BULL. AUSTRAL. MATH. SOC.
VOL. 26 (1982), 143-155.

60J85, 62M05

STOCHASTIC MODELLING IN CSIRO: TWO EXAMPLES*

I.W. SAUNDERS

Communicated by James M. Hill

Two examples of stochastic models arising from CSIRO consulting work are described. The first concerns the interaction of different strains of myxomatosis in a rabbit population. The effect of different infection rates in a simple epidemic model is investigated and the implications for rabbit control discussed. The second example is concerned with sampling iron ore from a moving belt. A stochastic model for flow rate and quality of ore is constructed and used to compare different sampling schemes.

1. Introduction

The Division of Mathematics and Statistics of CSIRO (Commonwealth Scientific and Industrial Research Organization of Australia) has a twofold role: to provide mathematical and statistical support to other scientists within the Organisation and also to carry out theoretical research aimed at widening the range of statistical and mathematical tools available. The wide variety of interests of the scientists in CSIRO implies that a wide variety of methods, mathematical, statistical or otherwise, will be required in analysing their data. The methods used depend on the nature of the problem. Precise measurements on well understood physical systems can

Received 16 March 1982.

* This paper is based on an invited lecture given at the Australian Mathematical Society Applied Mathematics Conference held in Bundanoon, February 7-11, 1982. Other papers delivered at this Conference appear in Volumes 25 and 26.

https://doi.org/10.1017/S0004972700005645 Published online by Cambridge University Press

often be explained using what might be called "applied mathematical" models - perhaps differential or integral equations - with any randomness effectively ignored. On the other hand, when living things or natural phenomena such as weather are involved, the random component will become more significant. Then the design of experiments to accommodate such random variation becomes important and analysis of variance techniques will often be appropriate. A third class of problems consists of those where the available data are so imprecise or incomplete that straightforward application of statistical methods is impossible or unproductive.

It is the final class that I shall be concerned with here. When detailed data are not available, some form of modelling is essential to an understanding of the situation. When randomness is likely to be a predominant feature, this must be incorporated. Thus we are led to consider stochastic models. I shall describe two problems that have arisen from my consulting work in which, for different reasons, complete data could not be obtained. In each case, we shall see that stochastic modelling enables us to derive useful information despite the lack of data.

2. Competition between strains of myxomatosis

A project for which CSIRO is well known is the use of myxomatosis for the control of rabbit populations. The control programme has had considerable success in reducing the vast population of rabbits and preserving pasture for livestock. In recent years however the use of myxomatosis has become less effective for a number of reasons. Among these is the appearance, through mutation, of new and less virulent strains of the disease. These are endemic in a number of regions. A rabbit infected by such an attenuated strain is likely to survive. If he does he will have acquired immunity, not only to reinfection by the attenuated strain, but also to subsequent infection by more virulent strains. Thus if a virulent strain is introduced, the population of rabbits susceptible to it will have been much reduced. This will slow the spread of the control strain since the immune rabbits can neither contract the disease themselves, nor pass it on to other susceptible rabbits.

In populations where an attenuated strain is endemic, little can be done. However such strains can also be introduced by natural means into previously susceptible populations. If this occurs in a population into

https://doi.org/10.1017/S0004972700005645 Published online by Cambridge University Press

which a virulent strain is to be introduced as a control measure, the two strains will effectively be competing for the susceptible rabbits. The question of interest is thus what features of the transmission processes of the two strains will determine the outcome of the competition. An answer to this will enable us to assess the usefulness of a particular strain of myxomatosis as a means of rabbit control.

Clearly there are many problems to be solved here. We need to consider the mechanism of disease transmission, the nature of the interaction between the two strains, the relative infectivities of the strains and the effectiveness of the method of introduction of the virulent strain. Here we shall concentrate mainly on the effect of relative infectivity and consider a much simplified model which, while it does not include the full complexity of the real situation, allows us to draw useful conclusions.

Since we shall be considering only the process of infection, we shall use an extension of the so-called simple epidemic model (*cf.* Bailey [1]) in which individuals, once infected, remain infectious forever. Our results therefore will apply to the initial stages of an outbreak, before deaths and recoveries have reached a significant level. We shall also suppose that the time scale of the outbreak is short enough that we can neglect births and deaths of susceptibles and assume a constant population size.

We suppose therefore that we have a population of n individuals and that at time t, $I_1(t)$ of them are infectious with disease type 1 and $I_2(t)$ of them with disease type 2, leaving $n - I_1(t) - I_2(t)$ susceptibles. The rate of spread of a disease will depend on the rate of occurrence of "infective contacts" between infectives and susceptibles and if the members of the population mix uniformly this rate will be proportional to the product of the numbers of infectives and susceptibles. We therefore assume that the rate of type k infections is $a_kI_k(t)S(t)$ for some a_k, where $S(t)$ denotes the number of susceptibles.

When a type k infection occurs, $S(t)$ is reduced by one and $I_k(t)$ is increased by one. We can thus define a Markov process modelling this situation through its infinitesimal transition probabilities:

$$\begin{aligned}\text{Prob}\big(I_1(t+dt) = j_1+1,\ I_2(t+dt) = j_2 \mid I_1(t) = j_1,\ I_2(t) = j_2\big) \\ = a_1I_1(t)S(t)dt + o(dt)\ ,\end{aligned}$$

$$\begin{aligned}\text{Prob}\big(I_1(t+dt) = j_1,\ I_2(t+dt) = j_2+1 \mid I_1(t) = j_1,\ I_2(t) = j_2\big) \\ = a_2I_2(t)S(t)dt + o(dt)\ ,\end{aligned}$$

$$\text{Prob(two or more infections in } t,\ t+dt) = o(dt)\ .$$

This process is rather intractible as it stands as a result of the nonlinear transition rates. We therefore seek a method of converting it into a linear process. Note that, since we are principally interested in the *relative* behaviour of the two epidemics, the time scale used is of secondary importance. A random change of timescale will therefore leave the conclusions unaltered. Define

$$(2.1)\qquad \begin{aligned} Z(t) &= \int_0^t S(u)du\ , \\ T &= \sup\big(t : S(t) > 0\big)\ , \end{aligned}$$

so that T is the duration of the process.

Since $S(t)$ is positive for $t < T$, we see that $Z(t)$ is increasing on $(0, T)$ and so has an inverse $Z^{-1}(t)$ defined on $\big[0, Z(T)\big)$. Note also that

$$\frac{dZ^{-1}(z)}{dz} = \big\{S\big(Z^{-1}(z)\big)\big\}\ .$$

Define a new process on the Z timescale by

$$(2.2)\qquad J_k(z) = I_k\big(Z^{-1}(z)\big) \quad \text{for} \quad z < Z(T)\ .$$

Then the infinitesimal transition probabilities of the J process are given by

$$\begin{aligned}\text{Prob}\big(J_k(z+dz) = J_k(z)+1 \mid J_k(z)\big) &= a_kJ_k(z)S\big(Z^{-1}(z)\big)\frac{dZ^{-1}}{dz}\,dz + o(dz) \\ &= a_kJ_k(z)dz + o(dz)\ .\end{aligned}$$

These are the transitions of independent linear birth processes and so the relative behaviour of the two diseases is the same as that of two such processes stopped at $z = Z(T)$ that is, when $J_1(z) + J_2(z) = n$. The properties of birth processes are well known, and those of the competing processes can now be deduced.

For a large population size n, we can use asymptotic results. A standard result is that, as $z \to \infty$,

$$J_k(z)e^{-a_k z} \to J_k(0)W_k \text{ with probability one,} \tag{2.3}$$

where W_k is a random variable with an Erlangian distribution, that is, the sum of independent exponential random variables. The expected value of W_k is 1 and so we see that the expected number of infectives satisfy

$$\left(EI_1(t)^{1/a_1}\right) / \left(EI_2(t)^{1/a_2}\right) = \left(I_1(0)^{1/a_1}\right) / \left(I_2(0)^{1/a_2}\right) \tag{2.4}$$

where the right hand side is independent of t.

In fact from (2.3) we have

$$\left(J_1(z)^{1/a_1}\right) / \left(J_2(z)^{1/a_2}\right) \to \left(J_1(0)^{1/a_1} W_1^{1/a_1}\right) / \left(J_2(0)^{1/a_2} W_2^{1/a_2}\right) , \tag{2.5}$$

with probability one, and in large populations, approximate equality will hold. The random variable $W_1^{1/a_1}/W_2^{1/a_2}$ can be shown to have median close to 1, provided a_1 and a_2 are not too different. Thus the relative behaviour of the means in (2.4) does reflect the typical behaviour of the processes. Clearly then the disease with the larger infection parameter is at a considerable advantage. For example, if $a_1 = 2a_2$ and initially 1% of the population was infected with disease 1, then 7% would need to be infected with type 2 to give ultimate equality in expectation. In the context of rabbit control it is clear that a strain to be used for control will need to have a high infectivity. Otherwise it will be necessary to infect a large proportion of the population artificially in order to achieve a useful result.

It may be in some situations that the infection rate is similar for the two strains. Then it will be of interest to consider how competing epidemics with equal infection rates behave. The asymptotic conclusion

https://doi.org/10.1017/S0004972700005645 Published online by Cambridge University Press

follows from the above results: the ratio of numbers of infectives satisfies

$$\frac{I_1(t)}{I_1(t)+I_2(t)} \to \frac{I_2(0)W_2}{I_1(0)W_1+I_2(0)W_2}$$

with probability one. In this case however we can go further and identify the limiting random variable. $I_1(0)W_1$ and $I_2(0)W_2$ are independent gamma random variables with common scale parameter 1. Thus the ratio has a beta distribution with parameters $I_1(0)$ and $I_2(0)$ and so when the infection rates are equal the proportion of the infectives that are of type 1 has an approximate beta distribution. We can in fact obtain an expression for the exact distribution of the numbers of infectives. Let $I_k(m)$ denote the number of type k infectives when the total number of infectives is m. Then $\left(I_1(m), I_2(m)\right)$ forms a bivariate Markov chain. This is true for arbitrary values of a_1 and a_2 but the situation is particularly simple in the case of equal infection rates, for then the Markov chain is a simple Polya urn scheme (see, for example, [2]). The distribution of $I_k(m)$ is therefore

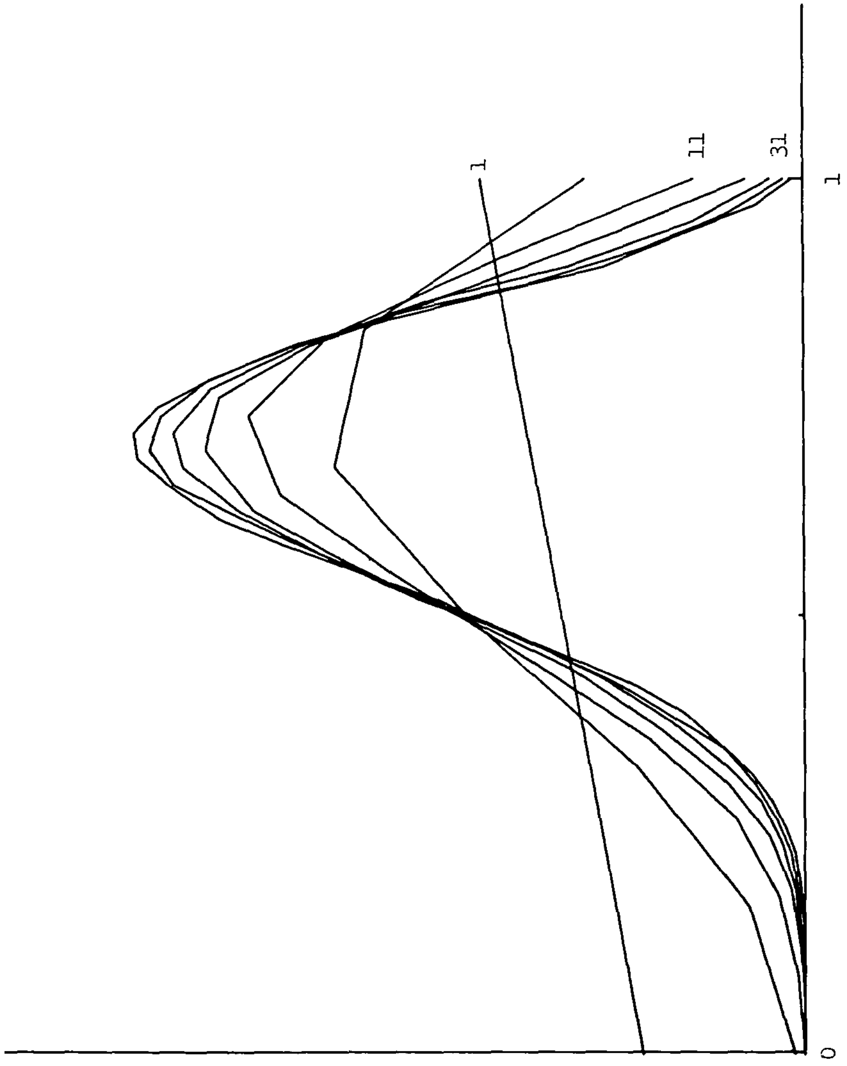
$$\text{(2.6)} \qquad \operatorname{Prob}\left(I_1(m) = I_1(0)+j\right) = \binom{m}{j} \frac{I_1(0)^{[j]}I_2(0)^{[m-j]}}{\left(I_1(0)+I_2(0)\right)^{[m]}} ,$$

where $i^{[j]}$ denotes $i(i+1) \ldots (i+j)$. This is a Polya-Eggenberger distribution and the limiting beta distribution can be deduced directly from (2.5). The speed of the convergence to the limiting distribution can be seen in Figure 1 where the distributions of $\left(I_1(m)-I_1(0)\right)/m$ are compared for $m = 1, 6, 11, \ldots, 31$ when $I_1(0) = 10$, $I_2(0) = 5$.

The most interesting feature of this result is that the distribution of the ultimate proportion of a particular type is concentrated around the proportion of that type among the initial infectives. The early introduction of the control strain will thus be at a premium.

We see that a fairly simple stochastic model has led to conclusions that may be useful in guiding both the use of myxomatosis as a control and

https://doi.org/10.1017/S0004972700005645 Published online by Cambridge University Press



Distribution of $(I_1(m)-I_1(0))/m$, $I_1(0) = 10$, $I_2(0) = 5$, $m = 1, 6, 11, 16, 21, 26, 31$

the directions of future study in this field.

Further mathematical detail may be found in [3] and [5] and a more detailed model of the transmission of myxomatosis is described in [4].

3. Sampling schemes for on-stream iron ore analysis

The CSIRO Division of Mineral Physics has recently been involved in the development of an on-stream analyser to measure the quality of iron ore sampled from a moving conveyor belt. This analyser can produce immediate estimates of average quality, which gives it an advantage over slower standard methods. If, for example, the ore being transported should prove to be of a lower quality than is required, immediate corrective action can be taken, thus avoiding penalties for failing to meet an agreed level of quality.

The method of analysis used by the new analyser requires that material be supplied to it at a constant rate, independent of the flow rate of the mail belt. This differs from standard sampling schemes which produce samples at a rate proportional to the flow rate of the belt. Thus it is necessary to investigate how the results obtained from the analyser will match up, in terms of accuracy and efficiency, with those from standard methods.

Once again, little detailed information is available; commercial ore transporters are not designed to provide the data that would be required for a complete study. Instead we must construct a model reflecting the important features of the process and use it to draw general conclusions.

We can get some idea of the sort of results we are likely to obtain by analogy with survey sampling from a stratified population. Suppose that population is divided into K "strata" and individuals in stratum i have an average value m_i for some quantity y which is of interest. Then the population mean of y will be $\sum N_i m_i / N$ where N_i is the population of stratum i and N is the total population size. Suppose a sample of the population is taken, consisting of n_i individuals from stratum i, $i = 1, \ldots, K$. Let y_i denote the sample mean of y for the individuals from stratum i. An unbiased estimate of the population mean will be

$\sum N_i y_i / N$. The sample mean is $\sum n_i y_i / n$ which will equal this unbiased estimate provided $n_i \propto N_i$. If the sample proportions in each stratum are not equal to the population proportions, then a weighted mean must be used. It can be shown that taking sample proportions equal to the population proportions also gives the most efficient estimator in terms of minimising the variance for a given sample size.

For on-stream sampling the quantity of interest is the quality, or percentage metal content, of the ore. This may be correlated with the flow rate and so its mean may vary as the flow rate changes. By analogy with the sampling scheme described above, we can consider the ore to consist of "strata" corresponding to particular flow rates. The "population size" of a stratum will be the mass of ore transported at the corresponding flow rate and the "sample size" will be the total mass of samples taken at that flow rate. Thus the above result suggests that the sampling scheme giving greatest efficiency will be the standard practice of taking samples at a rate proportional to the rate of flow of ore. Conversely, this suggests that constant rate sampling will be relatively inefficient. In order to discover how far these results do carry over to the on-stream sampling problem, and how serious the inefficiency is likely to be, we set up a model for the quality and flow rate of the ore and use it to compare the different sampling schemes.

We first define some notation:

(i) $w(t)$: the belt flow rate at time Δt ;

(ii) $g(t)$: the quality of the ore passing the sampling point at time t ;

(iii) T : the total duration of the process;

(iv) $t_1, t_2, \ldots, t_n$: the sampling times;

(v) Δt_i : $t_{i+1} - t_i$, the length of the ith sampling interval;

(vi) W_i : the mass of ore passing the sampling point during the ith sampling interval;

(vii) M_i : the mass of iron contained in W_i .

https://doi.org/10.1017/S0004972700005645 Published online by Cambridge University Press

We shall compare two different sampling schemes.

I. Time based sampling: samples are taken at fixed time intervals and the mass of a sample is proportional to the current belt flow rate. This gives a sampling rate proportional to the flow rate.

II. Single increment constant rate sampling: the samples are stored in a hopper from which ore is taken at a constant rate to be analysed. A new sample is taken whenever the level in the hopper reaches a "low" mark. The mass of a sample is again proportional to the current flow rate on the belt. The length of the sampling interval is thus $\Delta t_i = Cw(t_i)$ for some constant C .

Since we know little about the actual behaviour of the quality $g(t)$ and the flow rate $w(t)$ we endeavour to keep our assumptions simple and reasonably realistic. They are chosen to agree with the small amount of data that is available. We shall suppose that $g(t)$ varies sufficiently smoothly that its general trend during a single sampling interval can be adequately approximated by a straight line while its short term variations can be considered completely random. This leads us to model $g(t)$ by a Wiener process (or Brownian motion) within each sampling interval. For $w(t)$ we shall assume only that its variations are slow on the time scale of the sampling intervals. We take terms such as $w'(t)\Delta t$ to be negligible compared with $w(t)$. We assume further that the values of $g(t_i)$, $w(t_i)$ and $W(t_i)$ are all known exactly. Sampling variation will of course occur in practice, but will affect both of the sampling schemes.

Our interest is really centred on the total mass of iron $\sum M_i$ in the ore transported. The mean quality $\sum M_i/\sum W_i$ can be immediately found from this, and since $\sum W_i$ is the same for both sampling schemes, we shall compare the schemes on their estimates of $\sum M_i$.

Define

$$\hat{M}_i = -\left(g(t_i)w(t_i)+g(t_{i+1})w(t_{i+1})\right)\Delta t_i .$$

This is the natural estimate of M_i using the trapezium rule approximation to the integral of the flow rate of iron $g(t)w(t)$. It is not difficult

https://doi.org/10.1017/S0004972700005645 Published online by Cambridge University Press

to show that, conditional on the observed values of g, w and W, $\hat{M}_i - M_i$ has a normal distribution with mean approximately zero and variance

$$\text{var}\left(\hat{M}_i - M_i\right) = \sigma^2 V_i$$

where σ^2 is a constant and

$$V_i = W_i^2 \Delta t_i \; .$$

Since σ^2 does not depend on the sampling scheme, our comparison will be based on $V = \sum V_i$.

In order to compare the values of V resulting from the two schemes, we exploit the slow variation we have assumed for w and approximate V by an integral. Let V_1 and V_2 denote the values of V for scheme I and scheme II respectively. Then

$$\begin{aligned} V_1 &= \sum W_i^2 \Delta t \\ &\simeq (\Delta t)^2 \sum w\left(t_i\right)^2 \Delta t \\ &\simeq (\Delta t)^2 \int_0^T w(u)^2 du \; , \end{aligned}$$

while

$$\begin{aligned} V &= \sum W_i^2 \Delta t_i \\ &\simeq \sum w\left(t_i\right)^2 \Delta t_i^3 \\ &= c^2 \sum w\left(t_i\right)^4 \Delta t_i \\ &\simeq c^2 \int_0^T w(u)^4 du \; . \end{aligned}$$

To obtain a fair comparison, we must ensure that the two schemes are taking the same number of samples. Again we use an integral approximation. Let n_1 and n_2 denote the number of samples taken using scheme I and scheme II respectively.

https://doi.org/10.1017/S0004972700005645 Published online by Cambridge University Press

Then

$$n_1 = T/\Delta t$$

and

$$n_2 = \sum_1^{n_2} \Delta t_i/Cw(t_i)$$

$$\simeq C^{-1} \int_0^T w(u)^{-1} du \ .$$

Thus if $n_1 = n_2$ then we have

$$\int_0^T \left\{\frac{1}{w(u)} - \frac{C}{\Delta t}\right\} du = 0 \ .$$

The difference in the V values of the two schemes is

$$V_1 - V_2 = (\Delta t)^2 \int_0^T w(u)^4 \left(w(u)^{-1} - (C/\Delta t)\right)\left(w(u)^{-1} + (C/\Delta t)\right) du \ .$$

The integrand is positive when $w(u) < \Delta t/C$ and negative otherwise. Hence, after a little manipulation, we see that

$$V_1 - V_2 \le \int_0^T 2(\Delta t)^5 \left(w(u)^{-1} - (C/\Delta t)\right) du/C^2$$

$$= 0$$

since $n_1 = n_2$.

Equality can hold only if $w(u) = C\Delta t$ for all u in which case the two schemes coincide.

Thus we see that, unless the flow rate of ore is constant, scheme II is less efficient than scheme I. Without more detailed assumptions on the flow rate w we cannot determine the amount of the loss of efficiency. Saunders and Lwin [6] describe a model in which w and g are specified more completely. They find that, for realistic values of the parameters, approximately 10% more samples must be taken by scheme II than by scheme I to attain the same accuracy.

https://doi.org/10.1017/S0004972700005645 Published online by Cambridge University Press

4. Conclusions

We have looked briefly at two examples where stochastic modelling has enabled us to obtain useful conclusions in situations where, for one reason or another, complete data were unavailable. My aim in presenting this paper has been treefold:

1. to describe some of the work in stochastic modelling that is going on in CSIRO;
2. to show how useful answers to rather vague questions can be obtained; and finally
3. to show that interesting mathematics can sometimes arise from practical problems.

References

[1] Norman J.F. Bailey, *The mathematical theory of infectious diseases and its applications*, second edition (Griffin, London; Hafner [Macmillan], New York; 1975).

[2] Norman L. Johnson, Samuel Kotz, *Urn models and their application. An approach to modern discrete probability theory* (John Wiley & Sons, New York, London, Sydney, 1977).

[3] W. Kendall and I.W. Saunders, "Epidemics in competition II: the general epidemic", submitted.

[4] I.W. Saunders, "A model for myxomatosis", *Math. Biosci.* 48 (1980), 1-15.

[5] I.W. Saunders, "Epidemics in competition", *J. Math. Biol.* 11 (1981), 311-318.

[6] I.W. Saunders and T. Lwin, "Sampling schemes for quality estimation of a flowing material", *Internat. J. Mineral Process.* (to appear).

Division of Mathematics and Statistics,
CSIRO,
PO Box 310,
South Melbourne, Victoria 3205, Australia.

https://doi.org/10.1017/S0004972700005645 Published online by Cambridge University Press