4 Model Validation

4.1 General Concepts

From a methodological point of view, it is widely accepted that simulation is a useful technique when it comes to providing an experimental test bed in which to compare different system designs and replace experiments in the physical system with experiments that involve their formal representation in a computer by means of a simulation model. The outcome of the computer experiment thus provides the basis for quantitative support for decision-makers. According to this conception, the simulation model can be seen as a computer laboratory where experiments can be conducted with the model of the system, with the aim of drawing valid conclusions for the real system. In other words, the simulation model is used to answer "What if?" questions about the system.

Simulation may thus be seen as a *sampling experiment* of the real system through its model (Pidd, 1992). In other words, assuming that the evolution over time of the model correctly imitates the evolution over time of the system modelled, samples of the observational variables of interest are collected, from which conclusions on the system's behaviour can be drawn using statistical analysis techniques. Figure 4.1 illustrates this method conceptually.

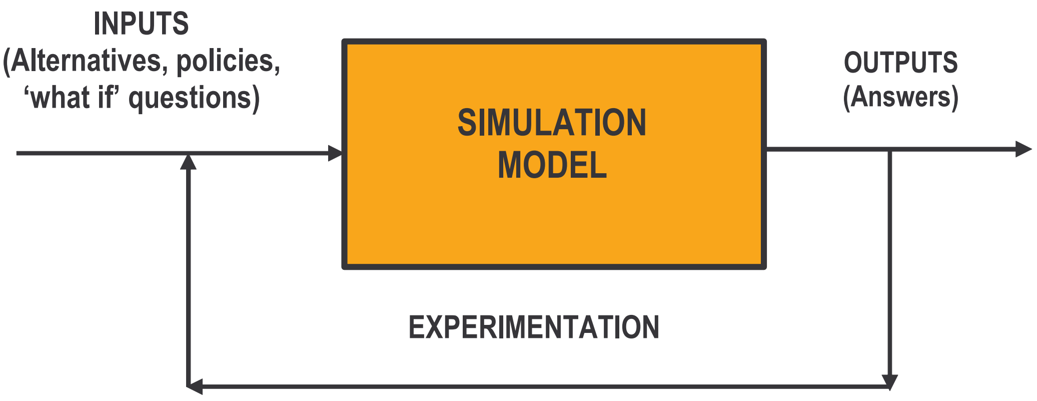

Figure 4.1. Experimental nature of the simulation

The reliability of this decision-making process depends on the ability to produce a simulation model that represents the system's behaviour closely enough for the model to be used as a substitute for the actual system for experimental purposes. This is true for any simulation analysis in general and obviously for traffic simulation. The process of determining whether the simulation model is close enough to the actual system is usually achieved through the validation of the model, an iterative process that involves calibrating the model's parameters and comparing the model to the actual system's behaviour. The discrepancies between the two and the insight gained are used to improve the model until its accuracy is thought to be acceptable.

The validation of a simulation model is a concept that should be taken into account throughout the model-building process.

According to Law and Kelton (1991), the key methodological steps for building valid and credible simulation models are the following:

- **Verification**, which consists in determining that a computer simulation program performs as intended and is concerned with *building the model properly***.**
- **Validation**, which consist in determining whether the conceptual simulation model (as opposed to the computer program) is an *accurate representation of the system under study.* Validation involves *building the right model.*
- A model is *credible* when its results are accepted by the user and are used as an aid in making decisions. Animation is an effective way for an analyst to establish credibility.

Balci (1998) defines a successful simulation study as "(…) the one that produces a sufficiently credible solution that is accepted and used by decision makers". This implies the assessment of the quality of the simulation model through the verification and validation of the simulation models.

Verification usually implies running the simulation model under a variety of input parameter settings and checking to see whether the output is reasonable. In some cases, certain measures of performance may be computed exactly and used for comparison. Animation can also be of great help for this purpose. With certain types of simulation models (traffic models are a good example), it may be helpful to observe an animation of the simulation output to establish whether the computer model is working as expected. In validating a simulation model the analyst should not forget that

- A simulation model of a complex system can only be an *approximation* to the actual system. *There is no such thing as an absolutely valid model of a system*.
- A simulation model should always be developed for a particular set of purposes.
- A simulation model should be validated relative to those measures of performance that will actually be representative of these purposes.
- Model development and validation should be carried out alongside each other throughout the entire simulation study.

Validation means the process of testing the model to see if it does actually represent a viable and useful alternative means to real experimentation. This requires *calibrating the model*, that is, adjusting model parameters until the resulting output data agree closely with the observed system data. The validation of the simulation model will be established on the basis of the comparison analysis of the observed output data from the actual system and the output data provided by the simulation experiments conducted with the computer model.

Model calibration and validation is inherently a statistical process in which the uncertainty due to data and model errors should be accounted for. Depending on the variables selected, the system and simulated data available, their characteristics and statistical behaviour, a variety of statistical techniques either for paired comparisons or for multiple comparisons and time series analysis, have been proposed. The conceptual framework for this validation methodology is described in the diagram in Figure 4.2 (adapted from Balci, 1998). According to this reasoning, when the results of the comparison analysis are not acceptable to the degree of significance defined by the analyst, the rejection of the simulation results implies the need for recalibrating certain aspects of the simulation model. The process is repeated until a significant degree of similarity, according to given statistical analysis techniques, is achieved.

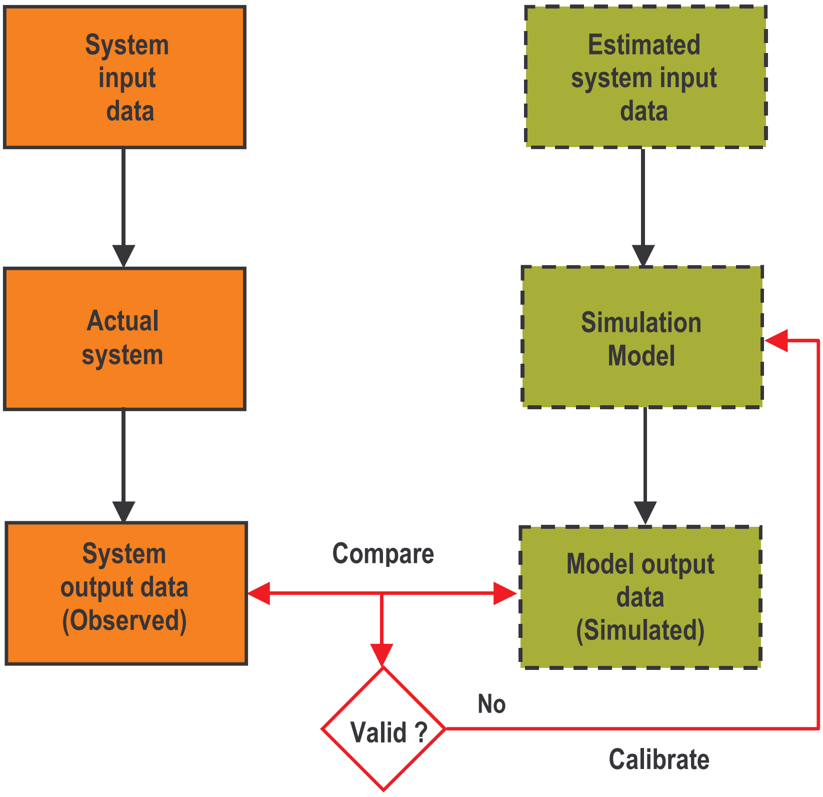
System input data
Estimated system input data
Actual system
Simulation Model
System output data (Observed)
Compare
Model output data (Simulated)
No
Valid ?
Calibrate

Figure 4.2. Logic diagram for model validation

4.2 Specifics for the Verification and Validation of Traffic Simulation Models

In the case of traffic systems, the behaviour of the actual system is usually defined in terms of traffic variables, that is, flows, speeds, occupancies, queue lengths and so on, which can be measured by traffic detectors at specific locations in the road network. To validate the traffic simulation model, the simulator should be able to emulate the traffic detection process and produce a series of simulated observations which, when compared to the actual measurements, will be used to determine whether the desired accuracy has been achieved in reproducing the system's behaviour. Rouphail (2003) proposes the following set of guiding principles:

- The analyst must be aware that calibration and validation are conducted in particular contexts.
- Depending on the context, the model requires specific sets of relevant data.
- Both models and field data contain uncertainties**.**
- Feedback is necessary for model use and development.
- Model validation must be carried out on a data set that is independent from the calibration data set.

The analyst will have to identify which data are relevant for the planned study, collect them, identify the uncertainties, filter out the data accordingly, and use two independent sets of data. *The first set should be used for calibrating the model parameters and the second for running the calibrated model and then for validating the calibrated model*.

The key question in Figure 4.2, "Is the model valid?", can then be reformulated as, "Do the model's results faithfully represent reality? The statistical techniques provide a quantified answer to this question. Its quantification can, according to Rouphail (2003), be formally stated in the following terms: the probability that the difference between the "reality" and the simulated output is less than a specified tolerable difference within a given level of significance:

$$P\{\ |\text{"reality" - simulated output}\ | \leq d\ \} > \alpha$$

where *d* is the tolerable difference threshold indicating how close the model is to reality, and α is the level of significance that tells the analyst how certain the result achieved is.

This formulation immediately raises the questions of what "reality" is and how *d* and α should be set. In this framework, the analyst's perception of the reality relies on the information gathered during the data collection and the subsequent data processing to account for the aforementioned uncertainties. The data available and its uncertainties will determine what can be said about *d* and α. To produce input data for the simulation model that is of the quality that is required to conduct an accurate statistical analysis, a careful data collection process is necessary to ensure that the desired correspondence is achieved at an acceptable significance level. Detailed examples of data collection for microscopic simulation can be found in Hughes (1998 and 2002).

4.3 Validation of Traffic Simulation Models

Statistical methods and techniques for validating simulation models are clearly explained in most textbooks and specialised papers (Balci, 1998), (Kleijnen, 1992, 1995, 1999 and 2000), (Law and Kelton, 1991). In the general methodology, the following three main principles are used to establish a framework for model validation:

- The data measured in the actual system should be split in two data sets: the data set that will be used to develop and calibrate the model, and a separate data set that will be used for the validation test.
- Specify the data collection process in the system as well as in the simulation model: traffic variables or MOEs (i.e. flows, occupancies, speeds, service levels, travel times, etc.), whose values will be collected for the calibration and validation phases, and the collection frequency (i.e. 30 seconds, 1 minute, 5 minutes, etc.)
- According to the methodological diagrams in Figure 4.2, validation should be considered an iterative process. At each step in the iterative validation process, a simulation experiment should be conducted. Each of these simulation experiments should be defined by the data input to the simulation model, the set of values of the model parameters that identify the experiment and the sampling interval.

The validation model could be approached in two different ways: in the first approach, the validation is based on a standard statistical comparison between the model and system outputs, and the second approach is based on time series analysis.

4.3.1 Validation based on a standard statistical comparison

The validation based on a standard statistical comparison between model and system outputs could be carried out using the comparison based on global measurements and/or the comparison based on disaggregated measurements.

4.3.1.1 Comparison based on global measurements

A method that has been widely used in validating transport planning models, for the typical scenario in which only aggregated values are available (i.e. flow counts at detection stations aggregated to the hour), is to analyse the scattergram or alternatively to use a global indicator such as the GEH index, which is widely used in the United Kingdom (Greater London Council, 1966). The GEH index for *n* pairs of (observed-simulated) values was calculated by the following algorithm:

For i = 1 to n calculate

```
            $\text{GEH}_i = \sqrt{\dfrac{2(\text{ObsVal}_i - \text{SimVal}_i)^2}{\text{ObsVal}_i + \text{SimVal}_i}}$
                If $\text{GEH}_i \le 5$ Then $\text{GEH}_i = 1$
                Otherwise $\text{GEH}_i = 0$
                Endif

    Endfor


            Let $\text{GEH} = \dfrac{1}{n}\sum_{i=1}^{n} \text{GEH}_i$
    If $\text{GEH} \ge 85\%$ then ACCEPT the model
    otherwise REJECT the model
     Endif
```

It needs to be noted that the GEH statistic is an “intuitive” and “empirical engineering” measure, not necessary a measure that a professional statistician would recognise or deign to use. The criterion of 85% or 80% has been established by practitioners as a rule of thumb (FHWA, 2003).

Figure 4.3 depicts an example of such an analysis. The regression line of observed versus simulated flows is plotted along with the 95% prediction interval. The R^2 (regression coefficient) value of 93.6, and the fact that only three points lie out of the confidence band would lead to the conclusion that the model could be accepted as significantly close to reality. Examples of the use of scatter plots and regression analysis combined with the RMSE can be found, for example, in the validation of MITSIM models in Yang and Koutsopoulos (1997).

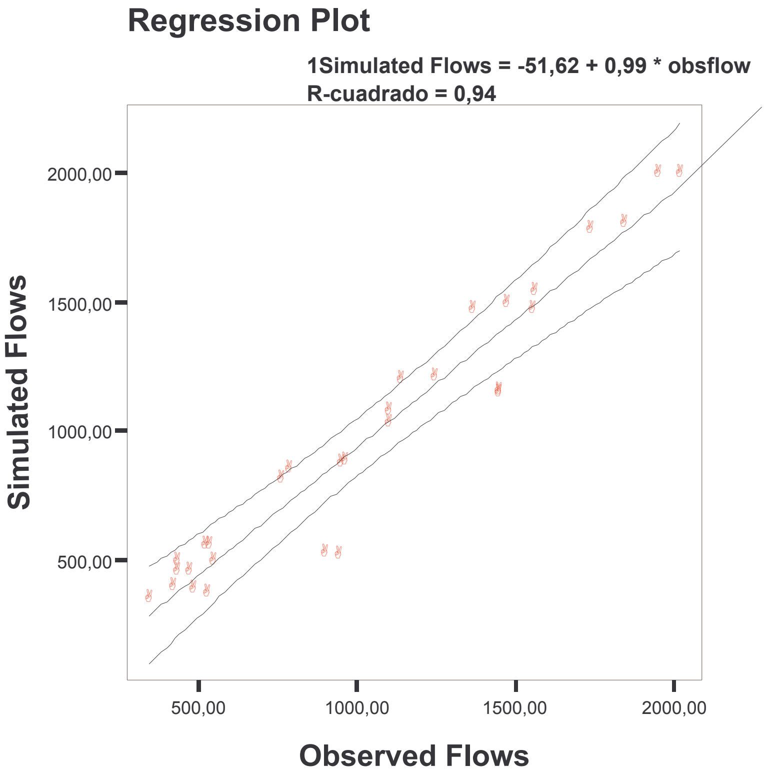

Figure 4.3. Scattergram that compares observed and simulated flows

Independently of the considerations of whether one criterion is better or more valid than another, one should draw attention to the fact that this type of indicator can only be considered a primary indicator for acceptance or rejection in the case of microscopic simulation models. As indicators working with aggregated values, they do not capture what is considered to be the essence of the microscopic traffic simulation: the ability to capture the time variability of traffic phenomena. Therefore, other types of statistical comparison should be proposed.

4.3.1.2 COMPARISONS BASED ON DISAGGREGATED MEASUREMENTS

For example, assuming that in the definition of the simulation experiment the sampling interval is five minutes, that is, that the model statistics are gathered every five minutes, and that the sampling variable is the simulated flow *w*, the output of the simulation model will be characterized by a set of values w_{ij} of the simulated flow at detector *i* at time *j*, where index *i* identifies the detector (i = 1,2,…,*n*, *n* being the number of detectors) and index *j* identifies the sampling interval (j = 1,2,…,*m*, *m* being the number of sampling intervals in the simulation horizon *T*). If v_{ij} are the corresponding actual model measures for detector *i* at sampling interval *j*, a typical statistical technique to validate the model would be to compare both series of observations to determine whether they are close enough. For detector *i*, the comparison could be based on testing whether the difference

$$D_{ij} = w_{ij} - v_{ij},\ j=1,\ldots,m$$

has a mean $\overline{\mathrm{d}}_{\mathrm{i}}$ that is significantly different to zero or not. This can be determined using the *t* statistic:

$$\bar{t}_{m-1} = \frac{\bar{d}_i - \delta_i}{\bar{s}_d \big/ \sqrt{m}}$$

where δ_I is the expected value of $\bar{d}_i$ and $\bar{s}_d$ is the standard deviation of $\bar{d}_i$, which is used to test the following null hypothesis:

$$\mathrm{H}_0 : \delta_i = 0 \quad \left(\left| \bar{t}_{m\text{-}1} \right| > t_{m-1;\alpha/2} \right)$$

- If for δ_I = 0 the calculated value $\bar{t}_{m-1}$ of Student's *t* distribution is significant to the specified significance level α, then we have to conclude that the model is not reproducing the system behaviour's closely enough. Thus the model must be rejected.

- If δ_I = 0 gives a non-significant $\bar{t}_{m-1}$, then we must conclude that the simulated and the real means are "practically" the same, so the simulation is "valid enough".

This process will be repeated for each of the *n* detectors. The model is accepted when all detectors (or a specific subset of detectors, depending on the purposes of the model and taking into account that the simulation is only a model and therefore an approximation, so δ_I will never be exactly zero) pass the test.

However, there are certain considerations that should be taken into account in the case of the traffic simulation analysis.

1. The statistical procedure assumes identically and independently distributed (i.i.d.) observations whereas the actual system measures and the corresponding simulated output are time series. Therefore, it would be desirable that at least the *m* paired (correlated) differences $d_{ij} = w_{ij} – v_{ij}$, *j=1,…,m* were i.i.d. This can be achieved when w_{ij} and the v_{ij} are average values of independently replicated experiments.

2. The bigger the sample is, the smaller the critical value $\bar{t}_{m-1;\alpha/2}$, and this implies that a simulation model has a higher chance of being rejected as the sample grows bigger. Therefore, the t statistic may be significant and yet unimportant if the sample is very large and the simulation model may be good enough for practical purposes.

These considerations mean that it is unadvisable to rely on one type of statistical test for validating the simulation model. Other authors (Rao et al., 1998) have proposed other, less stringent validation tests for traffic simulation based on the classic comparison of two means.

4.3.2 AN ALTERNATIVE APPROACH BASED ON TIME SERIES ANALYSIS

Time series are a family of statistical tests for the validation of traffic simulation models that are rooted in the observation that the measured series and the simulated series, v_{ij} and w_{ij} respectively, are time series. In this case, the series measured could be interpreted as the original series and the simulated series the “prediction” of the observed series. The quality of the simulation model could therefore be established in terms of the quality of the prediction, and that would mean resorting to time series forecasting techniques for that purpose. If one considers that what is observed as the output of the system and the output of the model that represents the system are dependent on two types of components—the functional relationships governing the system (the pattern) and the randomness (the error)—and that the measured and the observed data are related to these components by the relationship

Data = pattern + error

then the critical task in forecasting can be interpreted in terms of separating the pattern from the error component so that the former can be used for forecasting. The general procedure for estimating the pattern of a relationship is through fitting some functional form so as to minimize the error component. This could be achieved by regression analysis.

If for detector *i*-th the error of the *j*-th “prediction” is $d_{ij} = w_{ij} – v_{ij}$, $j = 1,…,m$**,** then a typical way of estimating the error of the predictions for the detector *i*-th is the root mean square error , rms_i, which is defined as

$$\mathbf{rms_i} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(w_{ij} - v_{ij}\right)^2}$$

This has perhaps been the most frequently used error estimate in traffic simulation, and although obviously the smaller rms_i is, the better the model is, it has a quite significant drawback in the fact that, because it squares the error, it emphasises large errors. Therefore, it would be helpful to have a measure that both considers the disproportionate weight of large errors and provides a basis for comparison with other methods. It is quite common, in traffic simulation, for neither the observed values nor the simulated ones to be independent, namely when only single sets of traffic observations are available (i.e. flows, speeds and occupancies for one day of the week during the rush hour). A good example of the autocorrelation analysis of observed and simulated flows is the simulation study of the I-35W freeway in Minneapolis (Hourdakis and Michalopoulos, 2002).

Theil’s U statistic (Theil, 1996) is the measure that achieves the aforementioned objective of overcoming the drawback of the rms_i index, if we explicitly consider the fact that we are comparing two autocorrelated time series, and therefore the objective of the comparison is to determine how close both time series are.

In general, if X_j is the observed and Y_j the forecasted series, $j = 1,\ldots,m$, then if $FRC_{j+1} = \frac{Y_{j+1} - X_j}{X_j}$ is the forecasted relative change, and $ARC_{j+1} = \frac{X_{j+1} - X_j}{X_j}$ is the actual relative change, Theil's U statistic is defined as

$$U = \sqrt{\frac{\sum_{j=1}^{m-1}(FRC_{j+1} - ARC_{j+1})^2 \Big/ (m-1)}{\sum_{j=1}^{m-1}(ARC_{j+1})^2 \Big/ (m-1)}}$$

then taking expression of FRC_{j+1} and ARC_{j+1}, Theil's U statistic is

$$U = \sqrt{\frac{\sum_{j=1}^{m-1}\left(\frac{Y_{j+1} - X_{j+1}}{X_j}\right)^2}{\sum_{j=1}^{m-1}\left(\frac{X_{j+1} - X_j}{X_j}\right)^2}}$$

An immediate interpretation of Theil's U statistic is the following:

$U = 0 \Leftrightarrow FRC_{j+1} = ARC_{j+1}$, and then the forecast is perfect.

$U = 1 \Leftrightarrow FRC_{j+1} = 0$, and the forecast is as bad as is possible.

In the latter case, the forecast is the same as that that would be obtained if no changes in the actual values were forecast. When forecasts Y_{j+1} are in the opposite direction to X_{j+1}, then the U statistic will be greater that unity. Therefore, the closer to zero Theil's U statistic is, the better the forecast series is, or, in other words, the better the simulation model. When Theil's U statistic is close to or greater than 1, the forecast series and therefore the simulation model should be rejected. Taking into account that the average squared forecast error

$$D_m^2 = \frac{1}{m}\sum_{j=1}^{m}(Y_j - X_j)^2$$

can be decomposed (Theil) in the following way:

$$D_m^2 = \frac{1}{m}\sum_{j=1}^{m}(Y_j - X_j)^2 = (\overline{Y} - \overline{X})^2 + (S_Y - S_X)^2 + 2(1-\rho)S_Y S_X$$

where $\overline{Y}$ and $\overline{X}$ are the sample means of the forecast and the observed series respectively, S_Y and S_X are the sample standard deviations and ρ is the sample correlation coefficient between the two series, the following indices can be defined:

$$\left.\begin{aligned} U_M &= \frac{\left(\overline{Y}-\overline{X}\right)^2}{D_m^2} \\ U_S &= \frac{(S_Y - S_X)^2}{D_m^2} \\ U_C &= \frac{2(1-\rho)S_Y S_X}{D_m^2} \end{aligned}\right\} \Rightarrow U_M + U_S + U_C = 1$$

U_M

U_M is the “bias proportion” index, which can be interpreted in terms of a measure of systematic error, U_S is the “variance proportion” index, which provides an indication of the forecast series’ ability to replicate the degree of variability of the original series or, in other words, the simulation model’s ability to replicate the variable of interest of the actual system. Finally, U_C or the “covariance proportion” index is a measure of unsystematic error. The best forecasts, and hence the best simulation model, are those for which U_M and U_S do not differ significantly from zero and U_C is close to unity. It can be shown that this happens when, in a regression, β_0 and β_1 do not differ significantly from zero and unity respectively.

The example of this detector demonstrates how these statistical techniques can reveal hidden information that is critical in certain aspects of validation, which traditional techniques cannot do. The results for the plot of the observed and simulated series are shown in Figure 4.4. Visual inspection reveals a very good agreement between both series confirmed by the value 0.9999 of R^2. The analysis of Theil’s coefficients corroborates the quality of the simulation model. There are very low values of U (0.015348), U_c (0,000362) and U_S (0.055073)), although the presence of a very high value of U_M (0.920005) reveals the presence of a systematic bias. There is an almost constant difference of four units between the observed and the simulated series, that is, the simulated series is shifted 4 units with respect to the observed one. The discrepancy could be explained in this case (see Hourdakis and Michalopoulos, 2002 for details) by the misplacement of the detector.

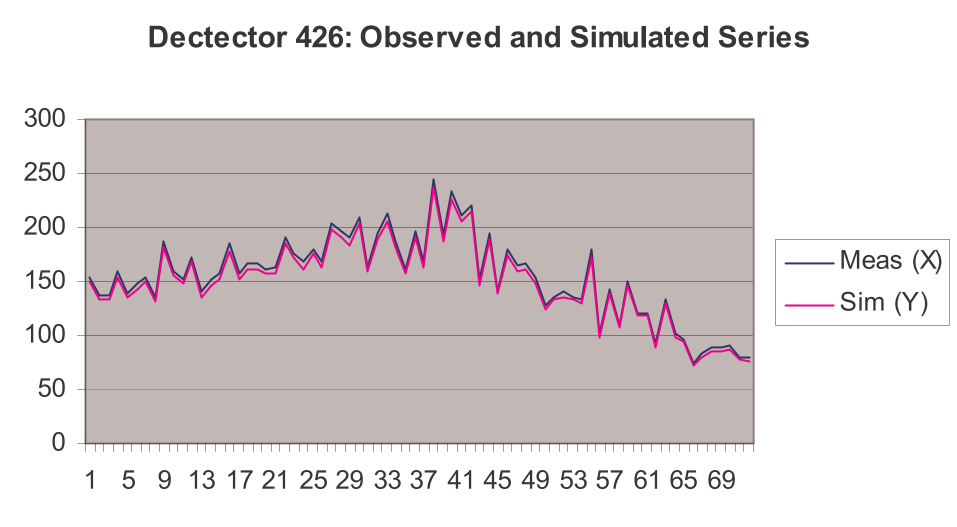

Figure 4.4. Observed and simulated time Series

4.4 DYNAMIC TRAFFIC ASSIGNMENT VALIDATION PARAMETERS

The statistical methods and techniques for validating traffic simulation models presented in the previous section give a validation of the simulation model, but a critical aspect in the calibration/validation of the dynamic traffic assignment model is determining the values of the dynamic traffic assignment parameters that enable a meaningful selection of paths. No formal convergence proof can be given for the dynamic traffic assignment proposed, since the heuristic network loading process based on microscopic simulation does not have an analytical form. The method proposed is based on the assumption that, insofar as the assignment described may be associated with a heuristic approach to a preventive dynamic equilibrium assignment (Xu et al., 1999), properly selecting the path should lead to such equilibrium. An assignment's progress towards equilibrium, and therefore the quality of the solution, may be measured using the relative gap function, *RGap(t)* (Florian et al., 2001 and Janson, 1991), which estimates, at time interval *t,* the relative difference between the total travel time actually experienced and the total travel time that would have been experienced if the travel time for all vehicles had been equal to the current shortest path, such that

$$RGap(t) = \frac{\sum_{i \in I} \sum_{k \in K_i} h_k(t)\left[s_k(t) - u_i(t)\right]}{\sum_{i \in I} g_i(t) u_i(t)}$$

where

t is the time interval used in the dynamic traffic assignment algorithm;

I is the set of all OD pairs;

$k \in K_i$ is the set of paths for *i*-th OD pair;

g*i* *is* the traffic demand of OD pair *i;*

$h_k(t)$ is the path flow assigned to path $k \in K_i$ that connects OD pair *i* at interval *t;*

$s_k(t)$ is the total travel time experienced of all vehicles assigned to path $k \in K_i$ that connects OD pair *i* at interval *t*; and

$u_i(t)$ is the total travel time experienced by all vehicles assigned to the shortest path that connects OD pair *i* at interval *t.*