ELSEVIER

European Journal of Operational Research 87 (1995) 21–34

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

Case Study

Statistical validation of simulation models

Jack P.C. Kleijnen [1]

Department of Information Systems and Center for Economic Research (CentER), Tilburg University, 5000 LE Tilburg, Netherlands

Abstract

Rigorous statistical validation requires that the responses of the model and the real system have the same expected values. However, the modeled and actual responses are not comparable if they are obtained under different scenarios (environmental conditions). Moreover, data on the real system may be unavailable; sensitivity analysis can then be applied to find out whether the model inputs have effects on the model outputs that agree with the experts' intuition. Not only the total model, but also its modules may be submitted to such sensitivity analyses. This article illustrates these issues through a case study, namely a simulation model for the use of sonar to search for mines on the sea bottom. The methodology, however, applies to models in general.

Keywords: Modeling; Simulation; Scenarios; Regression; Military

1. Introduction

This section answers the following questions. What is validation? What has the literature to say in general on this topic; how about case studies? What is the role of statistical techniques? What is this article's contribution; how is it organized?

Validation is defined in this article following a classic simulation textbook (Law and Kelton, 1991, p. 299): "*Validation* is concerned with determining whether the conceptual simulation model (as opposed to the computer program) is an accurate representation of the system under study". Hence, validation cannot result in a perfect model: the perfect model would be the real system itself. Instead, the model should be 'good enough', which depends on the goals of the model. For example, some applications need only relative (not absolute) simulation responses corresponding to different scenarios, as this article will demonstrate.

General discussions on validation of simulation models can be found in all textbooks on simulation. Examples are Banks and Carson (1984), Law and Kelton (1991, pp. 298–324), and Pegden, Shannon, and Sadowski (1990, pp. 133–162). A well-known article is Sargent (1991). Recent survey articles are Balci (1995), including 102 references, and Kleijnen (1995), including 61 references.

Case studies on validation, however, are rare. Kleijnen (1995) could retrieve only a few case studies.

[1] E-mail: kleijnen@kub.nl, fax: +31-13-663377.

SSDI 0377-2217(95)00132-8

There is a special need to further develop the theory on validation, especially in view of the great importance of validation in the practice of Operations Research (OR). A false model may generate output that is sheer nonsense, or worse, it may generate subtle nonsense that goes unnoticed. Such a model may lead to wrong decisions.

The present article is meant to contribute to the practice and the theory of validation. Though the case study concerns a specific simulation model, the methodology applies to models in general. This article discusses in detail how to apply familiar statistical techniques such as regression analysis, design of experiments, and *t*-tests. It further shows how large simulation models can be validated in two stages: in stage #1 individual modules are validated (see Section 3.1); in stage #2 the whole simulation model is treated as one black box, and is validated (see Section 3.2). Further, all simulation models with randomness lead to the question how to compare real and simulated responses through statistical tests. Other general topics in validation are briefly discussed in Sections 4.1 through 4.6.

Statistical techniques may yield reproducible, objective, quantitative data about the quality of a given simulation model. Experience shows that the correct use of mathematical statistics in operations research is not so simple. It is easy to apply the wrong statistical techniques: there is much statistical software, but that software does not warn against abuse (such as violations of the statistical assumptions). That software certainly does not instruct the operations researchers to apply mathematical statistics to validation problems. On hindsight the correct use of statistics may seem easy. Balci (1995) states: "False beliefs exist about testing... testing is easy... no training or prior experience is required." The statistical analysis in this article deviates from the analyses used in other naval studies. The latter studies are rather crude from the viewpoint of mathematical statistics. Notice that statistical techniques do not solve all problems in validation (Forrester and Senge, 1980).

Recently the interest in validation has shown a sharp increase in the USA defense community (Kleijnen, 1995, gives seven references; also see Balci, 1995). In Europe and China the defense organizations also seem to take the initiative (Wang et al., 1993). The renewed interest in validation is further illustrated by a recent monograph (Knepell and Arangno, 1993), and a Special Issue on 'Model Validation in Operational Research' of the *European Journal of Operational Research* (Landry and Oral, 1993).

Unfortunately, this interest has not resulted in a standard theory on validation. Neither has it produced a standard 'box of tools' from which tools are taken in a natural order (Landry and Oral, 1993). There does exist a plethora of philosophical theories, statistical techniques, software practices, and so on. Several *classifications* of validation methods are possible (Kleijnen, 1995, gives six references). The emphasis of the present article is on statistical techniques.

The case study in this article will illustrate that there are no perfect solutions for the problems of validation in simulation. The whole process has elements of art as well as science.

The study concerns a model for the use of *sonar* when searching for mines on the sea bottom. The model was developed for the Dutch Navy, by TNO-FEL (Applied Scientific Research – Physics and Electronics Laboratory); TNO-FEL is a major military research institute in The Netherlands. The model is called HUNTOP (mine HUNTing OPeration). Other countries have similar simulation models for naval mine hunting (the corresponding literature is classified).

The rest of this article is organized as follows. Section 2 discusses the HUNTOP model in some detail. This model includes several factors: the environment (namely, the mine field and acoustical characteristics of the sea water), the sonar system, the ship's course, and the human operator's performance. Section 3 validates the simulation model in two stages (these two stages correspond with the levels 2 and 3 in Balci, 1995). Section 3.1 gives sensitivity analyses of some modules, applying experimental design theory and regression analysis. Section 3.2 compares simulated detection probabilities – resulting from the model as a whole – with real probabilities. This comparison encounters statistical complications, such as dependencies among estimated detection

probabilities of different mines. It is important to measure the scenarios (environments) that drive the simulation and the real-life test respectively. Section 4 discusses remaining issues. In practice, a validation project has limited time and financial resources; so at the end of the project there remain problems to be investigated. Sections 4.1 through 4.6 discuss issues that arise in the validation of simulation models in general (namely, screening, risk analysis, Gaussian approximations, type I and II errors, less stringent statistical validation, and animation); Sections 4.7 through 4.13 briefly present remaining validation problems that are specific for naval mine hunting models. Section 5 gives a summary and conclusions. Some conclusions hold for simulation and modeling in general, whereas some results apply only to this particular case study.

2. Naval mine hunting model HUNTOP

HUNTOP is a complicated simulation model that reflects the combined knowledge of a number of experts in naval mine hunting. This article is not meant to discuss all the intricacies of naval mine hunting. Instead, this article focusses on the validation of HUNTOP. Therefore this section is limited to those aspects of HUNTOP that are necessary to understand this article's approach to the validation of simulation models.

Naval mine hunting is performed by ships equipped with *sonar*. Conceptually, a sonar may be viewed as a *torchlight*: in the 'dark' a certain area becomes 'lighted' or 'insonified', so objects within that area may become visible on a sonar display. Hence, as the ship with its sonar moves, new areas become visible, while previous areas move out of sight.

Dispersed over the area are *mines and NOMBOs*, NOn-mine Minelike Bottom Objects, which are harmless objects that look like mines. Both types of objects can be detected only if they are within the insonified area.

Operationally, an imaginary straight line or *track* is drawn over the area. The ship tries to follow this track; however, navigation errors do occur. To cover the whole mine field, several tracks may be planned. At both sides of each track, the area is subdivided into *strips* that are mutually exclusive and exhaustive.

Sonar stands for SOund NAvigation Ranging: the device detects objects by the *sound* waves these objects reflect (see the definition in 'The Random House Dictionary'). Physics theory proves that sound velocity varies with water temperature and salinity. Obviously, this temperature and salinity vary with the water depth. The experts use a *Sound Velocity Profile* (SVP), which maps sound velocity as a function of depth. In HUNTOP an SVP is a simple piecewise-linear function that is kept constant during the whole simulation run; a simulation run is one voyage of the ship over the whole mine field. In practice, however, the SVP varies along the track. And even at the same place, the SVP will show seasonal and daily variations. So from the start of the validation project, the SVP seemed an important factor.

When an object is insonified, its echo appears on the sonar screen with a certain *contrast*. Technically, this contrast is determined by the following three components (a full understanding of these components is not essential for this article, but terms displayed in italics will be used further on):

(i) The *echo* of the object itself. This echo depends deterministically on several factors, for example, the object's size.

(ii) *Reverberation*: the echo of the object's environment, that is, reflections from the sea bottom, the water surface, and the water itself. Reverberation depends deterministically on the *grazing* angle (the angle at which the sonar beam hits the bottom), the *bottom type* (a rocky bottom reflects sound more than sand does), and some more factors.

(iii) *Acoustic noise*: sounds generated by the ship, waves, marine life, and so on. Acoustic noise may generate random, *spurious* contrasts.

Mines may be hidden behind hills on the sea bottom. So the *bottom profile* seems to have an important effect on the mine detection probabilities.

A contrast may be missed by the *human operator*. Human behavior shows noise and is therefore

represented by statistical distribution functions, called operator curves. An *operator curve* gives the detection probability of an echo as an increasing function of the time that the echo has been visible.

An object is *visible* only during a certain time, which depends on the *sonar window* (comparable with the light circle of a torch) and on the *ship position* (position of the object relative to the ship's course). When the object becomes invisible or the operator is busy, the detection probability drops to zero.

Actually the model uses *several* operator curves. For example, if there are many echoes on the sonar screen, then the detection probability of an individual object is lower, all other things being equal.

(Whenever a detection occurs, the operator must classify the observed contrast as either a mine or a NOMBO. That classification may be true or false. HUNTOP, however, does not cover this classification stage nor any other follow-up operations such as sending an unmanned mini-submarine to identify a classified object or to neutralize or destroy the mine. See also Section 4.11.)

The *laws of physics* (for example, Snell's law) that govern sonar beam propagation are well known and are *deterministic*. The *environment*, however, is not well known: accurate information on the current SVP and the sea bottom's profile is hard to obtain. The simulation uses a single SVP within one run; the bottom profile is modeled by a simple geometric pattern, which is fixed within a single simulation run. So, even if the model is perfect, it may give the wrong answer when fed with the wrong inputs (problem of data validity). This type of *uncertainty* must be distinguished from *random noise*, which occurs in the operator module (and in other modules that will follow); also see Kleijnen (1994).

In many physics laws, *time* is continuous. The simulation model, however, is programmed with time sliced into periods of fixed length (discrete event simulation, such as queueing simulation, uses time steps of variable length). In other words, the model consists of difference equations, not differential equations. (The time slice has a length of three seconds if the ship's speed is two meters per second; at this time step, numerical accuracy is acceptable.)

The model is *calibrated*: a parameter without physical interpretation is used to modify the computed contrasts such that the model's outputs are closer to the outputs observed in practice (also see Section 4.13).

The model is meant to be used for diverse *purposes*. One goal may be to compare different *tactics* for mine hunting; for example, the tilt angle of the sonar may be changed (in the torch-light analogy, think of shining farther away, so a larger area is seen, albeit with less intensity). Moreover, a given tactic may give different results depending on the *environment*. Therefore the non-controllable, environmental factors should also be investigated. Besides these relative responses, *absolute* predictions are of interest: the expected detection probabilities in a given situation may be used to determine the 'huntability' of the mine field and to assess the performance of a particular sonar system. The presence of several goals complicates the validation: see the definition of validation in Section 1.

Altogether the simulation model has nearly *40 inputs*; some were mentioned above. That model is organized into a number of *modules or subroutines*. Examples are the ship's position, the operator's state, the object's visibility, and the object's contrast; the latter three modules give the inputs for the detection probability module.

(There are actually several model options. For example, the SVP may be either input to the model or it may be calculated as a function of salinity and temperature. This article, however, concentrates on the SVP as input. Other examples are reverberation and noise, which are also modeled in two ways. Moreover, there is an analytical variant of this model. The simulation results can be used to check this analytical model.)

For reasons of confidentiality this article does not give more details on HUNTOP (those details are presented in a classified report, Alink and Vermeulen, 1991).

Summary. In the HUNTOP model the mine detection probabilities depend primarily on the following factors.

(i) Environmental factors: the mine field (including the number of mines and NOMBOs), the sea (depth, SVP, and noise level), and the sea bottom (type and profile).
(ii) The sonar system (technical specifications and operational settings such as tilt angle).
(iii) The ship's course (including navigation error).
(iv) The operator's performance.

3. Validation

The present section is more or less a *chronological* account of issues that arose in the HUNTOP validation study. As this section will show, validation may proceed in two stages: in stage #1 individual modules are validated (see Section 3.1); in stage #2 the whole simulation model is treated as one black box, and is validated (see Section 3.2).

3.1. Sensitivity analysis per module

Some modules within the model give *intermediate* output that is hard to observe in practice, and hence hard to validate. Sensitivity analysis may be applied to such modules, in order to check if certain factor effects have signs (directions) that agree with experts' prior qualitative knowledge.

(If the real system being simulated does not yet exist, then real-world data are not available. In that case sensitivity analysis should be applied to the whole model too.)

In *practice*, sensitivity analysis is done *ad hoc*. Often a base case is selected. Next each factor is changed, one at a time. Two or three values are simulated for each quantitative factor. For qualitative factors a few 'values' are simulated. The resulting responses are analyzed crudely, for example, 'eye-balled'. Van Groenendaal (1994) gives several examples of such 'practical' sensitivity analysis.

This article, however, advocates the following *scientific* approach. First specify a *regression metamodel* or *response surface*, that is, approximate the input/output behavior of the complicated simulation model by a simple function. Examples of such approximations are: (i) models with only main or first-order effects, (ii) models augmented with two-factor interactions or cross-products between pairs of inputs, and (iii) models further augmented with pure quadratic effects (which quantify curvature of the response surface) (Kleijnen, 1987, 1994, 1995). Examples will be given for HUNTOP.

Based on the regression model with (say) Q effects, select an *experimental design*, that is, a combination of (say) n input values for the simulation model. Obviously, the more parameters the metamodel has, the more combinations are required. For example, a first-order approximation with three inputs (x_1, x_2, x_3) requires that four combinations be simulated: there is the dummy factor $(x_0 = 1)$ with its 'grand effect' or 'intercept' (say) β_0; so in total there are four effects to be estimated, namely β_0 through β_3.

Next estimate the factor effects β from the simulated input combinations. Apply the well-known *least squares* algorithm.

Then check if the estimated metamodel approximates the simulation model adequately. That *fit* can be simply quantified through the well-known multiple correlation coefficient R^2. (A more complicated procedure is cross-validation; see Kleijnen and Van Groenendaal, 1992).

If the fit is not good enough, *transform* the inputs; examples are the logarithmic transformation $(\log x)$ and the inverse $(1/x)$.

Once the fit of the approximation has been checked, study the *individual estimated factor effects* β_1, β_2, etc. Qualitative knowledge about the simulated module often suggests that these effects should have specific signs; for example, deeper water gives a wider sonar window (see β_2 in the sonar window module below).

Experience shows that this methodology is flexible enough in practice.

(The importance of sensitivity analysis is also emphasized by Fossett et al. (1991, p. 719). They investigate three military case studies, but do not present any details.)

Because of time constraints, only two modules

are examined in the HUNTOP case study: (i) sonar window, and (ii) visibility.

Sonar window module

The sonar window module has as *response* variables the minimum and maximum distances of the area on the sea bottom that is insonified by the sonar beam. *Factors* are selected as follows (this selection depends on prior knowledge of the simulated system, not on mathematical statistics): the sonar rays hit the bottom under the grazing angle (see Section 2), which is determined deterministically by three factors, namely SVP denoted by x_1, average water depth or x_2, and tilt angle or x_3. SVP is treated as a qualitative factor.

As the first response variable (say) y take the *minimum* distance from the sonar to the insonified area on the sea bottom (actually the sonar position is projected onto the imaginary flat sea bottom).

Specify a second-degree polynomial in x_2 and x_3 per SVP type (or x_1 'value'). Such a polynomial seems a good compromise between a simple first-degree polynomial (which misses interactions and has constant marginal effects) and a higher-order polynomial (which is difficult to interpret and requires many more simulation runs).

To estimate the six regression parameters of this polynomial ($Q = 6$), use a classical central composite experimental design with nine input combinations ($n = 9$); also see Kleijnen (1987) and Kleijnen and Van Groenendaal (1992).

The fitted second-degree polynomial turns out to give an acceptable approximation: the multiple correlation coefficient R^2 ranges between 0.96 and 0.98, for the four SVPs simulated. (Otherwise, the regression variables x_2 and x_3 could have been transformed.)

Expert knowledge suggests that certain factor effects have specific signs: $\beta_2 > 0$, $\beta_3 < 0$, and $\beta_{23} < 0$. The corresponding estimates turn out to have the correct signs. So this module has the correct input/output behavior, and the validity of this module need not be questioned.

(The pure quadratic effects are not significantly different from zero. So on hindsight, simulation runs could have been saved, as there is no curvature in the response surface. Once the simulation model has been validated, the signs of the effects in the metamodel may also help the decision makers in the optimization of their policies (Kleijnen and Van Groenendaal, (1992).)

For the second response, maximum distance, similar results hold. The exception is one SVP that results in an R^2 of only 0.68 and a non-significant β_2.

Visibility module

An object is visible if it is within the sonar window and it is not concealed by the bottom profile. HUNTOP represents the bottom profile through a simple geometric pattern, namely hills of fixed heights with constant upward slopes and constant downward slopes. A fixed profile is used within a single simulation run. Intuitively, the orientation of the hills relative to the ship's course and to the direction of the sonar beam is important: does the sonar look down a valley or is its view blocked by a hill?

The *response* variable of this module is the time that the object is visible, expressed as a percentage of the time it would have been visible were the bottom flat (in which case no concealment could occur). This response is random because the ship's course shows navigation error. Navigation error is modeled by a normal distribution with the desired course over the track as the mean value.

Six *inputs* are varied: water depth, tilt angle, hill height, upward hill slope, downward hill slope, and object's position on the hill slope (top, bottom, or in between). The SVP and the orientation of the bottom profile are kept constant. Navigation error is eliminated in this sensitivity analysis (not in stage #2; see Section 3.2).

Again specify a quadratic metamodel for this module. To estimate the 28 regression parameters, use a central composite design with 77 input combinations. R^2 turns out to be 0.86. The upward hill slope has no significant effects: no main effect, no interactions with the other factors, no pure quadratic effect. These results agree with the experts' qualitative knowledge. So the validity of this module is not questioned either.

3.2. Real versus simulated detection probabilities

This subsection answers the questions: (i) what are the correct probabilities to be estimated, and (ii) how can the estimated (correct) probabilities be compared statistically?

Relevant probabilities

Let M denote the number of mines in the simulated mine field, and R the number of simulation runs. A simulation run is one voyage of the ship over the whole mine field (see Section 2). During that run an individual mine is either detected or not. So define the *simulation binary variables*:

$$x_{ij} = \begin{cases} 0 & \text{if simulated mine } i \text{ is not detected in simulation run } j, \\ 1 & \text{if simulated mine } i \text{ is detected in simulation run } j, \end{cases}$$

$$(i = 1, \ldots, M \text{ and } j = 1, \ldots, R) \tag{1}$$

This equation leads to the following definition of the *simulation detection probability* p_i for mine i that holds for all simulation runs j:

$$P(x_{ij} = 1) = p_i, \tag{2a}$$

$$P(x_{ij} = 0) = 1 - p_i. \tag{2b}$$

Analogous to R, define K as the number of so-called *field runs* which are performed during a real mine sweep at sea (through a mine field that has been constructed for research and training purposes). Assume that the number of simulated mines in the validation stage (namely M) equals the number of mines in the field runs. (Once the model is validated, the number of mines in the model can change.) Analogous to x_{ij} in (1), define the *real-life binary variables* y_{ik}:

$$y_{ik} = \begin{cases} 0 & \text{if real mine } i \text{ is not detected in field run } k, \\ 1 & \text{if real mine } i \text{ is detected in field run } k, \end{cases}$$

$$(i = 1, \ldots, M \text{ and } k = 1, \ldots, K). \tag{3}$$

Analogous to (2), define the *real-life detection probability* q_i for mine i:

$$P(y_{ik} = 1) = q_i, \tag{4a}$$

$$P(y_{ik} = 0) = 1 - q_i. \tag{4b}$$

A major problem in this case study is the use of different *environments* in the simulation model and the field test respectively. Firstly, consider the SVPs. HUNTOP uses crude approximations of the SVPs in the real world, namely simple piecewise-linear functions, kept constant during the whole simulation run (see Section 2). The real SVPs are poorly measured. Secondly, consider the *mine fields*. In the real world, mines have locations that are not known exactly. So on one hand, an echo is not counted as a detection if its origin is 'far' away from the assumed locations of the real mines. On the other hand, 'false' echoes (NOMBOs and spurious contrasts) are counted as detections if their origins are close to the assumed location of a real mine.

So environmental conditions are uncertain in the real world, and they are crudely represented in the model. Obviously, the modeled and the real detection probabilities depend on uncertain but deterministic inputs such as the SVP and the mine field. These inputs define *scenarios*. (These inputs must be distinguished from the stochastic inputs, namely navigation error, spurious contrasts, and human performance; see Section 2.)

There are numerous scenarios, denoted by (say) S_h with $h = 1, 2, \ldots$ So analogous to x_{ij} in (1), define the *scenario dependent* simulation binary variables:

$$x_{ijh} = \begin{cases} 0 & \text{if simulated mine } i \text{ is not detected in simulation run } j \text{ under scenario } h, \\ 1 & \text{if simulated mine } i \text{ is detected in simulation run } j \text{ under scenario } h. \end{cases} \tag{5}$$

Analogous to p_i in (2), define simulation detection probabilities for mine i, *conditional on scenario* h:

$$P(x_{ij} = 1 \mid S_h) = p_{ih}. \tag{6}$$

To estimate p_{ih} from R simulation runs, keep the scenario fixed at S_h and use pseudorandom num-

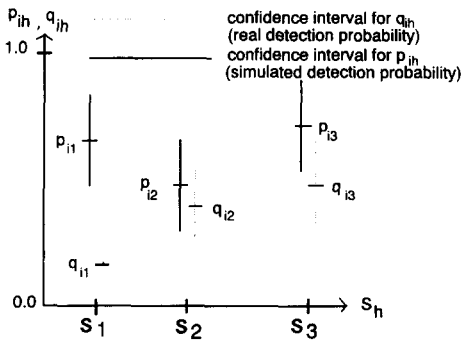

Fig. 1. Sensitivity of detection probabilities p and q for scenario S, for mine i.

bers to sample navigation errors, spurious contrasts, and human operator performance. This yields the estimator $\hat{p}_{ih}$.

To define q_{ih}, the real-life detection probability for mine i *conditional on scenario* h, replace x by y in (6).

Obviously the conditional probabilities p_{ih} in (6) and the unconditional probability p_i in (2) are connected by

$$p_i = \sum_h P(x_{ij} = 1 \mid S_h) P(S_h). \tag{7}$$

If it were desired to estimate this p_i (the average over all scenarios), then scenarios would be sampled too. In validation, however, the *conditional* probabilities p_{ih} and q_{ih} should be compared, not the unconditional probabilities p_i and q_i, as Fig. 1 demonstrates. In this figure the conditional probabilities -simulated and real ones- depend on the *scenario*: $p_{i1} > p_{i2}$ and $p_{i1} < p_{i3}$, while $q_{i1} < q_{i2} < q_{i3}$. The figure also displays confidence intervals for the corresponding estimators: see the vertical lines.

At scenario S_1 the model is not valid: the confidence intervals for p_{i1} and q_{i1} do not overlap. At S_2 the model is valid: the confidence intervals around p_{i2} and q_{i2} overlap largely; that is, the real and simulated probabilities are the same, practically speaking. At S_3 the model *might* be acceptable, depending on the practical purpose of the simulation study (also see the definition of validation in Section 1).

Suppose that the model is run with scenario S_2 (see the confidence interval for p_{i2}), whereas the field test uses scenario S_1 (see the interval for q_{i1}). Then the model is incorrectly rejected (for S_2 the model is valid).

So the validation procedure must estimate how much the simulated and the real detection probabilities respond to different scenarios. If the detection probabilities are found to be sensitive to the scenario, *then scenarios must be measured accurately*. If such measurement is infeasible, only less stringent validation tests are possible (see Section 4.5).

(In queuing simulations, scenarios may correspond with traffic loads, and estimated detection probabilities with average waiting times. Obviously, if the traffic load is not measured, it is virtually impossible to validate the queueing model. Further, Figure 1 shows that the estimated simulation responses (here: $\hat{p}_{ih}$) may lie within the range of estimated real-life responses (here $\hat{q}_{ih}$), so without measurement of the scenario the model cannot be rejected.)

The importance of the environment is also emphasized by Fossett et al. (1991, p. 714). A similar issue is discussed, in the context of ecological models, by Fleming and Schoemaker (1992). Also see the monograph by Knepell and Arangno (1993, pp. 2–9).

In the practice of naval mine hunting, several field runs are made, each in a different period. Measurements show that the SVPs (or scenarios) vary within a run; they also vary from period to period.

The HUNTOP model turns out to give estimated detection probabilities for some mines that are *not* sensitive to the scenario: for some mines these probabilities are always zero; for some other mines these probabilities are always one, whatever the scenario is.

Statistical test

Once the simulated and the real detection probabilities are obtained, it turns out that these estimates are not exactly the same. Is this difference to be explained by noise or by a systematic deviation between model and reality, given a specific scenario? In validation it is hoped that this difference is explained by noise, practically speaking.

Mathematical statistics may be applied to obtain quantitative data about the quality of a model (see Section 1). So specify the *null-hypothesis*: the model and the real system give the same detection probabilities under a specific scenario, for each mine. In symbols:

$$H_0: \quad p_{ih} = q_{ih} \quad (i = 1, \ldots, M \text{ and } h = 1, 2, \ldots). \tag{8}$$

The probabilities p_{ih} and q_{ih} are estimated, by simulation and by field runs respectively, assuming the scenario can be fixed at S_h. First consider a single mine (say) mine i (after Eq. (13) the case of more than one mine will follow).

The estimator $\hat{p}_{ih}$ R is *binomially* distributed with parameters p_{ih} and R. (The R simulation runs give independent responses, since x_{ijh} and $x_{ij'h}$ are independent for $j \neq j'$ where j and j' run from 1 to R; however, the detection probabilities of mine i and i' may be dependent within the same run; this dependence will be taken care of in Eq. (15)). So the variance of this binomial variable is

$$\text{var}(\hat{p}_{ih}) = p_{ih}(1 - p_{ih})/R. \tag{9}$$

Analogously, the field runs give binomial variables $\hat{q}_{ih}$ K with parameters q_{ih} and K.

The simulated and real estimators $\hat{p}_{ih}$ and $\hat{q}_{ih}$ are independent, because the simulation outputs depend on pseudorandom numbers whereas the real outputs depend on completely different random events.

Next consider the *variance* of the *difference* $\hat{p}_{ih} - \hat{q}_{ih}$ under the hypothesis that the simulated and the real probabilities are equal indeed (see Eq. (8)). Denoting these equal probabilities by r_{ih} gives

$$\text{var}(\hat{p}_{ih} - \hat{q}_{ih} \mid p_{ih} = q_{ih} = r_{ih}) = \frac{r_{ih}(1 - r_{ih})}{R} + \frac{r_{ih}(1 - r_{ih})}{K} = \frac{r_{ih}(1 - r_{ih})(K + R)}{RK}. \tag{10}$$

To estimate this *common* parameter r_{ih}, use the *pooled* estimator or weighted average

$$\hat{r}_{ih} = \frac{\hat{p}_{ih}R}{R + K} + \frac{\hat{q}_{ih}K}{R + K}. \tag{11}$$

Next remember the basic relation

$$E(\hat{r}_{ih}^2) = \text{var}(\hat{r}_{ih}) + [E(\hat{r}_{ih})]^2 = \frac{r_{ih}(1 - r_{ih})}{R + K} + r_{ih}^2. \tag{12}$$

Finally, derive the *unbiased* variance estimator of $\hat{p}_{ih} - \hat{q}_{ih}$:

$$\hat{\text{var}}(\hat{p}_{ih} - \hat{q}_{ih} \mid p_{ih} = q_{ih} = r_{ih}) = \frac{\hat{r}_{ih}(1 - \hat{r}_{ih})(R + K)^2}{(R + K - 1)(RK)}. \tag{13}$$

Obviously the hypothesis in (8) requires a *two-sided test*.

Actually there are *several mines*: $M > 1$. Hence the null-hypothesis in (8) is a so-called composite hypothesis, requiring simultaneous testing. Therefore apply *Bonferroni's inequality*, which means that the hypothesis is rejected if one or more mines have estimated simulation and real detection probabilities that differ significantly. Further, each individual mine is tested at a *per comparison* error rate of α/M, where α denotes the *experimentwise* type I error rate. A typical value for this α is 0.20 (independent of M). Details are given in Kleijnen (1987, p. 42) and Miller (1981).

Notice that Bonferroni's inequality permits the use of univariate techniques; so – contrary to Balci (1995) statement – multivariate procedures are not a 'must'.

For convenience, approximate the distribution of the difference between two binomial variables by a *Gaussian*. Assume that this approximation is good enough, since the Central Limit Theorem applies (see Eq. (14) below), and many other approximations are used in the whole process of model building and validation. Also see Section 4.3.

Obviously, the estimated mean of this normal distribution is

$$\hat{p}_{ih} - \hat{q}_{ih} = \sum_{j=1}^{R} x_{ijh}/R - \sum_{k=1}^{K} y_{ikh}/K. \tag{14}$$

The nuisance parameter, namely the variance of this normal distribution, is estimated through (13).

Let z_α denote the 'upper α point' or $1-\alpha$ quantile of the standard normal distribution. Then *reject* the null-hypothesis in (8) if

$$\max_i \left[\frac{|\hat{p}_{ih} - \hat{q}_{ih}|}{\{\hat{\text{var}}(\hat{p}_{ih} - \hat{q}_{ih})\}^{1/2}} \middle| p_{ih} = q_{ih} = r_{ih} \right] > z_{\alpha/M}. \quad (15)$$

Notice that *Bonferroni's inequality* applies, even though the M estimated probabilities within a given simulation run are *dependent*. Indeed, if the operator is busy with one mine, then there is a higher chance that he misses the next mine. Similarly the estimated probabilities for various mines within a particular field test may be dependent.

It is well-known (Balci, 1995) that when *testing* the validity of a model, there are two classical *error sources*, namely the type I or α error and the type II or β error:

α = probability of rejecting the model if the model is valid, (16a)

β = probability of accepting the model if the model is not valid. (16b)

The β error probability increases as the α error probability decreases, given fixed sample sizes (R and K). The complement of the type II error probability, $1-\beta$, is called the *power* of the test. This power increases, as the *model specification error* $\delta = |p_{ih} - q_{ih}|$ increases. The power also increases as the sample sizes increase. Simulated sample sizes can be large (here R); so it might seem that small, unimportant specification errors δ will be declared significant. However, the real sample sizes are always relatively small (here K), so a model is rejected only if the specification error δ is relatively large. Kleijnen (1995) further discusses the appropriateness of statistical tests in validation; also see Section 4.4.

Unfortunately, the *outcomes* of these validation tests can not be presented here, as they are classified. The statistical analysis described above, deviates from the analyses used in other naval studies. The latter studies are rather crude from the viewpoint of mathematical statistics. Those analyses are confidential too, so details cannot be given.

4. Remaining issues

Validation is a continuing process: the environment keeps changing, so the model must be updated and revalidated. Indeed, this case study concerns an ongoing modeling effort at TNO-FEL (also see Section 4.7).

On the other hand, a validation *project* has limited time and financial resources, so at the end of the project there remain issues to be investigated. Sections 4.1 through 4.6 discuss issues that arise in the validation of simulation models in general, whereas Sections 4.7 through 4.13 examine validation problems that are specific for naval mine-hunting models.

Knepell and Arangno (1993) also discuss the ongoing character of validation and its project character.

4.1. Screening

Sensitivity analysis was applied to only two modules (see Section 3.1). So not all 40 factors of the total model were systematically investigated. *Screening* of so many factors can be done through the sequential technique based on aggregation, explained in Bettonvil and Kleijnen (1994). This technique was applied to a military model by Leermakers (1993) and to an ecological model by Bettonvil and Kleijnen (1994).

4.2. Risk analysis

Sensitivity analysis shows which inputs are really important. Collecting information on those inputs is worthwhile. However, if it is impractical to collect reliable information on those inputs, then *risk analysis* may be applied. In such an analysis, a probability distribution of inputs is derived from the experts' knowledge. Next Monte Carlo sampling yields a probability distribution of output values. See Kleijnen (1994) and also Forrester and Senge (1980, pp. 225–226).

4.3. Gaussian approximation

Section 3.2 used a Gaussian approximation for the distribution of the *difference* between two *binomial* variables.

Brenner and Quan (1990) give an exact confidence intervals for a single binomial parameter p or q, not the difference $p-q$. They use the original binomial distribution, not the Gaussian approximation. Moreover, they do not follow the traditional approach that accounts for the discrete character of the binomial distribution and gives a conservative confidence interval. Instead they follow a Bayesian approach, assuming no prior information on the binomial parameter.

Louis (1981) discusses the special case of observing no successes ($x = 0$ or $y = 0$).

4.4. Type I and II errors

Given the sample sizes R and K, the type I error probability α, and the model error δ, it is possible to compute the β error probability; see Section 3.2. To decrease both error probabilities, it is necessary to increase the *sample sizes*. In this case study, R (number of simulation runs) may be increased; K (sample size of field test), however, is usually given.

4.5. Less stringent statistical validation

The null-hypothesis in (8) states that the simulated and the real detection probabilities are equal. Now, however, hypothesize that the *estimated* simulation and real probabilities are only *positively correlated*: if a mine has a relatively high estimated detection probability in the field run, then the estimated simulation probability should also be relatively high.

To test this hypothesis formulate the *regression model*

$$\hat{p}_{ih} = \beta_{0h} + \beta_{1h}\hat{q}_{ih} + \epsilon_{ih} \tag{17}$$

where ϵ_{ih} denotes 'white noise' (independently distributed Gaussian noise with mean zero and variance, say, σ_h^2). So if scenarios are measured, then plot $\hat{p}_{ih}$ as a function of $\hat{q}_{ih}$. Use ordinary least squares to estimate the intercept and slope of the straight line that passes through the 'cloud' of M points.

An *ideal* simulation model would mean that in this regression model the residuals are zero ($\epsilon_{ih} = 0$) so R^2 is one, while the intercept is zero and the slope is one.

Of course, such an ideal model is utopian. Therefore formulate the new *null-hypothesis*.

$$H_0: \quad \beta_{1h} \leq 0. \tag{18}$$

To test this hypothesis, use the standard t-statistic that is given in any textbook on regression analysis. So *reject* this null-hypothesis and *accept* the simulation model, if there is strong evidence that the estimated simulation and real detection probabilities are positively correlated (Kleijnen, Bettonvil, and Van Groenendaal, 1995).

There would indeed be M points to estimate the regression model (17), if the *scenario* could be kept constant during the whole field test. When scenarios are not fixed, then collecting all data in a single diagram creates extra noise (technically, the index h is deleted in Eqs. (17) and (18))

The weaker validation requirement of this subsection makes sense if the model is used to predict relative responses (as is the case in sensitivity analysis of tactics and sonar design), not absolute responses (needed to gauge the 'huntability' of a mine field). In the latter situation, input data of higher accuracy are necessary.

4.6. Animation

Animation may be used to present the simulation model and its results; see Kleijnen (1995) for references.

Animation may get naval experts involved in the model construction, verification, validation, and operational implementation.

4.7. Sound Velocity Profile (SVP)

In this particular case study, the SVP is a factor that certainly requires more research. In practice that factor is hard to measure sufficiently, since the SVP depends on time and place. In the model the SVP is treated as a *qualitative factor*. Such a nominal scale indicates lack of knowledge. Moreover, the simulation uses a single SVP per run, which is certainly unrealistic.

It would be useful to develop a real-time *measurement device* for SVPs and to install that device on board of the ship. Its measurements would

provide time and space dependent input to the simulation model, which would then become a decision support system (DSS). The Dutch Navy has acknowledged this need and has proceeded to acquire such a system.

Once the model is validated, it is a challenge to find *robust* mine sweeping procedures, which are not sensitive to the varying SVPs.

So validation must continue, as the model keeps changing (also see the beginning of Section 4).

4.8. Sea bottom

The bottom *profile* is another qualitative factor (see Section 4.7). The model uses a simple geometric pattern, whereas the real bottom is erratic (fractiles might be used to model that profile more realistically).

Moreover *bottom type* (sand, rock, etc.) is modeled crudely: bottom type is scaled from one to four, whereas it is actually a qualitative factor.

4.9. Navigational error

The simulated navigational error was found not to have the desired mean. Therefore navigational error may be modeled by specifying *positive correlation* as follows. Let y_t denote the actual ship's position at time t and e_t the navigational error at that time. Then specify

$$y_t = \mu + e_t, \tag{19}$$

where the error forms a time series

$$e_t = \rho e_{t-1} + z_t, \tag{20}$$

where ρ is the (positive) autocorrelation coefficient and z_t denotes a normally independently distributed variable with mean zero and variance (say) σ_z^2 such that e_t has the prespecified variance σ_e^2 (see Kleijnen and Van Groenendaal, 1992); in other words, the ship's position is a weighted average of the desired course μ and the previous position y_t, augmented with an independent normal error with zero mean:

$$y_t = (1-\rho)\mu + \rho y_{t-1} + z_t. \tag{21}$$

4.10. Measurement errors and data validity

Data validity in general is discussed in Knepell and Arangno (1993). In the *field runs* of this case study, a circle with a given radius is drawn around the location of the mine, assuming that location is exactly known. For validation purposes, the mine is supposed to be detected if and only if the operator records a contrast within that circle. Consequently, if the operator sees a false contrast (minelike object or spurious contrast) that falls within the circle, that echo is counted as a detection. On the other hand, a detection may be recorded outside the circle, and then it is not counted.

This procedure may be refined by giving higher *weights* to a recorded detection, the closer it lies to the true position of a mine. Until now weights were zero or one. (The weight function could be some bivariate distribution with means equal to the true coordinates and with such a shape that the weights decrease as specified by the naval experts. For example, with 90% probability a mine may be counted as being detected, if a recorded object lies no more than 20 meters from a true location. Multivariate distributions of many shapes are surveyed in Johnson 1987.)

4.11. Mine classification and destruction

The current model ends at the phase of mine detection, excluding the follow-up operations of classification and destruction (see Section 2).

In practice, any contrast that the operator interprets as a mine (even if that detection is caused by a minelike object or a spurious contrast) and that is 'close' to an actual mine, may become a success in the follow-up phase. However, calling 'mine!' all the time would generate a success probability of one; yet it would also waste much time and energy in the follow-up phase ('Peter and the wolf').

It seems better to separately measure *true* detections caused by whatever echo close to the true location of a mine, and *false* detections caused by minelike objects and spurious contrasts only (these detections resemble the type I and II errors in hypothesis testing; see Eq. (16)). True

and false detections should be measured, not only in the field runs but also in the simulation runs. In the current model, however, spurious contrasts are never counted as successes.

4.12. Other measures of effectiveness

This article has concentrated on the detection probabilities of the M individual mines; in practice, however, other responses are also of interest.

A closely related measure is the *average* detection probability per *strip* (strips and tracks were defined in Section 2). The detection probabilities are usually assumed to be the same for all mines within the same strip.

The probabilities per strip can be further aggregated into the *overall* detection probability for the *whole mine field*. Notice that in general, aggregation means loss of information. However, the width of the strips is debatable.

(Naval experts are interested in the *characteristic detection width* and the *characteristic detection probability*, which they denote by A and B. They derive these quantities from $p(v)$, the function that expresses the detection probability p as a function of (say) v, the athwart distance of the mine to the track. Obviously, this $p(v)$ generally decreases as v increases. They use the following equations to determine A and B:

$$W = \int_{-\infty}^{\infty} p(v)\, \mathrm{d}v,$$

$$\int_{-W_1}^{W_1} p(v)\, \mathrm{d}v = \tfrac{2}{3}W, \qquad (22)$$

$$A = 3W_1, \quad B = W/A.$$

To estimate these A and B, they process the estimated detection probabilities, once the simulation has been finished. (The estimators of A and B are negatively correlated, since $B = W/A$.) They collect all M estimated probabilities of a particular field test ($\hat{q}_i$) and their athwart distances (v_i), ignoring measurement errors of v. Since they further ignore the scenario (S_h), the resulting cloud of M observations (v_i, $\hat{q}_i$) is very erratic. It seems better to estimate $p_h(v)$ from the estimated probabilities *per* scenario; for example, a mine farther away from the ship has a smaller detection probability, *given* a certain SVP. To validate the simulated A and B, the actual scenarios should be measured in the field runs; see the discussion of Fig. 1. This estimation is possible, provided a real-time measurement device is installed on board of the ship; see Section 4.7.)

4.13. Calibration

Improvements of the current model should make it possible to eliminate the *artificial* calibration parameter, which was introduced to get better fit between simulated and field results (see Section 2).

5. Summary and conclusions

This article discussed a *case study*, namely a simulation model of mine hunting at sea, developed by TNO-FEL for the Dutch navy. This model, called HUNTOP, includes the environment (namely, the mine field and acoustical characteristics of the sea water), the sonar system, the ship's course, and the human operator's performance.

Simulation models can be validated in *two stages*. Since no data were available for individual modules of HUNTOP, *sensitivity analysis per module* was performed in stage #1. This analysis can use experimental design theory and regression analysis. The results for two modules (sonar window and object visibility) corroborate the validity of these HUNTOP modules, since the input/output behavior of these modules agreed with the experts' qualitative knowledge.

In stage #2 the *model as a whole* can be validated. So simulated detection probabilities were compared with real-life probabilities. A statistic was derived to test the null-hypothesis of equal expectations for estimated simulated and real probabilities. It was emphasized that it is important to measure the environmental *scenarios* that drive the simulation and the field test respectively.

Finally a review followed, discussing *remaining issues* in the validation of simulation models in

general and in naval mine-hunting models in particular.

Acknowledgements

Gustav A. Alink, a former employee of FEL-TNO, was a tremendous help during the whole project that is described in this paper. Marcel Das, a Tilburg doctoral student, gave valuable comments on a preliminary draft of this paper.

References

Alink, G.A., and Vermeulen, J.F.J. (1991), "Validation of the Mine Hunting Model HUNTOP", Report No. 1 FEL-91-A096 (confidential, except for abstract), TNO Physics and Electronics Laboratory, The Hague.

Balci, O. (1994), "Validation, verification, and testing techniques throughout the life cycle of a simulation study", *Annals of Operations Research*.

Balci, O. (1995), "Principles of simulation model validation, verification, and testing", *International Journal in Computer Simulation*.

Banks, J., and Carson, J.S. (1984), *Discrete-event System Simulation*, Prentice-Hall, Englewood Cliffs, NJ.

Bettonvil, B., and Kleijnen, J.P.C. (1994), "Identifying the important factors in simulation models with many factors", Tilburg University, Tilburg, Netherlands.

Brenner, D.J., and Quan, H. (1990), "Exact confidence intervals for binomial proportions – Pierson and Hartley revisited", *Statistician* 3, 391–397.

Fleming, R.A., and Schoemaker, C.A. (1992), "Evaluating models for spruce budworm-forest management: Comparing output with regional field data", *Ecological Applications* 2/4, 466–477.

Forrester, J.W., and Senge, P.M. (1980), "Tests for building confidence in system dynamics models", in: A.A. Legasto, J.W. Forrester, and J.M. Lyneis (eds.), *System Dynamics*, North-Holland, Amsterdam, 209–228.

Fossett, C.A., Harrison D., Weintrob H., and Gass, S.I. (1991), "An assessment procedure for simulation models: A case study", *Operations Research* 39, 710–723.

Johnson, M.E. (1987), *Multivariate Statistical Simulation*, Wiley, New York.

Kleijnen, J.P.C. (1987), *Statistical Tools for Simulation Practitioners*, Dekker, New York.

Kleijnen, J.P.C. (1994), "Sensitivity analysis versus uncertainty analysis: When to use what?", in: J. Grasman and G. van Straten (eds.), *Predictability and Nonlinear Modeling in Natural Sciences and Economics*, Kluwer, Dordrecht, Netherlands, 322–333.

Kleijnen, J.P.C. (1995), "Verification and validation of simulation models", *European Journal of Operational Research* 82/1, 145–162.

Kleijnen, J.P.C., Bettonvil, B., and Van Groenendaal, W. (1995), "Validation of simulation models; regression analysis revisited", Tilburg University, Tilburg, Netherlands.

Kleijnen, J.P.C., and Van Groenendaal, W. (1992), *Simulation: A Statistical Perspective*, Wiley, Chichester, UK.

Knepell, P.L., and Arangno, D.C. (1993), *Simulation Validation: A Confidence Assessment Methodology*, IEEE Computer Society Press, Los Alamitos, CA.

Landry, M., and Oral, M. (1993), "In search of a valid view of model validation for operations research", *European Journal of Operational Research* 66/2, 161–167.

Law A.M., and Kelton W.D. (1991), *Simulation Modeling and Analysis*; Second Edition, McGraw-Hill, New York.

Leermakers, W. (1993), "Gevoeligheidsanalyse van modellen met veel parameters: een case-study", (Sensitivity analysis of models with many parameters: a case study), TNO Physics and Electronics Laboratory, The Hague, December 1993.

Louis, T.A. (1981), "Confidence intervals for a binomial parameter after observing no successes", *American Statistician* 35, 154.

Miller, R.G. (1981), *Simultaneous Statistical Inference*, Revised Edition, Springer-Verlag, New York.

Pegden, C.P., Shannon, R.E., and Sadowski, R.P. (1990), *Introduction to Simulation using SIMAN*, McGraw-Hill, New York.

Sargent, R.G. (1991), "Simulation model verification and validation", *Proceedings of the 1991 Winter Simulation Conference*, 37–47.

Van Groenendaal, W. (1994), "Investment analysis and DSS for gas transmission on Java", Tilburg University Press, Tilburg.

Wang, W., Yin, H., Tang, Y., and Xu, Y. (1993), "A methodology for validation of system and sub-system level models", Department of System Engineering and Mathematics, National University of Defense Technology, Changsha, Hunan, 410073, P.R. China.