

# CALIBRATION OF A MONTE CARLO SIMULATION MODEL OF DISEASE SPREAD IN SLAUGHTER PIG UNITS

*<http://www.sp.dk/~ej/efita.zip>*

**Erik Jørgensen**

*Department of Biometry and Informatics, Research Centre Foulum,  
Danish Institute of Agricultural Sciences,  
PO Box 39, DK-8830 Tjele  
e-mail: [ej@dina.sp.dk](mailto:ej@dina.sp.dk) <http://www.sp.dk/~ej>*

Abstract: The use of new resampling methods to improve the handling of stochastic simulation models is demonstrated. A Monte Carlo simulation model of disease spread within a slaughter pig herd is used as example. The model parameters reflect the disease spread and comprise, for example infection risk given diseases, and the positioning of the animals. The setting of the prior distribution of the parameters using expert knowledge is complicated, because the expert knowledge is generally based on the resulting dynamics rather than the underlying parameters. The paper shows how the prior distribution of model parameters can be made consistent with the knowledge concerning model output, using methods such as importance sampling and Markov Chain Monte Carlo techniques. Based on these methods, different management strategies are compared.

## **1 Introduction**

Mathematical models of animal production systems, such as Monte Carlo simulation models, can be seen as representations of expert knowledge of the systems. This knowledge comprises model output, model parameters and finally the structure and equations in the model. With the model, a diverse set of output variables can be predicted as a consequence of a set of decision rules. When these models are used for decision purposes, the use is often based on maximisation of expected utility or some related measure, such as expected income. Within this framework it is natural to consider the expert knowledge as a representation of the expert's subjective probability distribution. This concept can be generalized to any function of model output, and therefore to any use of the models. Consequently, the whole range of methods within Bayesian statistics becomes available, for example model testing, updating and calibration of model parameters to new observations and estimation of the uncertainty in the model predictions. Until recently, this has not been possible due to the complexity of the calculations. However, advances within Bayesian Statistics, and improvements in calculation speed have removed at least some of these obstacles, see e.g. Givens (1993). Such methods will be applied in the following to a model of disease spread in slaughter pig production units.

## **2 The simulation model**

The model used in the present context is a Monte Carlo simulation model of an animal production herd and can simulate the effect of different production strategies. In addition, it can handle disease spread within the herd. The model follows a long tradition within research in animal production systems, early examples are Singh (1986) and de Roo (1987). The present model was originally developed to study the information flow within pig production units, and the impact of different information system and production management as described in Jørgensen & Kristensen (1995). The model of the disease spreading is an add-on to this simulation model, and utilizes its basic elements. The addition of the disease spreading has been made by defining an additional object class, called a generic disease model.

## 2.1 *The generic disease model*

The simulation model consists of three parts. One part that describes the disease dynamic at the individual pig level. Another part that describes the spread of germ cells in the herd, and finally one part that describes the movement and positioning of the animals in the herd. Several parameters are used to describe these model parts, see Jørgensen et al. (1995). The first parameter ( $\Phi_1$ ) describes the period between the infection of the pig and the subsequent release of germ cells from the pig. Another important parameter is the risk of infection ( $\Phi_2$ ), i.e. if a pig is in contact with an infectious pig, what is the risk of becoming infected during a given period. Other parameters describe the spread of the disease, i.e. how much contact is there between neighbouring pens ( $\Phi_3$ ), how important is the airborne infection compared to the contact infection ( $\Phi_4$ ), how much does the airborne infection pressure decrease with distance, ( $\Phi_5$ ). The final part of the model comprises the movement of the animal. Depending on the management strategy, this is well defined. Based on these parameters the result of model can be expressed in several output parameters, such as the number of pigs produced, daily gain and average slaughter weight. In the present context only number of infected pigs,  $\Omega$ , is used. These values correspond closely to the production traits that can be observed in pig units, and can easily be used for economic evaluation of the cost of the diseases under different management strategies. As the term generic disease model implies, the underlying model does not depend on the actual disease in question. The parameter values in the model do, however, differ between diseases. Our initial modelling effort has been concentrated on the disease Atrophic Rhinitis, partly because an attempt of modelling this disease has already been made, Turner et al. (1993).

## 2.2 *Treatment comparison*

The example that is used in this paper is based on our modelling efforts for this disease. As the method shown has been implemented during this model building, the actual process has not been as straightforward, as we present it. An experiment was carried out in the Danish Applied Pig Research (DARP) scheme, where the effect of closed pen partitionings were studied in commercial slaughter pig units. Usually pigs are housed in pens with open partitionings, thus allowing contact between the pigs in neighbouring pens. This experiment was used as a starting point. Using the model, the effect of pen partitionings on the initial disease spread in the herd could be studied. Usually, such a question can be investigated only under very expensive experimental conditions. Note that the mechanistic approach allowed us to model the effect of closed pen partitionings simply by setting the parameter  $\Phi_3 = 0$ . A housing system consisting of a sectioned production system with 10 pens and 16 pigs within each pen was used. The production cycle consisted of keeping the growing pigs in the section for a 105 days production cycle. The production management closely resembled Danish sectioned management. In each of the simulation runs a single newly infected pig were introduced with the other 159 non-infected pigs. The output variable,  $\Omega_{ij}$ : *the number of infected pigs*, were in the range from 0 to 159.  $i$  denotes the treatment, i.e. 1: *Open pen partitionings* and 2: *Closed pen partitionings*.  $j$  is the simulation run.

The procedure started out with the traditional approach used when building simulation models, but the shortcomings of this method lead us to a redefinition of the concept behind our modelling. New techniques within the field of Bayesian statistics were adapted, mainly inspired by the work of Raftery et al. (1995) and described in detail in Givens (1993). Also the developments within Markov Chain Monte Carlo methods have been used, see Gilks et al. (1996).

## 3 **Method**

The following exposition of the method is very brief. An elaborate version can be found in Jørgensen (1997).

### 3.1 *Elements of the Monte Carlo Simulation Method*

The Monte Carlo simulation model is a method for evaluating an integral

$$\Psi = E_{\pi}\{U(X)\} \tag{1}$$

and it involves generating random draws  $X = x^{(j)}$  from the target distribution  $\pi$  and then estimating  $\Psi$

by  $\{U(x^{(1)}) + \dots + U(x^{(k)})\}/k$ . In our context,  $X = \{\Theta, \Phi\}$  is a vector consisting of decision parameters,  $\Theta$ , and system parameters and state variables,  $\Phi$ .  $U()$  is some response function, e.g. a utility function. The Monte Carlo method is thus a numeric method for finding the integral in eq. (1). Often it is an advantage to reformulate the integral in eq. (1) by splitting  $\Phi$  into the so-called state of nature,  $\Phi_O$ , and parameters and state variables  $\Phi_s = \{\Phi_{1s}, \Phi_{1s}, \dots, \Phi_{Ts}\}$  that are calculated by the model. A subset of  $\Phi_s$ ,  $\Omega$  is called the output of the model. The elements of  $\Omega$  are, in general, observable. In our current model context, the state of nature consists of the 5 parameters mentioned previously, while the state variables are such values as the disease state at each time for each pig, the weight at each time etc. The number of infected animals is such an output parameter. This splitting of the parameter vector leads to a reformulation of eq. (1) e.g.  $\Psi = E_{\pi_O} \{E_{\pi_{s|O}} \{U(X)\}\}$ . The dimension of  $\Phi_O$  will in general be fixed by the model structure, while the number of elements in  $\Phi_s$  will vary with different decisions and different combinations of the other elements in  $\Phi$ . Disregarding the problem of dimensionality, the integration with respect to  $\Phi_O$  is well behaved and lends itself to techniques other than simple Monte Carlo simulation. In contrast, the integration with respect to  $\Phi_s$  is of a complexity that is only feasible to solve using the Monte Carlo method.

The model can be seen as a representation of expert knowledge concerning the causal structure ( $U()$ ), the parameter values ( $\Phi$ ) and the output ( $\Omega$ ). Expert knowledge has contributed to the structure of the model, and based on experimental results prior distributions of values of the model parameters ( $\Phi_O$ ) can be found. The expert knowledge is often based on the elements of  $\Omega$ , or rather  $E(\Omega|\Theta')$ , where  $\Theta'$  is a part of the possible decision combinations. (The extrapolation from knowledge based on  $\Theta'$  to decisions in general is typical for the modelling approach. The use of the mechanistic simulation models for this extrapolation is considered more robust than the use of simple empiric models). This gives us a source of information about output parameters that are independent from the model. The problem the model developer faces is, how to make this knowledge consistent, i.e. to specify the correct joint prior distribution of  $(\Phi_O, \Omega|\Theta)$ . Raftery et al. (1995) introduces the concept *post-model* distribution to this joint distribution, while the term *pre-model* distribution refers to the prior distribution of  $\Phi_O$  before the information based on the output parameters has been used.

The usual method is as follows. The prior distribution of the parameters  $\pi_a(\Phi_O)$  are specified. The  $k$  model calculations are carried out using random drawings from this prior distribution, and resulting values of the (observable) output variables are calculated i.e. the simulation runs calculate  $\mathbf{S}_0 = \{x^{(j)}\}$ . If the distribution of the output variables does not 'seem' likely, the prior distribution of the parameters is adjusted and another  $k$  simulation runs are carried out, i.e.  $\mathbf{S}_1$  is generated. This process is iterated with corresponding new  $\mathbf{S}_i$  until the model builder and (possible) the user of the model is satisfied with the validity of the model (or tired of making new simulation runs). In fact with models of animal production systems, it is customary to use point estimates of the parameters in the state of nature,  $\Phi_O$ , and only allow the parameters and state variables  $\Phi_s$  to include stochasticity. This approach is dubious, especially when the model is used for decision support purposes, because it under estimates the risk in the decision making.

### 3.2 The resampling approach

The techniques presented in the following are an alternative approach to these model building steps. They ensure that the process is coherent and can be documented. In addition, they allow for a reuse of the costly simulation runs. When a series of simulation runs has been made, the resulting data set  $\mathbf{S}$  is a random sample from the joint prior distribution,  $\pi_a$ , used in the sampling process (e.g. the *pre-model* distribution). At each point we can calculate the corresponding density  $\pi_a(x^{(j)})$ . If we are interested in a sample from another probability distribution,  $\pi_p$  (e.g. the *post-model* density), we do not need to discard  $\mathbf{S}$  but can reuse the elements. The basic method is to estimate the proportion between  $\pi_p(x^{(j)})$  and  $\pi_a(x^{(j)})$  (in some cases only up to a norming constant) in each of the sampled points. If we are interested in e.g. estimating the expected utility based on this density we can use *importance sampling* (i.e. weigh each observation with the weight factor when the average is calculated). If we are interested in a new sample that follows the new density we can use *acceptance-rejection sampling*, i.e. to select observations from the old data set with a probability depending on the weight factor. This process can be done several times using the *SIR algorithm*, Rubin (1987). Using the *SIR algorithm* (Sampling/Importance Resampling) the individual observations in the data set is no longer independent, i.e. some observations will be replicated. The *Metropolis-Hastings independence sampling* can be used to draw samples from

the density by keeping the next observation in the data set, if it is more likely (i.e. has a higher weight) than the current. Otherwise, the current observation is replicated with a probability depending on the ratio of the weight factors. The *Metropolis-Hastings algorithms* will also result in dependency between observations. See Liu (1996) for a short review and comparison between some of the techniques.

The prior density of the input parameters,  $\pi_a$  will be known, as it is input to the model, but  $\pi_p$  will depend on the *pre-model* distribution of the output parameters, and has to be estimated, based on the model runs. In general, density estimation in a high dimensional parameter space is not tractable. But the special problems that are found within the Monte Carlo simulation models will often make it possible to base the inferences on density estimation with low dimensionality, often univariate kernel density estimation will suffice. What is needed are models for the conditional expectation  $E_{\pi_{S|O}} \{\Omega | \Phi_O\}$  and kernel density estimation of the marginal density of  $\Omega$  and  $E_{\pi_a}(\Omega)$ . Note that Raftery et al. (1995) is only concerned with deterministic models. The modelling of  $E_{\pi_{S|O}} \{\Omega | \Phi_O\}$  is thus trivial in their case. Usually only a small fraction of the elements of  $\mathbf{S}$  will be used, but different resampling approaches can be used to generate new observations from  $\pi_p$  with a higher success rate, e.g. the *Metropolis-Hastings random-walk* algorithm. In each of these cases a new data set can be generated, but the individual elements  $\{x^{(j)}\}$  will not be independent. The approach can be extended to sensitivity calculations, i.e. the effect of a slight change in prior distribution can easily be calculated using e.g. *importance sampling*. (Note that the so-called score function approach, Rubinstein & Shapiro (1993), is another obvious choice for sensitivity estimation). However, this approach will not be applied here.

## 4 The Simulation Study

### 4.1 Prior distributions of model parameters

The prior distribution of the parameters was quantified following discussions with domain experts. The parameters were transformed into something meaningful to the experts. In the model, the parameter  $\Phi_1$  is actually the transition intensity between the state *Infected, but not infectious* and *Infected and infectious*. This was transformed into the median time from *infection to infectious*. Similarly, the daily infection rate with one infectious pig in the pen was transformed into the infection rate in 100 days, corresponding to the production period of the animal, and similarly for the other parameters. The resulting prior distributions are shown in figure 1a-e, as the pre-model distributions.

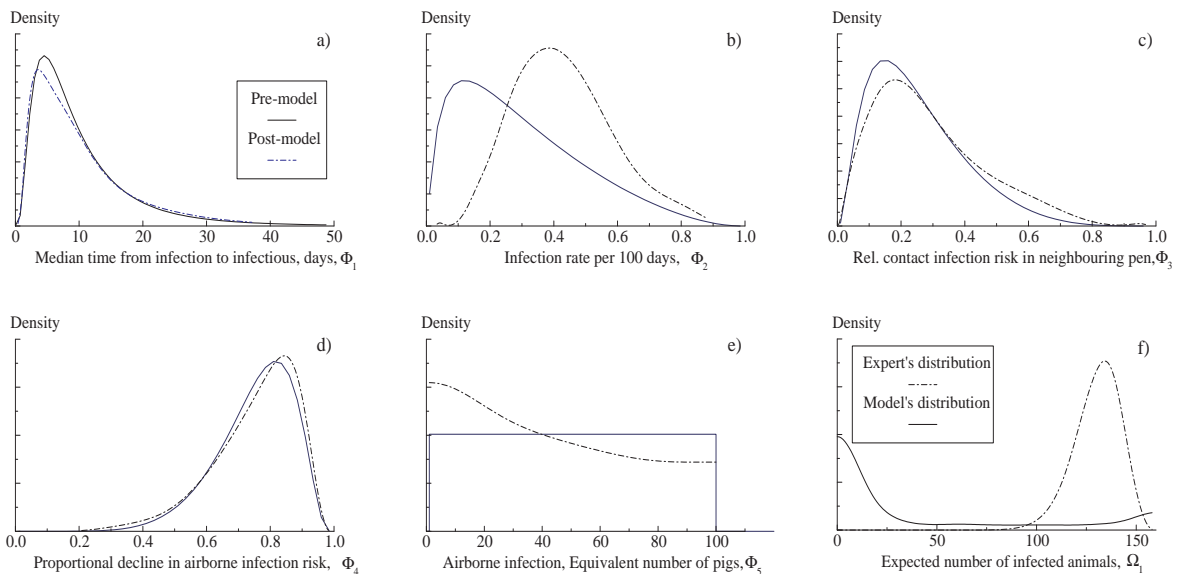


Figure 1: Prior distribution (pre- and postmodel) of parameter values in the model

With respect to the resulting spread of the infection the expert seemed more confident, although they differed slightly. The resulting distribution, which tries to capture their different opinion, is shown in figure 1f.

## 4.2 Simulation runs

Initially 1727 simulation runs were performed generating the data set  $\mathbf{S}_0$ . The duration of each simulation run was approx. 100 sec. Based on these 1727 simulations a logistic quadratic response surface was fitted to the data to obtain a local approximation to  $E_{\pi_{S|O}}\{\Omega|\Phi_O\}$ . The expected number of infected animals were predicted for each of the 1727 simulations runs, and the density of the expected number estimated using a kernel smoother, and shown in figure 1f (the model distribution). This pre-model density was combined with the expert opinion and resampling with the SIR algorithm were used. 10 times repeated sampling of the whole database leaving a total of 1375 or approximately 8 percent of the initial simulation runs per sampling cycle. Approximately 15 percent of to initial simulation runs are represented in this sample with an average representation of 5.4 of those included. 2.4 percent of the initial runs are included in every resampling cycle. Resulting kernel estimates of the parameter distributions were made and are presented in figure 1 as post-model distributions.

## 4.3 The revised distribution

As shown in figure 1f the distribution of the expected number of infected animals was not very informative based on the pre-model distribution of the parameters. As a result, the resampling weights depended almost entirely on the expert's prior distribution of the output values. The distribution of parameter  $\Phi_2$  (figure 1b) is much affected by the introduction of the expert knowledge concerning the output, while the other parameters seem relatively unaffected. Even though the marginal distributions were modified only slightly, dependencies between the variables are introduced, as shown in figure 2b. These dependencies are very sensible, e.g. a given number of infected pigs can be a result of either high risk of being infected for each pig, or a large risk of getting infected from other pigs. Note that the calibration of the model could have been done by simply changing the distribution of the parameter values. Probably by using a prior distribution of parameter  $\Phi_2$  resembling the post model distribution in figure 1b would have sufficed. This would, however, falsely have maintained the independency between the parameter values.

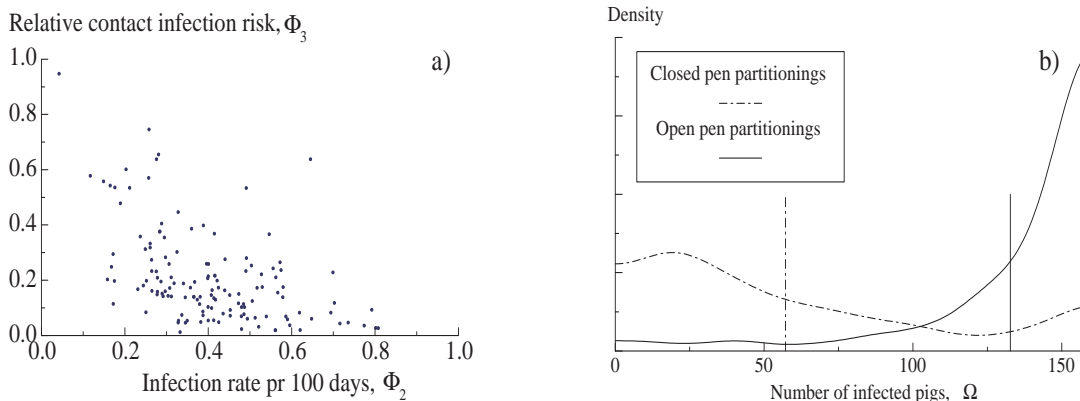


Figure 2: a) Post-model dependency between  $\Phi_2$  and  $\Phi_3$ . b) Treatment comparisons, estimated mean indicated by vertical lines

## 4.4 Results of treatment comparisons

The effect of the treatment was to eliminate the contact spread between neighbouring pens, i.e.  $\Phi_3 = 0$ . Here the simulation model had to be run again with this value for  $\Phi_3$ . However, the other parameters should follow the post-model distribution. Again the SIR approach was used. The sampled observations were used as initial values for new simulation runs except that  $\Phi_{3j} = 0$ . This generated  $\Omega_{2j}$ .  $\Omega_{1j}$  was already in the data set. Thus was a series of paired observations ( $\Omega_{1j}, \Omega_{2j}$ ) generated. The resulting distribution are shown in figure 2b. In total 831 additional simulation runs were obtained. Because of the negative correlation between  $\Phi_2$  and  $\Phi_3$  as shown in figure 2a,  $\Omega_2$  has a high variation. If this correlation was not taken into account, e.g. if the calibration had been done simply by changing the distribution of  $\Phi_2$ , the result would have been a lower variation of  $\Omega_2$ , because the variation due to

$\Phi_3$  had been removed. Of course this would be against common sense, the highest precision would be expected where the expert knowledge had been used in calibration, i.e.  $\Omega_1$ .

## 5 Discussion and conclusion

The methods presented seems to be an ideal framework to use, when working with the complex models that are used within research in animal production systems. These models will often be very detailed because the researchers want to maintain an almost one to one correspondence between the real system and the model, even though a much simpler model with only a few parameters usually could make predictions just as well, at least within a restricted problem domain.

The possibilities for reuse of the simulation runs can be used for studying a wide range of questions, e.g. sensitivity analysis, importance of different expert view. Givens (1993) lists several other possibilities. The method can also be applied to Markov Chain models, to include parameter uncertainty. One obvious application of the method is to adapt a simulation model to the production level of an individual herd. A general database can be generated using the time-consuming simulation model runs. Using this database, the adaptation to different herds can be made simple by weighing the elements of the general database according to the level of each herd. This possibility could be explored for on-farm use of the simulation models or even for web-based solution, where the storing and handling of large general databases should be without problems, and the response times should be adequate for on-line purposes. With the current high activity in research in numerical methodology within Bayesian statistics and analysis of highly structured stochastic systems, further improvements and potential applications should be expected in the future

## References

- Gilks, W., Richardson, S., & Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.
- Givens, G. (1993). *A Bayesian Framework and Importance Sampling Methods for Synthesizing Multiple Sources of Evidence and Uncertainty Linked by a Complex Mechanistic Model*. Ph.D. dissertation. Department of Statistics, University of Washington, Seattle.
- Jørgensen, E. (1997). Techniques for reuse of monte carlo simulation runs.  
[http://www.sp.dk/~ej/monte\\_tech.zip](http://www.sp.dk/~ej/monte_tech.zip)
- Jørgensen, E. & Kristensen, A. (1995). *An object oriented simulation model of a pig herd with emphasis on information flow*, In *FACTs 95 March 7, 8, 9, 1995, Orlando Florida, Farm Animal Computer Technologies Conference*, pages 206–215.
- Jørgensen, E., Kristensen, A., & Vestergaard, E.-M. (1995). Modelling of respiratory disease. working paper: Generic disease model. *Dina Notat* **40**, 1–18.
- Liu, J. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing* **6**, 113–119.
- Raftery, A., Givens, G., & Zeh, J. (1995). Inference from a deterministic population dynamics model for bowhead whales. *J. Amer. Statist. Assn.* **90** (430), 402–430.
- de Roo, G. (1987). A stochastic model to study breeding schemes in a small pig population. *Agr. Syst.* **25**, 1–25.
- Rubin, D. (1987). Comment on 'the calculation of posterior distributions by data augmentation'. *J. Amer. Statist. Assn.* **82**, 543–546.
- Rubinstein, R. & Shapiro, A. (1993). *Discrete Event Systems. Sensitivity analysis and stochastic optimisation by the score function method*. John Wiley & Sons, Chichester.
- Singh, D. (1986). Simulation of swine herd population dynamics. *Agr. Syst.* **22**, 157–183.
- Turner, L., Wathes, C., & Audsley, E. (1993). *Dynamic Probabilistic Modeling of Atrophic Rhinitis in Swine. Paper No. 93-4559*, In Anonymous, editor, *Proceedings 1993 International Winter Meeting of ASAE, Chigaco Illinois*, pages 1–19. ASAE, 2950 Niles Rd. St. Joseph, MI 49085-9659 USA.