modify the service time mean (on each day), and queueing systems are sensitive to such changes, especially when the servers are heavily utilized.
- The breakdown of travel-speed randomness between daily variability and call-specific variability is not important, except when the system is heavily loaded. Ambulance utilizations are often on the order of 20 or 30%, so in such cases it does not appear to be important to determine the breakdown of travel-speed randomness. However, as mentioned earlier there are situations where ambulance utilization can exceed 50%, and in such cases the breakdown will play a more important role.

We offer one final thought to close this section. It is not conceptually difficult to create a simulation with a relatively complicated travel speed model. The difficulty really lies in fitting the parameters of the model to the real system. So one could just code up the complicated model and try the extremes of parameter settings, exactly as we did in this section for various queueing models. If the performance sensitivity is low, then there is little need to go to the trouble of fitting the complicated model.

5 YACHT MATCH RACING

The America's Cup is a coveted sporting trophy with a long history. Competitors design and build their own boats under certain design rules that define the International America's Cup Class (IACC). All races are match races, which are races between only 2 boats. The exact format of the event varies, but invariably includes several series of match races.

The design of IACC yachts and the development of tactics is a sophisticated business. Scale models, computational fluid dynamics, and race-modeling programs (RMPs) are now standard aspects of the process. An RMP simulates a race at some level, computing estimates of various performance measures, including the probability that one yacht design will beat another. The state of the art in RMPs is prob-

ably ACROBAT (Philpott, Henderson, and Teirney 2003). This is a full simulation of a match race, combining yacht dynamics, tactics, and a model of wind behaviour over the course of a race. ACROBAT was created to assist in yacht-design decisions, such as the tradeoff between upwind and downwind speed through decisions of boat width, length and sail area.

How should one model the dynamic behaviour of wind on two yachts that are moving over a race course for approximately 2.5 hours? The answer to this question, as always, depends on the questions one is trying to answer. In the case of Philpott, Henderson, and Teirney (2003), the key questions are yacht-design questions. In recent work (Sheild, Henderson, and Philpott 2005), the key question is one of tactics. It turns out that one may want to use quite different models of wind in these two cases, as we shall see below.

5.1 Yacht Design

Consider first the question of how to model wind when one is looking at the question of yacht design. Here it is important to capture time-varying wind strength, as it is possible to design "specialist" boats that are most effective in specific wind conditions like light air or heavy air, as well as generalist boats that perform well in a variety of conditions. Without a model of wind where wind strength fluctuates with time, simulations would probably favour specialist boats. The model also needs to capture the effect of different wind conditions at different parts of the course. Occasionally a yacht can "sail into a hole," where there is very little wind, while the other yacht enjoys a reasonable breeze. This happens rarely, and always when the boats are widely separated. When the boats are close together, they see virtually identical wind conditions.

It is natural to consider building a model of wind that gives values for the wind speed and direction over a grid of locations covering the course at discrete time points that span the duration of a race. Philpott, Henderson, and Teirney (2003) did not pursue this approach for 2 main reasons. First, the data available was not up to the task of fitting such a spatially complex model. Second, it seemed that such a model would be quite slow to execute. Since the wind conditions were only needed at 2 locations on the course, a different approach was taken.

Philpott, Henderson, and Teirney (2003) developed a model that was consistent with several physical observations, i.e., their model *captured the physics*. First, the *Taylor hypothesis* states that the wind field evolves as if it were a non-varying vector field that moves in the mean wind direction at the mean wind speed (Lawson 1980). Any "yachtie" will tell you that it is false, but that it is approximately true: Wind conditions experienced upwind are repeated to quite a high degree downwind some time later. The repeated conditions are not identical to the earlier ones, and therein lies the approximation. Second, the wind speed and wind direction were found to be approximately statistically independent processes, so they were modeled independently. This reduces the model to one where we generate a process of 2-vectors of wind speeds (one component for each boat), and a process of 2-vectors of wind directions independently. The wind speed was then modeled as a bivariate autoregressive process, while the wind direction was modeled as a bivariate Markov chain to capture a phenomenon where wind direction tends to persist for some time before abruptly switching to a new direction.

The final aspect of this model was to employ the Taylor hypothesis to help model the spatial behaviour of wind. The wind conditions on the upwind boat were generated first. The Taylor hypothesis then asserts that these wind observations will percolate in the direction of the mean wind direction, and therefore, towards the downwind boat. The conditions on the downwind boat were generated conditional on these observations.

This model has a number of shortcomings, not least of which is that the modeled spatial-temporal dependence structure of wind is somewhat unclear. However, it served its purpose of providing a believable wind model that captured enough physical characteristics of wind to assist in design decisions. In that sense, it was successful.

Thus far we have focused on wind behaviour *during* a race, that is, on a time scale of hours. But in recent America's Cup tournaments, the winner between two yachts is the first yacht to win 5 races. Since there are typically 4 race days per week, a series can take a long time to complete. Should we model the nonstationary behaviour of weather on this longer time scale in the simulation? The answer seems to be no. We can view the simulation model as a subroutine that gives, for a given set of wind conditions, the probability that one boat design beats another in a single race. These single-race probabilities can then be combined using some higher-level model of the series to determine the probability that one boat design beats another in a race series.

5.2 A Tactical Question

In recent work, Sheild, Henderson, and Philpott (2005) consider a question of tactics. America's Cup races are run over 6 legs that alternate between upwind and downwind directions. At the start of the first leg, yachts vie for a particular position on the starting line and a starting tack, i.e., a direction to head from the starting line. Let us call the choice of starting position and tack the "starting decision." The starting decision is very important because it is a major factor in determining which yacht will be ahead of the other when the yachts next come close together in the first leg, which in turn has a very strong influence on the outcome of the race. The starting decision is usually based

on wind observations made prior to the race by members of a syndicate that are on boats upwind of the course. These people communicate a recommendation to the racing yacht 5 minutes before the race, after which communication between the racing yacht and elsewhere is no longer allowed. The goal of this work is to help with the starting decision by conditioning on the known weather conditions prior to the start of the race.

The nature of the data available to the model has changed since the work described in Philpott, Henderson, and Teirney (2003). We now have a time series of wind speed and direction on a coarse grid of locations over the course. The improved data availability and the different nature of our goal now ensures that we should model a time series of wind conditions over the entire course (and perhaps for some area upwind of the course), due to the need to condition on starting information.

A model that captures wind speed and direction, or equivalently the x and y components of wind speed, on a grid of locations on the course will have a large state space. This suggests that a general Markov chain model as used previously will require an enormous data set to fit the transition probabilities, unless we impose substantial additional structure. Therefore, we abandon the type of model used in Philpott, Henderson, and Teirney (2003), which was very physically motivated, and instead look for a well-understood stochastic process that has the features we desire.

The desire for tractable conditional distributions suggests that some sort of Gaussian model might be appropriate. Since wind speeds are nonnegative we model the log of the wind speeds as Gaussian. Our goal is a process $(X(t, x, y) : t \geq 0, (x, y) \in \Re^2)$, where $X(t, x, y) \in \Re^2$ gives the log-components of wind velocity at the point (x, y) at time t.

The specific model that Sheild, Henderson, and Philpott (2005) propose uses a vector autoregressive Gaussian process for the wind velocities on a grid, with additional variables for the wind velocities at the yacht locations. The structure of the coefficient matrix in the autoregression is tailored to the Taylor hypothesis, in that downwind conditions depend on upwind conditions at earlier time steps, thereby linking the spatial and temporal dependencies.

The model is, in essence, a simplified version of a Gaussian process in space and time. It is simplified because it is not easy (computationally speaking) to fit and generate Gaussian processes in general. A key difficulty lies in the problem that one must condition on all values that have been generated thus far when generating a new value. From the usual Cholesky decomposition approach to generating Gaussian random vectors, one can see that conditioning on all previous values can lead to an increasing computational load as the simulation progresses. So there is a tradeoff between computational tractability and faithfulness to the true Gaussian joint distribution. More research is needed to develop flexible classes of Gaussian models that are both easily fitted to data, and from which random quantities can be generated rapidly.

The details of the model beyond the sketch above are not important for our discussion. What *is* important is the process by which we arrived at this point. The model choice was determined by the goals of the study and data availability. The model has limitations: It is unlikely that it can capture the persistence of wind directions that was well-modeled in Philpott, Henderson, and Teirney (2003). This sacrifice gains us a great deal: The conditional distributions are straightforward to compute, as are forecasts of future wind conditions given present information.

6 REFLECTION

In this paper we have described a modeling philosophy, and applied it to a variety of examples. Some of the key themes, along with some additional thoughts, are as follows.

- *Capturing the physics* of the system is probably the most reliable way to ensure that complex dependence relationships are captured accurately. This often leads to a reduction of the dimensionality of the parameters of a model that must be fitted, thereby simplifying the calibration phase.
- Analysis of simple models can inform decisions about the required level of model detail.
- Gaussian and Poisson random fields have an important role to play in simulation models of spatial phenomena, due to their relative tractability and physical interpretations. Models that are both easily fitted to data and efficiently generated are in short supply.
- We must often attempt to calibrate models with very little (relatively speaking) data. This occurs because we are trying to model a multidimensional random vector (or even a full time series) rather than a univariate random vector. The curse of dimensionality is the key problem. This difficulty with calibration suggests that methods for addressing input uncertainty will play an important role in simulations involving random vectors with complicated joint distributions.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation Grant DMI-0400287. Much of this work was completed while I was visiting the Department of Engineering Science at the University of Auckland. I would like to thank the department for their hospitality and support.

REFERENCES

Abate, J., G. Choudhury, and W. Whitt. 1999. An introduction to numerical transform inversion and its application to probability models. In *Computational Probability*, ed. W. Grassman, 257–323. Boston: Kluwer.

Avramidis, A. N., A. Deslauriers, and P. L'Ecuyer. 2004. Modeling daily arrivals to a telephone call center. *Management Science* 50 (7): 896–908.

Bell, C. E., and D. Allen. 1969. Optimal planning of an emergency ambulance service. *Journal of Socio-Economic Planning Science* 3:95–101.

Biller, B., and S. Ghosh. 2004. Dependence modeling for stochastic simulation. In *Proceedings of the 2004 Winter Simulation Conference*, ed. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 153–161. Piscataway NJ: IEEE.

Biller, B., and S. Ghosh. 2006. Multivariate Input Processes. In *Simulation*. Handbooks in Operations Research and Management Science. Elsevier. To appear.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association* 100:36–50.

Budge, S. 2004. *Emergency Medical Service Systems: Modelling Uncertainty in Response Time*. Ph. D. thesis, Department of Finance and Management Science, University of Alberta, Edmonton, Alberta, Canada.

Budge, S., A. Ingolfsson, and E. Erkut. 2003. Optimal ambulance location with random delays and travel times. Submitted for publication.

Chen, B. P. K., and S. G. Henderson. 2001. Two issues in setting call center staffing levels. *Annals of Operations Research* 108:175–192.

Cinlar, E. 1972. Superposition of point processes. In *Stochastic Point Processes: Statistical Analysis, Theory, and Applications*, ed. P. A. W. Lewis, 549–606. New York: Wiley Interscience.

Edie, L. C. 1954. Traffic delays at toll booths. *Journal of the Operations Research Society of America* 2:107–138.

Henderson, S. G. 2003. Estimation for nonhomogeneous Poisson processes from aggregated data. *Operations Research Letters* 31:375–382.

Henderson, S. G., and A. J. Mason. 2004. Ambulance service planning: simulation and data visualization. In *Operations Research and Health Care: A Handbook of Methods and Applications*, ed. M. L. Brandeau, F. Sainfort, and W. P. Pierskalla, 77–102. Boston: Kluwer Academic.

Hill, R. R., and C. H. Reilly. 2000. The effects of coefficient correlation structure in two-dimensional knapsack problems on solution procedure performance. *Management Science* 46 (2): 302–317.

Jongbloed, G., and G. Koole. 2001. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry* 17:307–318.

Law, A. M., and W. D. Kelton. 2000. *Simulation Modeling and Analysis*. 3rd ed. New York: McGraw-Hill.

Lawson, T. V. 1980. *Wind Effects on Buildings: Statistics and Meteorology*, Volume 2. London: Applied Science Publishers.

ODT Ltd. 2002. Transport model tuning. Working paper 2002A, Optimal Decision Technologies Ltd., Auckland, New Zealand.

Philpott, A. B., S. G. Henderson, and D. Teirney. 2003. A simulation model for predicting yacht match race outcomes. *Operations Research* 52:1–16.

Pritsker, A. A. B. 1998. Organ transplantation allocation policy analysis. *OR/MS Today* 25 (4).

Resnick, S. I. 1992. *Adventures in Stochastic Processes*. Boston: Birkhäuser.

Sheild, H., S. G. Henderson, and A. B. Philpott. 2005. A spatial-temporal wind model for yacht match race simulation. Working paper.

Steckley, S. G., S. G. Henderson, and V. Mehrotra. 2005. Performance measures for service systems with a random arrival rate. In *Proceedings of the 2005 Winter Simulation Conference*, ed. M. E. Kuhl, N. Steiger, F. B. Armstrong, and J. A. Joines. Piscataway NJ: IEEE. To appear.

Whitt, W. 1992. Understanding the efficiency of multi-server service systems. *Management Science* 38:708–723.

Whitt, W. 1999. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters* 24:205–212.

Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs NJ: Prentice Hall.

AUTHOR BIOGRAPHY

SHANE G. HENDERSON is an associate professor in the School of Operations Research and Industrial Engineering at Cornell University. He has previously held positions in the Department of Industrial and Operations Engineering at the University of Michigan and the Department of Engineering Science at the University of Auckland. He is an associate editor for the *ACM Transactions on Modeling and Computer Simulation*, *Operations Research Letters*, and *Mathematics of Operations Research*, and the secretary of the INFORMS Simulation Society. He likes cats but is allergic to them. His research interests include discrete-event simulation and simulation optimization. His e-mail address is <sgh9@cornell.edu>, and his web page is <www.orie.cornell.edu/~shane>.