Tests for Confidence Building

6

The models developed based on the concepts, methodology and techniques of system dynamics discussed in the earlier chapters must be tested to build up confidence in the models. This chapter presents the tests for confidence building in the system dynamics models. These tests are discussed under the broad heading of tests for structure, tests for behaviour and tests for policy implications. The logical sequences of conducting these testes are also presented.

6.1 Introduction

Once we have developed a model, how can we develop our confidence in the use of the model for developing scenarios and management strategies? How can we make others trust our model? This chapter deals with the tests for confidence building in system dynamics models. Model validation to develop confidence in the model is important, but it is a controversial aspect of any process-based model in general and system dynamics (Barlas 1996). The validity and usefulness of dynamic models should be judged, not against an imaginary perfection but in comparison with the mental and descriptive models which we would otherwise use (Forrester 1968). Models should be judged, not on absolute scale but on relative scale. If they succeed in clarifying our knowledge and insights into the systems for better understanding and management, the model should be accepted.

Tests for building confidence in system dynamics models essentially consist of validation, sensitivity analysis and policy analysis of the system dynamics models. The two important notions of building confidence in system dynamics models are testing and validation of system dynamics models. Testing means the comparison of a model to empirical reality for the accepting or rejecting the model, and validation means the process of establishing confidence in the soundness and usefulness of the model.

In testing mode, the model structures are compared directly to descriptive knowledge of real system structures, and model behaviour may be compared to

B.K. Bala et al., *System Dynamics*, Springer Texts in Business and Economics,
DOI 10.1007/978-981-10-2045-2_6

observed real system behaviour. In validation mode, the model behaves plausively and generates problem symptoms or modes of behaviour observed in the real world. The modeller's confidence needs to be transferred to the target audience.

Validation is complicated by many relevant audiences. For a scientist, a model is useful if it generates insight into the structure of real system, makes correct prediction and stimulates meaningful questions for future research. For the public and political leaders, a model is useful if it explains the causes of important problems and provides a basis for designing policy to improve the behaviour of the system. Validity meaning confidence in a model's usefulness is inherently relative concepts. One must choose between competing models.

In system dynamics model, behaviour testing is very common. This is because very often system dynamics models incorporate such variables for which no real-life data is available. Under such circumstances, assumed relationships are based on available literatures, and the appropriateness of such relationships is justified in the overall context of the generated model behaviour. Though system dynamics modellers try to make best use of the available data for parameter estimation, sometimes by statistical methods, vigorous use of statistical tests is rare.

The main reason why system dynamics methods are not statistically tested is that the system models include variables for which no real-life data exists. Since system dynamics models include both statistically valid variables and assumed variables, it cannot be claimed that the model behaviour would deteriorate because of the use of statistically non-tested variables. In fact, the use of the assumed variables would improve the behaviour generation capacity of the model, and it may be claimed that the model has passed improvement test. In situation where the modeller has clear idea and supporting data of the mode of behaviour of the system, the model building and validation becomes relatively an easy task.

The ultimate objective of model validation is to develop confidence in predictions. The first step of the ultimate objective of system dynamics model validation is to establish the validity of the model structure. The next step is the behaviour reproduction of the model compared to the real behaviour of the system, and it is meaningful only if we have sufficient confidence in the model structure (Barlas 1996). The logical sequences of model validation are shown in Fig. 6.1. The model behaviour changes to the change in parameter values, and policy issues are also important to understand how the model will behave under changing conditions. Keeping these philosophical aspects in mind, the tests for confidence building in system dynamics models are to be designed. The tests for building confidence in system dynamics models may be broadly classified as:

1. Tests for structure
2. Tests for behaviour
3. Tests for policy implications

One must realise that not all the tests are to be considered for validation of a model, but structure and behaviour pattern tests are essential, and policy

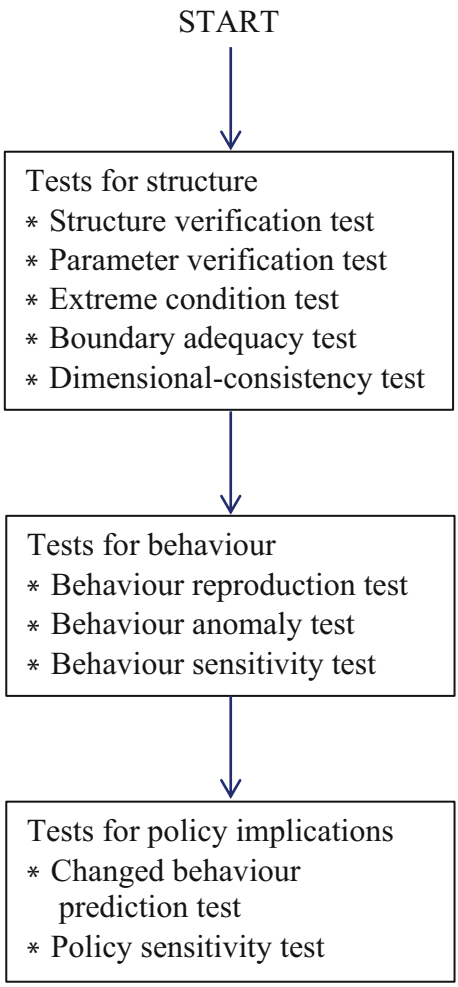

Fig. 6.1 Logical sequences of the tests for model validation

implication test makes the validation sufficient. It is rather those tests which are essential for establishing the creditability of the model are to be included.

6.2 Tests of Model Structure

The first step in the validation of system dynamics models is the structure validity tests, and this can be further classified as direct structure tests and structure-oriented behaviour tests. In direct validation structure tests, the validity of the model structure is assessed by direct comparison of the model structure with the knowledge of the real systems. It is accomplished by comparing mathematical equations and logical relationship with the available knowledge of the real systems. No

simulation of the model is needed. Direct structure tests can be classified as empirical tests and theoretical tests. Empirical tests are conducted by comparing the model structure with the information (quantitative and qualitative) obtained about the system, while theoretical tests are conducted by comparing the model structure with the generalised knowledge of the system from literature such as research reports and studies. Structure confirmation tests are the toughest tasks to do as we need to compare the equations of the model directly with the knowledge of the real systems.

In structure-oriented behaviour tests, the validity of the model structure is assessed by the comparison of the behaviour of model predicted with the knowledge of behaviour of the real systems expected and usually observed in reality. It is a qualitative validation of the model.

Broadly speaking, tests of model structure may be classified as:

1. Structure verification test
2. Parameter verification test
3. Extreme condition test
4. Boundary adequacy test
5. Dimensional consistency test

6.2.1 Structure Verification Test

The structure verification test applies as empirical means comparing the form of equations of the model with relationships that exist in the real systems. The structure of the model that is the relationships in the equations should be in line with the descriptive knowledge of the system. It may also be conducted as theoretical tests by comparing the model equations with the generalised knowledge of the systems in the literatures. All the equations should be well argued and based on available information. The structure of the model should match observable goals, pressures and constraints of the real systems. Verifying a model structure is an easier task and takes less skill than some other tests.

6.2.2 Parameter Verification Test

The second structure verification test is the parameter confirmation test, and it means evaluating the constant parameters against the knowledge of the real systems both conceptually and numerically. Every constant (and variable) should have a clear real-life meaning. The basic choice is formal statistical estimation or judgemental estimation. Econometrics or other methods may be used to estimate the parameters.

The choice of appropriate initial values for stock equations, values of constants and table functions is directly related to the model description, and the values should be based on the published data from various sources. Computer software

packages are now available to estimate and justify the exact values of the parameters to produce the expected behaviour of the system. Structure verification and parameter verification are interrelated, and both tests have the same basic objective.

6.2.3 Extreme Condition Test

The model should be robust under extreme conditions. There is an important direct structure test to the robustness of the model under direct extreme conditions, and it evaluates the validity of the equations under extreme conditions by assessing the plausibility of the resulting values against knowledge/anticipation of what would happen under similar conditions in real life. It is relatively easy to anticipate which variables and what values would these variables take under extreme condition in real systems.

The model must be capable to cope with external conditions. If the extreme conditions are incorporated in the model, the result is an improved model in the normal operating region. System dynamics model structure permits extreme combinations of socks in the system under study. A model should be questioned if extreme condition test is not met. It is not acceptable that extreme condition is not necessary on the plea that it does not occur in real life. Extreme condition test is effective for two reasons: (a) it is a powerful tool to detect the defect in model structure, and (b) it enhances the usefulness of the model for analysing policies that may force a system to operate outside historical regions of behaviour. Hence the extreme condition test is a strong test.

Let us consider that the supply chain model of rice milling systems is to be tested for extreme conditions to detect the defect in the model structure and enhance the usefulness of the model for policy analysis. Figure 6.2 shows simulated milling inventory, wholesale inventory and retail inventory under extreme condition of crop failure, i.e. zero crop production. Under this condition, the milling inventory and then wholesale inventory and retail inventory are reduced to zero since the rice production is zero. The model results confirmed to the expected patterns of results and realities. This model complied with the basic principles of supply chain management and was consistent with supply chain theory and research results. Thus, the model is able to provide qualitative and quantitative understanding of the supply chain performances of rice milling systems. Hence the model is reliable and validated under extreme conditions.

6.2.4 Boundary Adequacy Test

Boundary adequacy test considers structural relationships necessary to satisfy the model's purpose. Boundary adequacy test asks whether or not model aggregation is appropriate and if a model includes all relevant structure.

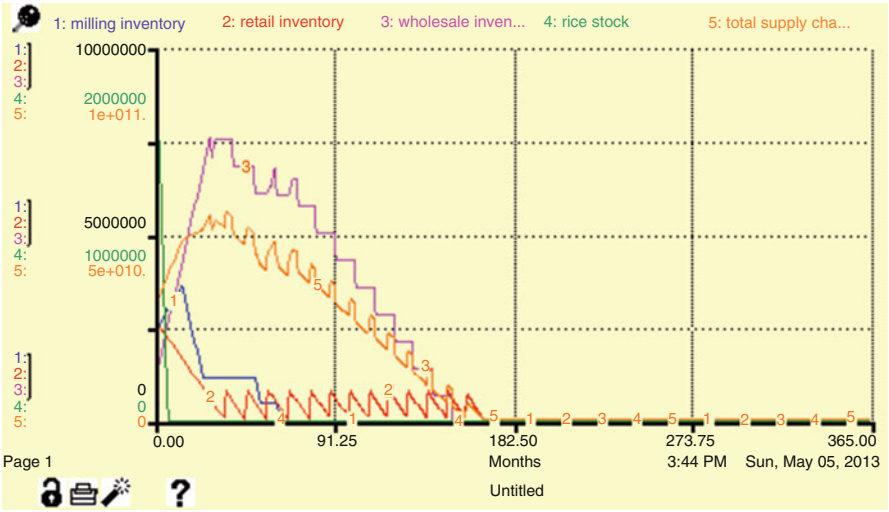

Fig. 6.2 Simulated milling inventory, wholesale inventory and retail inventory under extreme condition of crop failure, i.e. zero crop production

Once the model boundary is established, it is necessary to check whether any additional feedback loop has been omitted or not. If the additional feedback has any significant impact, it must be included. In essence, the model must include all variables and feedback loops encompassing the entire system under study which affect the dynamics or policy implications of the model.

6.2.5 Dimensional Consistency Test

Dimensional consistency test is one of the basic tests, and it must be carried out during the construction of the model. Dimensional consistency test involves checking the right-hand side and left-hand side of each of the equations of the model for dimensional consistency. It is better to specify the units of measure of each variable during construction of the model. The dimension of the left-hand and right-hand side of an equation should be the same for correct model formulation. Moreover, the dimensions of the variables should be close to their physical meaning; none of the dimension should be divorced from its actual meanings. Dimensional consistency test entails the dimensional analysis of a model's rate equation. Surprisingly many models fail this simple test. Hence, dimensional consistency test is the most powerful when applied in conjunction with the parameter verification test.

6.3 Tests of Model Behaviour

The second important step in the validation of system dynamics models is the model behaviour validity tests, and the tests for model behaviour should be conducted once the structural validation tests are completed successfully. The core tests of model behaviour may be classified as:

1. Behaviour reproduction test
2. Behaviour anomaly test
3. Behaviour sensitivity test

6.3.1 Behaviour Reproduction Test

Once the structure confirmation tests are completed successfully, the next test is the behaviour pattern tests to measure how accurately the model can reproduce the dynamic behaviour of the real systems. Behaviour reproduction tests compare how well the model-generated behaviour matches model-observed behaviour of the real system. Behaviour reproduction tests include symptom generation, frequency generation, relative phasing, multiple mode and behaviour characteristic. Behaviour reproduction tests become much more convincing when one can show why the tests are passed.

Many tools are available to assess the model behaviour to reproduce the system behaviour. Most common techniques are descriptive statistics to measure point by point fit. The most commonly used measure of the fit is the coefficient of determination (R^2), and it measures the fraction of variance explained by the model. The mean absolute error (MAE), mean absolute percent error (MAPE) and root mean square error (RMSE) all provide measures of the average error between the simulated and actual values. However, the emphasis should be on pattern rather than point prediction.

In the literature on modelling and simulation, there are a wide range of tests involving point by point comparisons of model-generated and model-observed behaviour. Despite widespread acceptance, such tests involving point by point measures of goodness of fit are generally less appropriate for socio-economic system dynamics models.

The reproduction of historical behaviour is the single most important test to build up confidence in models. Figure 6.3a shows the comparison between the predicted and historical behaviour of food self-sufficiency ratio in Malaysia. The model-simulated food self-sufficiency ratio agrees reasonably well with historical behaviour, and the model is reliable. Figure 6.3b shows the comparison of simulated and reported changes in wholesale price of rice in 2011 in Bangladesh. The model can simulate the actual behaviour of the system closely and can be used for policy analysis.

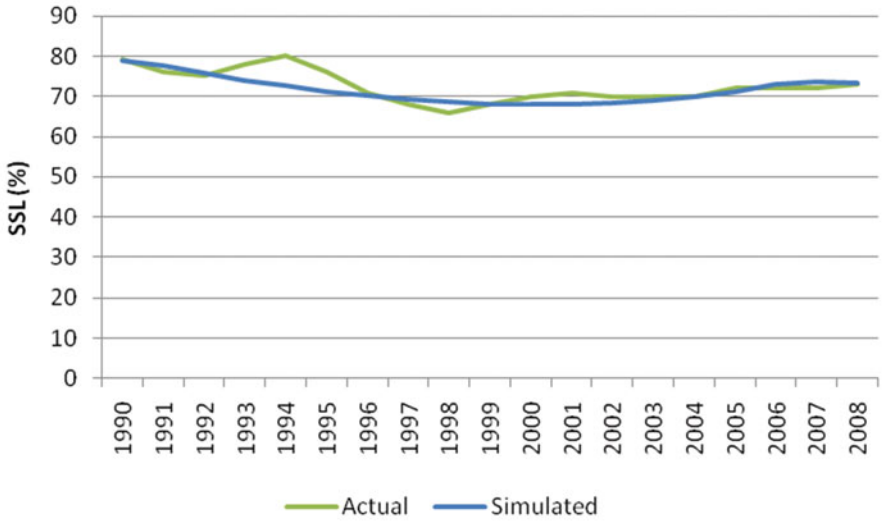

Fig. 6.3a Simulated and historical data of food self-sufficiency ratio in Malaysia

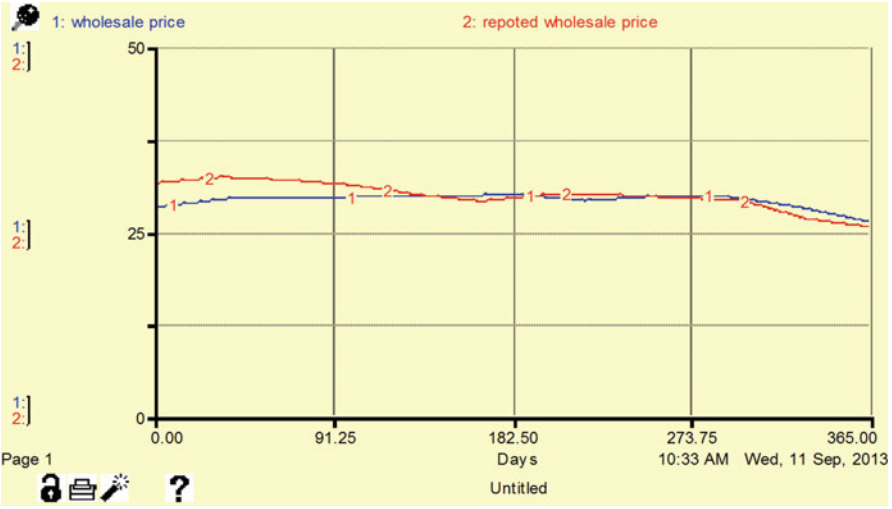

Fig. 6.3b Comparison of simulated and reported changes in wholesale price of rice in 2011 in Bangladesh

6.3.2 Behaviour Anomaly Test

While simulating a system dynamics model, one expects it to behave like real system under study, but frequently model builders face anomalous features of

model behaviour, and these contradict the behaviour of the real system. Whenever there is anomaly in model behaviour, there may be defect in model assumptions. One can often defend particular assumptions by showing how implausible behaviour arises if the assumption is altered. Loop knockout analysis is a common method to search for behaviour anomalies. Anomalous behaviour resulting from knockout test suggests the importance of the loop and may help to establish the plausibility of system behaviour.

6.3.3 Behaviour Sensitivity Test

The behaviour sensitivity test shows the sensitivity of the model behaviour to changes in parameter values. The parameter sensitivity test ascertains whether or not plausible shifts in parameter values cause a model to fail behaviour tests previously passed.

The behaviour sensitivity test is typically conducted by experimenting with different parameter values and analysing their impacts on behaviour. Typically, the behaviour of system dynamics models is insensitive to plausible changes in most parameter values. It appears that systems are insensitive. On the other hand, both real systems and models or real systems are sensitive to a few parameters. Finding a sensitive parameter does not necessarily invalidate the model. Even though it has a substantial effect on behaviour, plausible variations may not lead to failure of other behaviour tests.

The model of supply chain of rice milling systems was simulated to address the impacts of rice productivity on the supply chain performances. Here the rice productivity is the yield of rice per ha. Rice productivity may be reduced from crop damage due to floods or pest infestation, and also it can be increased by development of higher yield hybrid rice through research and development. Rice productivity for this policy is defined as

$$\text{rice production rate} = \text{area under rice} \times \text{yield of rice} \tag{6.1}$$

$$\text{yield of rice} = 1.81, 2.81 \text{ and } 3.81 \text{ tons/ha} \tag{6.2}$$

Simulated milling inventory, wholesale inventory, retail inventory and total supply chain cost for rice productivity of 1.81 tons per ha, 2.81 tons per ha (present average rice productivity) and 3.81 tons per ha are shown in Figs. 6.4a, 6.4b, 6.4c and 6.4d, respectively. Milling inventory is reduced to zero for most of the period of the reduced productivity of rice (1.81 tons per ha), while it is high for bumper production of rice (3.81 tons per ha) (Fig. 6.4a). Wholesale inventory is reduced significantly in the fourth quarter of the year for reduced rice productivity (1.81 tons per ha) (Fig. 6.4b). Total supply chain cost is also reduced in the third and fourth quarter of the year for reduced rice productivity (Fig. 6.4d). However, in all the cases, the retail inventory is almost the same except towards the end of the year when both milling and wholesale inventories are empty for reduced productivity

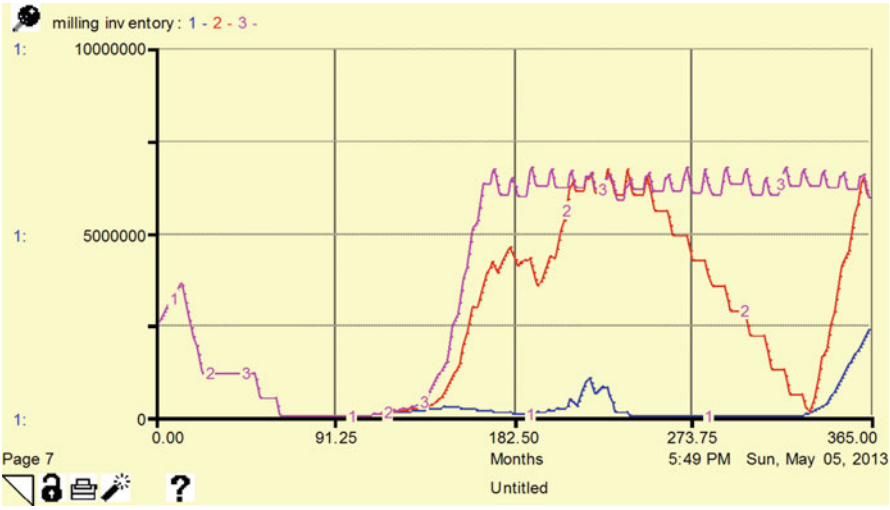

Fig. 6.4a Simulated milling inventory for rice productivity of 1.81 tons per ha, 2.81 tons per ha and 3.81 tons per ha

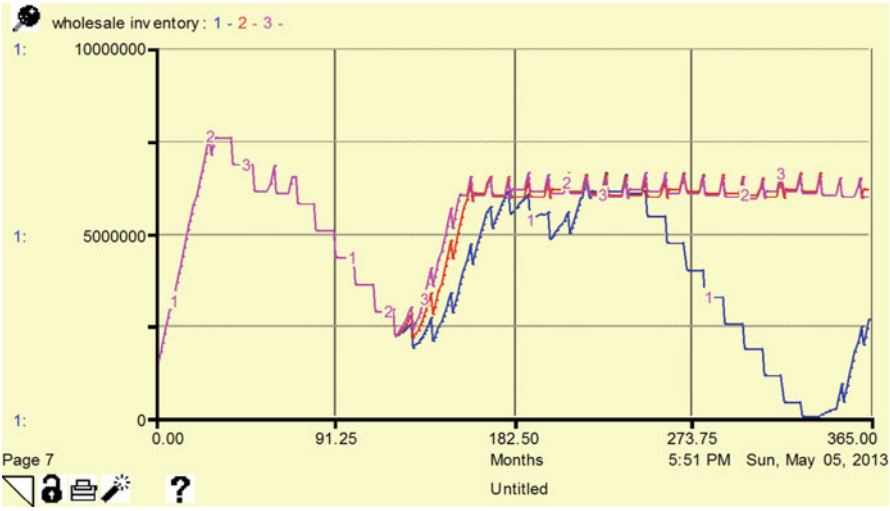

Fig. 6.4b Simulated wholesale inventory for rice productivity of 1.81 tons per ha, 2.81 tons per ha and 3.81 tons per ha

(Fig. 6.4c). Thus, increased wholesale inventory is a possible solution for the retail inventory to face the shortage of rice during the off-peak harvesting season of rice production. This implies that as long as wholesale inventory is available, the retail inventory is stabilised based on economic order quantity and reordering point

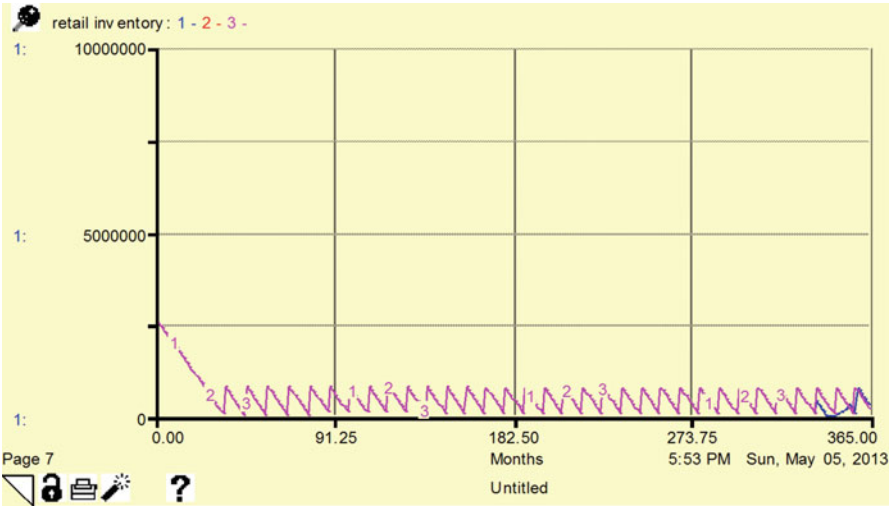

Fig. 6.4c Simulated retail inventory for rice productivity of 1.81 tons per ha, 2.81 tons per ha and 3.81 tons per ha

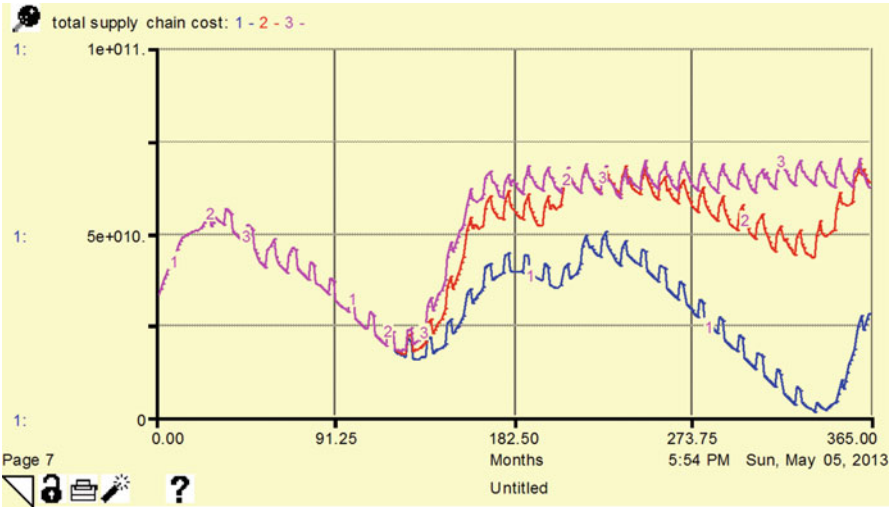

Fig. 6.4d Simulated total supply chain cost for rice productivity of 1.81 tons per ha, 2.81 tons per ha and 3.81 tons per ha

operation of milling, wholesale and retail inventory. This demonstrates that the policy based on economic order quality and reordering point can ensure the availability of rice even under reduced production of rice, i.e. during crop damage/failure unless both wholesale and milling inventories are empty.

6.4 Tests of Policy Implications

Tests should be conducted to build confidence in model's implications for policy. The core tests of policy implications may be classified as:

1. Changed behaviour prediction test
2. Policy sensitivity test

6.4.1 Changed Behaviour Prediction Test

The changed behaviour prediction test shows how well the model predicts the behaviour of the system if a governing policy is changed. The test can be conducted by changing policies in the model and verifying the plausibility of resulting behaviour changes. Alternatively one can examine the response of the policy already pursued to see how well model response agrees with the real system response. This test essentially shows the impacts of exogenous policies on the model behaviour.

Figure 6.5 shows the comparison of food self-sufficiency ratios for basic run and IPCC climate scenario for base year yield and yield increase of 6 tons per ha within next 50 years. The climate change impacts on food self-sufficiency level are small for all these runs. However, food self-sufficiency levels for both the base year yield and the yield increase of 6 tons per ha follow similar patterns, and in both the cases, the food self-sufficiency levels increase for about 10 years due to expansion

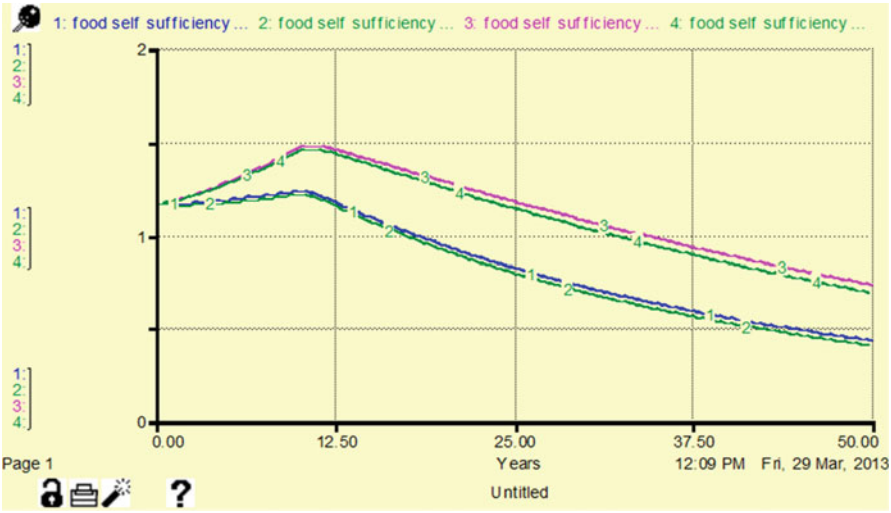

Fig. 6.5 Comparison of food self-sufficiency ratios for basic run and IPCC climate scenario for base year yield and yield increase of 6 tons per ha withinTests of policy implications:Changed behavior prediction test next 50 years

irrigation, and then it decreases as a result of the discard of irrigated area for its use for infrastructure development. The improved productivity increases the self-sufficiency level for about 12.5 years ahead, and the food self-sufficiency at the end of 50 years of simulation period increases from 43 to 73 %. Food self-sufficiency level is more seriously challenged by the decreasing cultivable land and growing population, and these are essentially demanding more increase in the productivity in the vertical direction due to the constraints of non-availability of additional cultivable irrigable land and more attention to control the growing population to improve rice self-sufficiency in the long run.

6.4.2 Policy Sensitivity Test

Researchers and decision makers have to make up their minds about where to concentrate their efforts to improve policies. When policy improvement is the desired objective, the question of policy sensitivity arises, although it may not be recognised as such. What kind of researchers should be involved? What mechanisms should be included or left out of formal and mental models? For which relationships and parameters should one seek better data and higher-quality estimates? These questions are best answered by policy sensitivity analysis.

The traditional and frequently used form of sensitivity analysis in system dynamics is to vary model assumptions and to observe how behaviour changes. In the branch of operations research using optimisation, sensitivity analysis is to vary model assumptions and to observe how optimal policies change. To avoid confusion between these two types of sensitivity, the terms behaviour sensitivity and policy sensitivity are used, of which the latter is the focus here. Policy sensitivity exists when a change in assumptions reverses the impacts or desirability of a proposed policy (Sterman 2000). For example, when one set of assumptions causes sustainable supply of palm oil, while another does not, the model exhibits policy sensitivity. If a particular policy change always produces improvement, regardless of changes in a sensitive parameter, then the policy recommendation is not affected (Forrester 1969). For example, when both sets of policy assumptions produce improvement of food security, the model exhibits policy insensitivity. These two statements clearly support and emphasise the sensitivity of the outcome of particular policies to uncertain parameters. To distinguish the two types of sensitivity analysis, we denote the latter policy outcome sensitivity.

Policy outcome sensitivity analysis is the most appropriate type of analysis when there remains uncertainty about parameter assumptions. This type of analysis can be expanded to include risk. The policy sensitivity we focus on here is the most appropriate for the purpose of model building. What parameters are most important for policy recommendations and require thorough analysis? What simplifications and aggregations are important for policies?

From a more practical viewpoint, if the simplified policy is the best one can do, or it is the type of policy that will be used in practice, then the bias is of less concern. Then it is interesting in itself to see how the optimised practical policy varies with changes in model assumptions. As a problem is demonstrated not to be

very sensitive to uncertain parameters, decision makers' confidence in the problem formulation increases and so does the likelihood that some first policy measures will be implemented. That this type of analysis is needed in renewable resources management, e.g. fisheries, is indicated by the long delays in implementing appropriate policies and by laboratory experiments showing tendencies towards misperception.

We start with the problem of model validation, where a major challenge is to choose between potentially large numbers of models that pass tests of falsification. Policy sensitivity analysis is of no direct help in this choice. However, it can be used to find out if the choice of model has policy implications. If it has not, the remaining uncertainty is mostly of academic interest.

This is a potentially important insight since a heated 'academic' debate about the correctness of models can confuse the policy debate. In principle, policy sensitivity analysis could even be used to investigate the importance of the 'unavoidable a priori' assumptions of different disciplines (Meadows 1980). If different disciplines prescribe different policies, policy sensitivity analyses could be used to identify the causes of disagreement and to direct further validation effort towards the identified causes.

Next, we consider model aggregation. In principle, aggregation is a very challenging problem. Most system dynamics models resort to aggregation of people, resources, perceptions, etc. On the other hand, in non-linear dynamic models, aggregation leads to model errors except in some very special cases (Krysiak and Krysiak 2002). For this reason, Krysiak and Krysiak argue that 'environmental economics as well as other fields of economics may benefit from using more complex models'. On the other hand, increased complexity has a cost in terms of efforts to validate, analyse and explain models. Policy sensitivity analysis can be used to identify the appropriate aggregation level for policies.

Finally, policy sensitivity test shows the degree of robustness of model behaviour and policy recommendations. Such testing can help to show the uncertainty in parameter values. In the worst case, the parameter change can invalidate the recommended policies that were given. However, the policy recommendations are not likely to be affected by uncertainties in parameters.

In summary, tests for building confidence in system dynamics models should be conducted in some logical sequences, and we should move one step to the next step only if we are able to establish sufficient confidence in the current step. The logical sequences of formal steps of model validation as suggested by Barlas (1996) capture the essentials of the validation of system dynamics models to build up confidence in system dynamics models. The logical sequences of formal steps of model validation as suggested by Barlas (1996) are shown in Fig. 6.6. Once the model has passed through structural tests, we should start behaviour pattern tests. If the model passes through both direct and indirect structural tests, but fails the behaviour pattern tests, then we should re-estimate certain parameter and/or input function.

Next it would be logical to skip the structural tests and apply behaviour pattern tests. Once we have reached the final step of the behaviour pattern test, we should

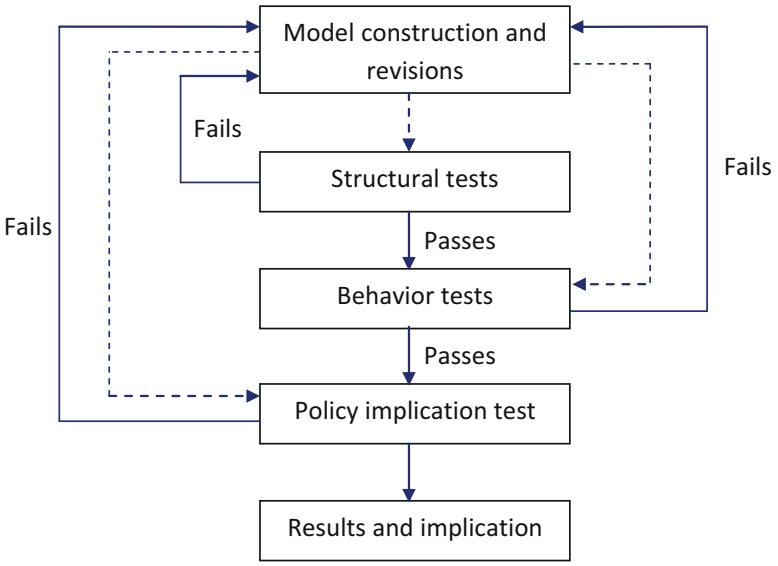

Fig. 6.6 Logical sequences of formal steps of model validation as suggested by Barlas (1996)

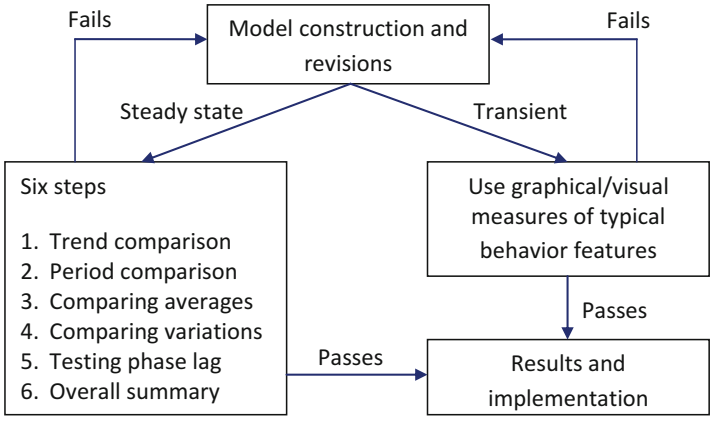

Fig. 6.7 Logical sequence of behaviour pattern tests

give emphasis on the accuracy of the pattern predictions. However, behaviour pattern tests are weak tests that provide no information on the validity of the structure of the model. Behaviour pattern tests must be carried out in logical order too. Figure 6.7 shows the logical sequence of behaviour pattern tests. If the problem involves transient and highly nonstationary behaviour, it is not possible to

apply any standard statistical measure. On the other hand, if the problem involves a long-term simulation, it is possible to apply standard statistical measures and tests. Note that if the model is considered to fail the behaviour pattern tests, once again we should go back to model revisions, and the model revisions involve parameter/input changes rather any other tests.

Exercises

Exercise 6.1 What are the tests for confidence building in system dynamics models? Describe the logical sequences of the tests of model validation.

Exercise 6.2 Describe tests of model structure with examples. Discuss the importance of verification tests. Explain why extreme condition test is important with examples.

Exercise 6.3 Describe tests of model behaviour with examples. Discuss the importance of behaviour reproduction tests with examples and also show logical sequence of behaviour pattern tests.

Exercise 6.4 Describe tests of policy implications with examples. Discuss the importance of policy sensitivity test.

Exercise 6.5 Describe the logical sequences of formal steps of model validation suggested by Barlas (1996).

References

Barlas Y (1996) Formal aspects of model validity and validation in system dynamics. Syst Dyn Rev 12(3):183–210

Forrester JW (1968) Principles of systems. MIT Press, Cambridge, MA

Forrester JW (1969) Urban dynamics. MIT Press, Cambridge, MA

Krysiak FC, Krysiak D (2002) Aggregation of dynamic systems and the existence of a regeneration function. J Environ Econ Manag 44:517–539

Meadows DH (1980) The unavoidable a prior. In: Randers J (ed) Elements of system dynamics method, Productivity Press, Portland

Sterman JD (2000) Business dynamics: systems thinking and modelling for a complex world. Irwin/Macgraw Hill, Boston

Bibliography

Bala BK (1999) Principles of system dynamics, First editionth edn. Agrotech Publishing Academy, Udaipur

Maani KE, Cavana RY (2000) Systems thinking and modelling: understanding change and complexity. Prentice Hall, Auckland

Mohapatra PKJ, Mandal P, Bora MC (1994) Introduction to system dynamics modelling. Universities Press, Hyderabad

Moxnes E (2005) Policy sensitivity analysis: simple versus complex fishery models. Syst Dyn Rev 21(2):123–145