NISS

Statistically-based Validation of Computer Simulation Models in Traffic Operations and Management

Jerome Sacks, Nagui M. Rouphail, B. Brian Park,
and Piyushimita (Vonu) Thakuriah

Technical Report Number 112
December, 2000

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Statistically-Based Validation of Computer Simulation Models in Traffic Operations and Management

by

Jerome Sacks, Ph.D.
Senior Fellow
National Institute of Statistical Sciences and Duke University
Research Triangle Park, NC 27709-4006
Phone: 919-685-9300 Fax: 919-685-9310 Email: sacks@niss.org

Nagui M. Rouphail, Ph.D.
Professor
Department of Civil Engineering, North Carolina State University
Campus Box 7908, Raleigh, NC 27695-7908
Phone: 919-515-1154 Fax: 919-515-7908 Email: rouphail@eos.ncsu.edu

B. Brian Park, Ph.D.
Research Fellow
National Institute of Statistical Sciences
Research Triangle Park, NC 27709-4006
Phone: 919-685-9326 Fax: 919-685-9310 Email: park@niss.org

and

Piyushimita (Vonu) Thakuriah, Ph.D.
Assistant Professor
Urban Transportation Center, University of Illinois at Chicago
412 S Peoria Street, Suite 340, Chicago, Illinois 60607-7036
Phone: 312-355-0447 312-Fax: 312-413-0006 Email: vonu-pt@uic.edu

Paper submitted to the Journal Transportation and Statistics
December 2000

ABSTRACT

The process of model validation is crucial for the use of computer simulation models in transportation policy, planning and operations. The obstacles that must be overcome and the issues that must be treated in performing a validation are laid out here. We describe a general process that emphasizes five essential ingredients for validation: *context*, *data*, *uncertainty*, *feedback*, and *prediction*. We use a test-bed to generate specific (and general) questions and to give concrete form to answers and the methods used in providing them.

The test-bed is the traffic simulation model, CORSIM; we apply it to assess signal-timing plans on a street network of Chicago. The validation process applied in the test-bed demonstrates how well CORSIM can reproduce field conditions, identifies flaws in the model that need to be overcome, and how well CORSIM predicts performance under new (untried) signal conditions. One specific conclusion: CORSIM, though imperfect, is effective in evaluating signal plans on urban networks, at least under some restrictions.

Statistically-Based Validation of Computer Simulation Models in Traffic Operations and Management

Jerome Sacks, Nagui M. Rouphail, B. Brian Park, and Piyushimita Thakuriah

1. Introduction

The validation of computer simulation models is a crucial element in assessing their value for making transportation policy, planning and operational decisions. Often discussed and sometimes practiced informally, the process is straightforward conceptually: data are collected that represent both the inputs and the outputs of the model, the model is run at those inputs, and the output is compared to field data. In reality, complications abound: field data may be expensive, scarce or noisy, the model may be so complex that only a few runs are possible, and uncertainty enters the process at every turn. Even though it is inherently a statistical issue, model validation lacks a unifying statistical framework.

The need to develop such a framework is compelling, even urgent. The use of computer models by transportation engineers and planners is growing; costs of poor decisions are escalating; and increasing computing power, for both computation and data collection, is magnifying the scale of the issues.

The opportunity is as great as the needs. Advances in statistical techniques for incorporating multiple types of information, while managing the multiple uncertainties, provide ground on which progress can be made in quantifying validation (Berliner et al., 1999; Lynn et al., 1998)

The purpose of this paper is to lay out a set of key issues faced in the validation of transportation models and advance a research effort to meet these issues. Many of the ones we describe are common to models and modelers in all areas of science and engineering:

- give explicit meaning to validation in particular *contexts*;
- acquire relevant *data*;
- quantify *uncertainties*;
- provide *feedback* to model use and development;
- *predict* performance under new (untried) conditions.

While easily said, the challenge is to meet these issues, by describing and developing approaches and methods that are effective and can be implemented. That there are many obstacles to surmount is no surprise to those who have attempted exacting validations. But there are successes as well and tools to exploit that are capable of overcoming the impediments.

In order to make our points clear, we will use a test-bed that generates questions a validation must address and, at the same time, accommodates analyses that respond to the key issues. The test-bed we use is the microsimulator CORSIM in an application to the assessment and selection of signal timing plans on an important street network in the City of Chicago.

The transportation science is clearly evident in the formulation, or meaning, of the validation as well as in data collection, feedback and prediction. Statistical issues are prominent in data collection and uncertainty quantification, but also play a vital role in formulation, feedback and prediction. The interplay between transportation science and statistics is clearly critical; on this ground future research needs to be built.

The paper is written so that general principles and issues are interwoven with their specific manifestation in the test-bed problems. At the end we are led to an outline of the validation process, a set of research issues that should be addressed to advance “validation science”, and the consequences of the validation in the context of the test-bed.

The five bulleted items above form an outline of the validation process. Within each are a number of research issues that emerge in the discussion in the following sections. Briefly they point to the need to

- formulate evaluation functions that capture transportation needs and are amenable to either direct, or indirect, observation in the field;
- measure and assess the impact of data quality on evaluation functions and performance;
- develop methods for treating a variety of problems connected with the analysis of uncertainties (see Section 7), including uncertainty of predictions.

The general conclusion from the test-bed is that, despite imperfections, CORSIM is effective as a model for evaluating signal plans on urban street networks at least under some restrictions. The basis of the statement is the validity of CORSIM prediction of performance under new conditions assessed by a second data collection – the “gold standard” of validation. The simplicity of the conclusion obscures the complexity of tuning the model to the specific network using data from “old conditions” and an initial data collection – thus the importance of the “feedback” step in the process.

We introduce the test-bed example and simulator in Section 2 along with the specific evaluation functions we use. Acquisition of data and the two field collections are described in Section 3. Estimation of the inputs to the model is described in Section 4. Section 5 covers the range of validation questions and the analyses relevant to them, including tuning, all based on the initial data collection. Section 6 discusses the prediction of performance under new conditions and the subsequent validation. Questions about uncertainty are discussed in Section 7; our conclusions are in Section 8.

2. The Test-Bed: CORSIM and Signal Timing on an Urban Street Network

CORSIM is a computer simulation model of street and highway traffic; it is the quasi-official USDOT platform upon which to gauge traffic behavior and compare competing strategies for signal control before implementing them in the field (FHWA, 1997).[1] For CORSIM to fulfill this purpose two crucial questions must be addressed:

(1) How well does CORSIM reproduce field conditions?
(2) Can CORSIM be trusted to represent reality under new, untried conditions (e.g., revised signal timing plans)?

Figure 1. TEST-BED NETWORK

Answering these two questions is challenging because of the localized and complex behavior signal plans induce on urban street networks. Flows on these networks, even small sub-networks, are highly complex: they encompass a variety of vehicles (autos, trucks, buses), pedestrian-vehicle interactions, driver behavior, and an assortment of network conditions (lane arrangements, stop signs, parking lots, one-way streets). Moreover, the traffic demands on the network are highly variable (minute-to-minute, hour-to-hour, day-to-day, month-to-month,…) as are many of the movements (even legal ones) of vehicles and pedestrians.

Since no simulator can be expected to capture real behavior exactly, formulating appropriate performance measures or evaluation functions is fundamental to the validation process. Variability, inherent in real traffic and also present in the computer model, compounds matters; choices of performance measures introduce subjective elements and thereby, potential sources of contention in assessment of the computer model.

To focus the issues we undertook a case study with the cooperation of the Chicago Department of Transportation (CDOT). The ultimate goal will be to optimize the signal plans for a network more extensive than the one below; the use of CORSIM to achieve this requires answers to our two questions. The test-bed for the study is the network depicted in Figure 1; the internal network (Orleans to LaSalle; Ontario to Grand) in Figure 1 is the key part of a planned RT-TRACS study to be carried out in Fall 2000. A different network had been studied earlier (Park et al., 2000), and helped guide some of the decisions made in the current test-bed.

Traffic in the network depicted in Figure 1 flows generally to the South and East directions during the morning peak, and to the North and West in the evening peak. This demand pattern is accommodated by a series of high-capacity, one-way arterials such as Ohio (EB), Ontario (WB), Dearborn (NB) and Clark and Wells (SB) in addition to

[1] CORSIM version 4.32 is the one used in this paper.

LaSalle (NB and SB). For reference purposes, the Chicago CBD is located southeast of the network.

2.1. *CORSIM Characteristics and Inputs*

CORSIM is a stochastic simulator. It represents individual vehicles, which enter the road network at random times, are moved (randomly) second by second according to local interaction rules describing governing phenomena such as car following and lane changing, response to traffic control devices, and turning at intersections according to prescribed probabilities. CORSIM can handle networks of up to 500 nodes and 1000 links containing up to 20,000 vehicles at *any one time*. The network of Figure 1 has 112 one-way links and 30 signalized intersections and about 38,000 vehicles move through it in an hour. Streets are modeled as directed links; intersections as nodes.

There are a variety of inputs or specifications that must be made, either directly or by default values provided in CORSIM. Inputs that must be *directly* made include

- specification of the *network* via fixed inputs describing the geometry (e.g., distance between intersections, number of traffic lanes, length of turn pockets), the placement of stop signs, bus stops, schedules and routes, and parking conditions;
- probability distributions of *inter-arrival times* governing the generation of vehicles at each entry node of the network -- the choices in CORSIM of arrival time distributions are limited, in essence, to have Gamma (Erlang) densities,

$$p(t \mid \lambda, k) = \frac{(\lambda k)^k}{(k-1)!} t^{k-1} \exp(-k\lambda t),$$

which are assumed independent (vehicle-to-vehicle, node-to-node), but allowed to be different for each entry node;

- *vehicle mix* -- auto or truck -- through independent Bernoulli trials with probabilities that can differ from entry node-to-entry node;
- probability distributions of *turning movements*, assumed to be independent, vehicle-to-vehicle and link-to-link, and different from link-to-link.

Default inputs are several. The chief ones are inputs that relate to driver characteristics such as car-following behavior, left-turn “jumpers”, acceptance of gaps between vehicles and lane-changing maneuvers. For example, *gap acceptance* is governed by a discrete distribution with 10 jumps. The default distribution can be taken or altered. Other inputs with default distributions that can be altered are dwell-times for buses, effects of pedestrians on turning vehicles, and short-term incidents, such as an illegally parked delivery truck.

Altering the default distributions by use of data is possible in some cases; but data that would better inform determination of driver characteristics are too elusive. For the test-bed study we assumed no pedestrian traffic (normally light on this network) and no incidents (but see below at Section 3).

Signal settings are direct inputs. We single them out as *controllable* factors since it is altering these inputs to produce improved traffic flow that drives the study. Signal settings consist of a cycle common to all signals, green times for movements at each intersection, and offsets (time differences between beginnings of cycles at intersections).

For validation the signal plan will be the one found in the field. For finding optimal fixed-time signal-timing plans[2], or comparing alternative plans, the signal parameters will necessarily be manipulated. Comparisons are best done through the simulator since field experiments are infeasible. Relying on CORSIM to select an alternative to a current in-place plan then raises the questions posed at the beginning of the section.

2.2. *CORSIM Output*

CORSIM comes equipped with an animation package (TRAFVU) that enables visualization of the traffic movements, a capability of great value in exploring the characteristics of the model and detecting problems and flaws. In addition to the visual output, CORSIM provides aggregated (over selected time intervals such as the signal cycle) numerical output for each link. The numerical outputs include

> *throughput* (the number of vehicles discharged on each link);
>
> *average link travel time*;
>
> *link queue time* (the sum over vehicles of the time, in minutes, during which the vehicle is stationary, or nearly so);
>
> *link stop-time* (sum over vehicles of stationary time);
>
> *maximum queue length* on each lane in the link over the simulation time;
>
> *link delays* (simulated travel time minus free-flow travel time, summed over all vehicles discharging the link).

Most of these statistics can be further segregated by movement or lane levels within each link. It is from these outputs that CORSIM performance measures will be taken.

One hour of simulation for the test-bed network takes about 40 seconds on a Pentium III-850 MHz PC. During this time, approximately 38,000 vehicles are processed through the network in an hour. While each run is fairly quick, the need for many runs to deal with

[2] Adaptive plans are under consideration as part of the RT-TRACS program and require extensive sensor capabilities to capture dynamic traffic conditions; models accommodating such plans are themselves subject to validation study.

the substantial variability induced by the stochastic assumptions makes the time for a single experiment non-trivial; the burden is magnified when a detailed uncertainty analysis is undertaken.

3. Data Collection

A crucial element in validation is designing and carrying out a data collection both for estimating inputs to the model and for comparing model output with field data. The challenge lies in managing costs while obtaining data that are useful and relevant for both estimation and validation.

For our test-bed example initial field data for the network were collected on a single day (Thursday, May 25, 2000) for three hours in the morning (7:00am-10:00am) and three hours in the afternoon (3:30pm-6:30pm). The processing of the data and the analyses were limited to the three time periods 8am-9am, 4pm-5pm, and 5pm-6pm. This covered the peak periods as well as a “shoulder” period.

Acquiring data for the inputs to CORSIM is a formidable task. Inputs such as driver characteristics are extremely difficult to gather and, in the test-bed example, we relied mostly on CORSIM default values. Some of the inputs, such as pedestrian effects, were ignored because there were few pedestrians. Incidents were not included despite the fact that there were illegally parked vehicles that did affect traffic flow. Because this was an endemic condition, they were handled by coding the network to account for their effect. Other parameters, such as free-flow speed were selected on the basis of posted speed limits (more about that later in Section 5.2). Signal timing plans and bus routes and stations were collected directly in the field and entered into CORSIM.

Traffic volume data were collected manually (by observers counting vehicles) and by video recording of traffic. Manual observation is notoriously unreliable but cost considerations did not allow video coverage of the full network. However, the video information was rich enough (covering all the links of the internal network of Figure 1) to allow adjustment of the manual counts determining the flow rate of vehicles at entry nodes of the network. On the other hand, turning movements outside the internal network could neither be confirmed nor reliably adjusted by video information. Extracting the video information took a considerable investment of time and personnel, rivaling the cost of acquiring the raw video data.

Supplemental validation data were collected on a similar schedule, on September 27, 2000. These were extracted primarily from video. The purpose: to produce a response to question 2 by analyzing the effectiveness of CORSIM to predict traffic behavior under the new conditions prevailing in September.

Collecting data for validation is most conveniently done simultaneously with data collection for inputs. The use of the same or closely related data for both input and validation is an issue that is rarely confronted. The conventional wisdom is that such dual-use of the data is “forbidden”. In fact, it can be done but how to attach computable uncertainties, essential to producing fully reliable results, is not straightforward. This issue is under study by a research team at NISS and Duke University; a Bayesian approach based on Bayarri and Berger (1999) holds great promise for producing methodology to treat the issue.

A problem as yet not addressed is assessing the impact of data of inferior quality. The problem is complicated by the need to specify the "brunt of the impact"; quantify scenarios of alternative collections of data; and design, execute and analyze computational experiments to measure the consequences, or sensitivities, of model output to wrong data inputs including incorrect signal settings or drifts in signal timing. This issue is not unique to transportation studies and research, it permeates virtually all of science.

4. Estimation of CORSIM Inputs from Initial (May) Data Collection

The direct fixed inputs required by CORSIM (see Section 2.1) including signal timing plans for each of the three one-hour periods were obtained from the field and entered into CORSIM. The direct inputs that needed estimation were treated as follows:

- vehicle mix at each entry node was estimated from one-hour (manual) counts for autos and trucks;
- turning probabilities (left-turn, right-turn, through) at each intersection were estimated from one-hour video counts (where video was available) and from manual counts at other intersections;
- inter-arrival rates (see equation in section 2.1) were estimated under the assumption that k=1 and the λ for each entry node and each of the three one-hour time periods was estimated as the total number of vehicles entering the (entry) link/3600.

Some λ's were later adjusted to reduce discrepancies between downstream counts generated by CORSIM and those observed by video – the discrepancies are believed due to inaccuracy of manual counts and the effects of parking lots. Turning movements were left at their field estimates. Measuring the ultimate effect on uncertainty of these modifications is an issue that remains to be explored.

5. Validation Process

Validation without purpose has little utility. For example, our interest in CORSIM here is for its value in assessing and producing good time-of-day signal plans. But CORSIM could also be used to evaluate traffic operations under disruptions (say a bridge closing) or to changes in the network (such as, strict enforcement of parking laws or truck restrictions). A more subtle use could be in measuring the impact of driver decisions when faced with a network modification such as a bridge closing. Some objectives may only reflect network changes, others may also implicate induced changes in demand.

Navigating through this variety of issues requires multiple tools:

Visualization and expert opinion can provide an overall impression whether the model output matches reality in a qualitative but highly subjective way. When video data are placed side by side with computer animations, discrepancies (and similarities) can often be seen directly, particularly if viewers are experts familiar with the network and its characteristics.

But the stochastic nature of CORSIM and of real traffic requires more. In particular, which random animation should be used to compare with the real traffic, and is the single day of traffic recorded by video "typical". More stringent comparisons based on *statistical analysis* are then crucial to help reduce the subjectivity, guide the visualization through choices of animation, and point to model flaws responsible for aberrant behavior. The challenge is then to provide statistical analyses that are appropriate to the desired ends.

There can be many competing "stories", one for each evaluation criterion as defined below in (5.1). Treating the *multiplicity* of comparisons in a coherent way is often disregarded – is the model flawed if it produces a poor match to reality at only one (five?) of one hundred links? Added complications come from comparisons based on evaluations of corridor and system characteristics as well as those of individual links.

Thus, an initial task is to set out criteria for selecting evaluation criteria. Comparison of field and model through selected φ's in the specific application of CORSIM to the network of Figure 1 will touch on the concerns and issues raised here as well as those noted below.

5.1. Evaluation Functions

Selecting an evaluation function φ is crucial and sometimes complicated by competing practical and theoretical considerations. First, is φ *relevant* to the purpose? Choosing among many relevant φ's is sometimes eased by requiring *feasibility* in both calculating model outputs for φ and collecting field data for calculating corresponding field value(s) of φ.

In our test-bed example, a good criterion for judging a signal-timing plan may be average link travel time, not straightforward to obtain in CORSIM and very costly to obtain in the field. The tactic of using probe vehicles, while in principle would work, is inhibited by the cost of using large numbers of vehicles and the need to account for the substantial variability connected with the use of probes. Computing travel time of vehicles from video records is highly labor-intensive; useful automatic area wide detection method, Mobilizer (Lall et al., 1994) are not yet widely available.

The evaluation function φ is likely to have versions at multiple time-scales and at different levels of spatial aggregation. For example, total queue-time per cycle per link could be aggregated over cycles and over links to form evaluations based on behavior over selected corridors, over the whole system and over distinct time periods. The choice of levels of *space-time resolution* adds to the determination of relevance and can be complicated by questions of feasibility.

Statistical analyses of the φ's must necessarily treat their *variability* arising from the intrinsic stochastic structure of simulators such as CORSIM.[3] But the field variability is also consequential and that cannot be so readily captured without an elaborate and costly field data collection. This is a confounding issue and only partly addressed below.

Travel times are very hard to obtain in the field as noted earlier. Stop-time per vehicle can be calculated for each link covered by video. Queue length per cycle can also be calculated, but queue-time is very difficult to obtain in the field although it is a standard part of CORSIM output.

We chose stop-time (stopped delay) on approaches to intersections as the primary evaluation function. It has been the typical measure by which intersection level of service (LOS) is evaluated (Highway Capacity Manual, 1994). Our choice was affected strongly by the comparative ease in collecting stop-time data from the video. The choice was further buttressed by the fact that other criteria such as throughput, delay, travel time, queue length are all highly correlated with stop-time.[4] In addition, we believe that drivers on urban street networks are particularly sensitive to stop-time**,** spurring traffic managers to seek its reduction. In fact, the Highway Capacity Manual selection of stopped delay for LOS designation is meant to reflect the user-perception of the intersection quality of service. We used *V* (the number of vehicles leaving an intersection, particularly "exit" nodes) as an auxiliary evaluation function. V is readily calculated from video and is also needed to calculate stop-time per vehicle discharged (*STV*) at a link.

At approach $a, STV(a) = \frac{Total\ stop\ time}{V(a)}$.

$$V(a) = V_0(a) + V_s(a)$$

where V_0 is the count of vehicles that do <u>not</u> stop on *a* while V_s is the count of vehicles that do stop on *a.* This raises the question of whether STV is an adequate reflection of the characteristics of the network (and signal plan) compared to the pair

$$P(a) = \frac{V_s(a)}{(V_0(a) + V_s(a))}$$

[3] Deterministic models will not have intrinsic randomness but will be exposed to variability either in assumptions about input parameters or from data used to estimate input parameters.

[4] Rejection of delay was also affected by CORSIM's calculation which fails to include vehicles left in the system at the end of the one-hour simulation period, potentially resulting in misleading numbers under congested conditions.

$STVS(a) = \frac{Total\ stop\ time}{V_s(a)}$ = stop-time per stopped vehicle.

We shall see that these quantities provide sharper understanding of the comparison between CORSIM and the field.

STV or STVS for aggregations of approaches (routes or corridors) is very difficult to obtain, requiring tracking of individual vehicles. But some concept of performance on aggregation could be important for example, long delay on one link may be compensated by short delay on the next downstream link leaving the corridor and the system as a whole unaffected. By summing over the individual links forming a corridor we could create a "pseudo stop-time" for the corridor. This will be close to a real stop-time provided vehicles turning off or onto the corridor exhibit little or no difference from those traveling straight through. However, the value of such "pseudo stop-times" is unclear and we only deal with individual links and approaches.

Multiplicity questions begin with selection of links or approaches to compare. We selected links on corridors that bore the heaviest traffic during the main peak period directions (East and South in the morning, West and North in the evening). A full treatment of multiplicity questions will not be done here**.**

5.2. Tuning

Tuning and calibrating a model are general terms, often used interchangeably, and sometimes confusingly. In Section 4 we treated estimation of inputs to the model directly from field data. When *model output* are used, either alone or with field data, to determine input parameters the process is often called calibration Tuning is a phrase commonly associated with adjusting input parameters to match model output. Formally the same as calibration the term, tuning, is frequently reserved for cases where the input parameters are unobservable or represent physical (and other processes) which the model does not (or cannot) adequately incorporate.

The practice of tuning is not only common but is often essential, especially for long-range study of the model and its associated phenomena. Some input parameters may neither be well-specified, nor capable of being estimated from the field data (for example, driver aggressiveness in our test-bed). Some assumptions about input parameters may be found to be erroneous after viewing the data and their modification may produce "better" simulations. Ultimately, what becomes problematic is the validation accompanying such tuning.

Two types of tuning were done in the test-bed example. The first addressed the blockage of turns at two intersections and subsequent gridlock. We altered the network by introducing sinks and sources that allowed the bypass of the blockage without affecting throughput. The second was stimulated by a substantial difference on one link (NB LaSalle/Ontario in Figure 1) between the field and CORSIM stop-times. This difference was largely resolved by changing the free-flow speed from 30 mph to 20 mph. The input of 30 mph was induced by the speed limit; the revision to 20 mph is consistent with the observed (from video) speed of vehicles on the corridor (LaSalle St).

5.3. Visual Validation

Where visualization is available as it is with CORSIM animation and with video field data, then a compelling approach to validation is to compare the two visually to see if traffic in CORSIM behaves like traffic in reality. To a great extent this produces a highly informal and subjective approach. Nonetheless, it is of great value in assessing CORSIM's capability to emulate reality as well as identifying sources of trouble or flaws in CORSIM, flaws that can sometimes be "fixed" by expert intervention in the coding.

Surely, the utility of visualization depends on the specifics of each application. What may be learned from the CORSIM example may pertain to other microsimulators, but not necessarily to other computer models.

A sign of problems in an application of CORSIM is the presence, in several of the replicate simulation runs, of spillback and gridlock in situations where such do not occur in reality. Spillback will occur on networks such as in Figure 1 (where near saturation conditions are present during peak periods) but recovery in the field usually takes place reasonably quickly. A difficulty with CORSIM is its apparent inability to recover readily from spillback, often resulting in gridlock. The effect on performance measures is usually to produce some large outliers in a repeated set of simulations, sometimes indicated by large run-to-run variance. A histogram of outputs can identify large outliers and following up with examination of the corresponding animations can often identify causes.

In two instances it was apparent that the cause was an inability of CORSIM to allow driver adjustment to left (or right) turn blockage resulting in spillback that would never clear up.

5.4. Numerical Comparisons

In 5.4.1 we discuss throughput; in 5.4.2 we deal with stop-time.

5.4.1. Throughput Comparison

In Table 1 we present test-bed results on throughput for internal network. The net change indicates that the field data reveals discrepancies showing less output in the morning and more output in the evening. This is due to the garage effect -- in the morning vehicles disappear to the parking lots and reappear from the parking lots in the evening.

The means of 100 replicated CORSIM runs are close to the observed counts in Table 2, except for EB Ohio/LaSalle in the morning and WB Grand/Wells in the evening. The first can be explained in large part by the "disappearance" of vehicles in the morning into parking lots along Ohio St., a major one-way eastbound corridor. The second, correspondingly, can be attributed to the "appearance" of vehicles from parking lots on Grand during the evening. In addition there is high enough variability in CORSIM runs to account for a considerable part of the apparent discrepancy (see Figures 2 and 3).

Table 1. Comparison of throughput on the internal network (veh/hr)
Table 2. Comparison of throughput on selected key links (veh/hr)
Figure 2. Link throughput at EB Ohio/LaSalle (8-9am)
Figure 3. Link throughput at WB Grand/Wells (5-6pm)

It would be incautious to view the closeness of real data to the model runs as evidence of the model's validity. Whether these internal throughputs are "good" evaluation functions is unclear. They are, however, relevant to STV and STVS because they determine the denominators of those measures. Not taken into account is the tuning of the model (as stated in Section 5.2) to help match inputs to the model with the flows observed in the video. How to do this formally is a matter of some delicacy and a research issue currently under investigation in a National Science Foundation sponsored research project at NISS.

Though field variability cannot be adequately captured we produced CORSIM and field time series of throughputs to examine whether CORSIM exhibits a degree of variability (over time) that is characteristic of the field data. In Figure 4, we exhibit such time-series. They are obtained as follows. There are 48 cycles and we elected to combine throughputs over every two-cycles (corresponding to 150 seconds of elapsed time). This leads to a time-series at 24 time points. CORSIM was run 100 times and the variation of each time series was computed as

$$\frac{\sum_{t=1}^{23}\left[Y(t+1)-Y(t)\right]^2}{23}$$

where Y(t) represents throughput during time interval t. The "representative" CORSIM time-series is selected to be the one whose variation is the median of the 100 variations.

Figure 4. Comparison of CORSIM vs. Field variation (EB Ohio at LaSalle, AM Peak)

CORSIM variability here (as well as on the link SB LaSalle at Ohio) is close to that of the field. Indeed, the variation of the field series is 116 and is at the 30%tile of the CORSIM distribution as shown in Figure 5.

Figure 5. CORSIM variation vs. field (EB Ohio at LaSalle, AM Peak)

5.4.2 Stop-time Comparisons

The distribution of stop-time at each approach has mass at 0 (the proportion of vehicles that do not stop) which is singled out in the first part of Table 3. Characteristics of the conditional distribution of stop-time (given that a vehicle stops) are given in Table 4. There are definite discrepancies on SB LaSalle at Ohio during the morning where

CORSIM generates fewer stops but longer stop-times for its stopped vehicles. On EB Ohio at LaSalle a similar (though somewhat reduced) discrepancy is apparent. While there appear to be differences on some of the other approaches, none appear very significant. For example, CORSIM stops fewer vehicles on NB LaSalle at Ontario in the 5-6 pm period but the stop-times are close.

These differences call for an explanation. Examination of video and CORSIM animation exposes the key cause: CORSIM does not fully reflect driver behavior. In particular, lane utilization in CORSIM is not consistent with lane utilization in the field; on some links vehicles in the field more often join long queues where they are stopped, but briefly. These vehicles typically do not appear in CORSIM simulation as having stopped. This accounts for smaller STVS in the field than in CORSIM. So, even though CORSIM does not fully reflect the field the key measure of how long are “truly stopped” vehicles delayed appears to match quite reasonably what is seen in the field.

Table 3. Comparison of stop rates on key links
Table 4. Comparison of key-links STVS (stop time per vehicle stopped)

6. Prediction and Validation

The most compelling form of validation is through confirmation by predictions in new circumstances. In the test-bed example a plan, different than the one in the field in May, was put in place in September, 2000. Under these new circumstances that is, new signal plan, predictions were to be made and data collection designed for September 27, 2000, a day expected to be similar to May 25, 2000, the date of the first data collection.

Prediction of the performance by the simulator requires specification of the inputs expected to prevail at the time of the new data collection. Believing that the conditions in the field for the September data collection would be the same the as in May we ran CORSIM with the May inputs, except for signals. After the data were collected in September we compared the results, first for throughput (Table 5) on several key links.

Table 5. Field-measured throughput comparison at key links

Except for the 13% disparity on SB LaSalle the throughputs are close. Whether or not the disparity in demand on SB LaSalle mattered awaited further analysis of stop-time. The predictions of September stop-time performance (using the May inputs) are in Tables 6 and 7. See also, Figures 6 and 7. Except for NB Orleans to Freeway, the STVS’ are reasonably close. For reasons discussed earlier (in 5.4.2) we have several disparities on stop rates.

To clarify these matters we first checked the effect of change in demand on SB LaSalle during the AM peak. We decreased the input demand there by 10%, reran CORSIM 100 times, and obtained essentially no change in output (the stop rate on SB LaSalle at Ohio went from 30.3% to 30.9% while STVS went from 22.0 to 22.3 sec/veh).

Next we explored the disparity on NB Orleans at the Freeway in the PM peak and observed, through video, that drivers effectively used green time of 20 seconds instead of the displayed green time of 16 seconds. Introducing this modification changed stop rates from 74% to 65% and average STVS from 51.9 to 40.8 sec/veh with s.d. of 6.8. The difference between 31.4 (the field STVS) and CORSIM's average of 40.8 is neither statistically significant (within 2 s.d.'s) nor practically significant (same Level of Service, see Table 9). Nonetheless we examined the NB Orleans link more carefully. We noted that CORSIM has difficulty in dealing with storage of vehicles on short, congested links that are just downstream of a wide intersection---exactly the characteristics of NB Orleans at Freeway (intersection at Ohio is 60 ft, the entire link is 240 ft, and the link is highly congested). We could have brought the CORSIM predictions more closely in line with the numbers in the field by altering the length of the link, but we regarded such tuning as potentially misleading.

Table 6. Comparison of stop rates on key-links
Table 7. Comparison of STVS (stop time per vehicle stopped)
Figure 6. Link STVS at SB LaSalle/Ohio
Figure 7. Link STVS at NB LaSalle/Ontario

A highly informative evaluation function of CORSIM is the change in CORSIM predictions, ΔCORSIM (September STVS - May STVS) compared to the corresponding change in the field values, ΔField. Even though the CORSIM predictions were not always accurate, the Δ's are close and of the same sign (Table 8). This is particularly important for comparing the performance of competing signal plans: predictions of improvements (in two links), no change (in two links) and degradation (on one link) in CORSIM jibes with the changes in reality.

Table 8. Δ CORSIM vs. Δ Field

7. Analysis of Uncertainty

A more exacting treatment of validation requires closer attention to:

- *uncertainties* inherent to the simulator as well as from parameter estimates used to define input distributions;
- the *dual-use* of data for both estimating input parameters and for validating;
- *multiplicity* questions arising from the use of multiple evaluation functions (for example, the multiple link/approaches in Tables 1&2).

These issues can be addressed through a Bayesian analysis. For instance, in the test-bed example, the first uncertainty question can be dealt with by specifying prior distributions for the λ's in the equation in section 2 as well as for the probabilities p of turning movements. Posterior distributions of λ,p can then be computed given field data. Before each CORSIM run, a draw from the posteriors can be made, leading to a selection of λ,p which then provides the needed inputs for the run. The resulting variability in 100 runs, say, will then incorporate both the inherent CORSIM variability as well as the uncertainty stemming from the use of the field data in estimating λ,p. Such an analysis is underway by a research team at NISS and Duke.

Deeper Bayes methods based on Bayarri & Berger (1999) can be deployed to treat the impact of dual-use of the data. This is a topic of ongoing research.

Treatment of multiplicity requires appropriate formulation. Methods as described in Westfall and Young (1993), used in Williams et al. (1999), or False Discovery Rate approaches (Benjamini and Hochberg, 1995) are not clearly applicable because of the high level of dependence among the evaluation functions.

The application of these methods demands introducing loss structures that take into account "practical significance". For example, a difference of 5 seconds in stop-time may be minor, while a difference of 15 seconds may be major. A good starting point may be a comparison of the field and (CORSIM) predicted LOS. Criteria for LOS based on stopped time are captured in the 1994 Highway Capacity Manual and are shown in Table 9.

Table 9 LOS Designation in the Highway Capacity Manual (1994)

8. Conclusions

There are two sets of conclusions. The first set in 8.1 is about the validation process, the second, in 8.2, is about the specific test-bed model, CORSIM

8.1 Validation Process

The validation process has five key elements: *context, data, uncertainty, feedback*, and *prediction*. Context is critical: it drives the formulation of evaluation functions or performance measures that are ultimately the grounds on which validation must take place and affect interpretations of uncertainty. For example, statistically significant disparities may, in the context of an application, be practically insignificant. In addition context and the specified evaluation functions can affect the selection or collection of data, both field and model output, to be used for evaluation. Conversely, the availability or feasibility of data collection can determine the choice of evaluation functions. These factors may then converge in the calculation of uncertainties stemming from noisy data and model imperfections. The outcomes of the evaluations and the associated uncertainties point to possible flaws in the models and feedback to model adjustments that correct or, perhaps, circumvent the flaws. Ultimately, it is through prediction that validation of a model is reached.

The process we described is effective and applicable in general. Of course, implementing the particulars, done for the most part in the test-bed example, will require filling a number of gaps, most specifically in determining uncertainties but also in designing data collection, assessing the impact of (lack of) data quality, and detecting flaws.

8.2 Test-Bed

Two questions were addressed: Does CORSIM represent “reality” when properly calibrated for field conditions? Does CORSIM adequately predict traffic performance under revised signal plans?

Comprehensive calibration of CORSIM is infeasible; there are just too many parameters that can (and some that cannot) be calibrated with field data. Our approach was to focus on key input parameters such as external traffic demands, turning proportions at intersections, and effective number of lanes (due to illegal parking) and the like using CORSIM default values for other inputs.

We found that CORSIM was effective yet flawed. A major difficulty is CORSIM’s propensity to turn spillback into grid-lock; inadequately described driver behavior led to intersection blockage far too frequently. CORSIM does not accurately model lane distribution of traffic: lane selection in reality was much more skewed than in CORSIM. CORSIM tends to stop more vehicles than indicated in the field: in reality drivers “coast” to a near-stop then slowly accelerate through the signal, but the behavior is much more abrupt in CORSIM.

The first of these flaws was circumvented by modifying the network. The second flaw had some effect but a relatively minor one. The third flaw manifested itself in disparate stop-rates but did not seriously affect stopped time per vehicle stopped (STVS).

Overall, despite the shortcomings, CORSIM effectively represented field conditions. Even when the field observations lie outside the domain of the CORSIM distributions, as in Figures 2 and 3, there is virtually no difference in the estimated levels of service (Table 9) between the field and CORSIM – practically insignificant even if statistically significant.[5]

The predictability of CORSIM was assessed by applying revised (September) signal plans to the May traffic network. CORSIM estimates of STVS were reasonably close to field estimates and, indeed, the CORSIM LOS was the same as in the field. More importantly, CORSIM successfully tracks changes in traffic performance over time: on five links for which field data were available, two links exhibited a reduction in STVS, one link an increase, and two had no significant change -- CORSIM’s predictions were the same.

In summary, a candid assessment of CORSIM is that with careful calibration and tuning CORSIM output will match field observations and be an effective predictor.

[5] The CORSIM distribution does not reflect the additional uncertainty induced by the field data estimates of model input parameters; so statistical significance here is overstated.

9. Acknowledgements

The authors are very grateful for the support and assistance of Mr. Christopher Krueger and Mr. Thomas Kaeser of the Chicago Department of Transportation. We also thank the Urban Transportation Center at the University of Illinois at Chicago for their help in the data collection. This research was sponsored in part by grants DMS-9313013 and DMS-9208758 from the National Science Foundation to the National Institute of Statistical Sciences.

10. References

Bayarri, M. J., and J. O. Berger (1999). Quantifying surprise in the data and model verification. In *Bayesian Statistics 6*, 53-82. (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, etds.). Oxford University Press, London.

Benjamini, Y. and Y. Hochberg (1995) Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, 57, 1, 289-300.

Berliner, L. M., J. A. Royle, C. K. Wikle, and R. F. Milliff. (1999). Bayesian methods in the atmospheric sciences. In *Bayesian Statistics 6*, 83-100. (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, etds.). Oxford University Press, London.

FHWA (1997). CORSIM User’s Manual, U.S. Department of Transportation.

Lall, B., K. Dermer and R. Nasburg. (1994). Vehicle Tracking in Video Image: New Technology for Traffic Data Collection, Proceedings of the Second International Symposium on Highway Capacity, R. Akcelik, Ed., Sydney, Australia, pp. 365-383.

Lynn, N., N. Singpurwalla and A. Smith (1998). Bayesian assessment of network reliability. *SIAM Review 40*, 202-227.

Park, B., N. Rouphail, J. Hochanadel, and J. Sacks, Evaluating the Reliabilit of TRANSYT-7F Optimization Schemes, ASCE Journal of Transportation Engineering, Forth coming.

Westfall, P. H. and S. Young (1993) Resampling based multiple comparison procedures. Wiley-interscience.

Williams V. S. L., L. V. Jones and J. W. Tukey (1999) Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. J Educ Behav Stat 24: (1) 42-69.

Table 1. Comparison of throughput on internal network (veh/hr)

Period	Direction	Field (veh)	CORSIM (veh)	
			Avg	s.d.*
8-9am	In	11805	11895	48.1
	Out	11330	11877	52.8
	Net	-475	-18	-
4-5pm	In	10834	10805	39.7
	Out	10990	10796	40.6
	Net	156	-9	-
5-6pm	In	11431	11449	61.9
	Out	11756	11422	71.9
	Net	325	-27	-

Note: 1) Field data obtained from video on May 25, 2000.
2) S.D. is the estimated (from 100 runs) standard deviation of a CORSIM run.
3) Averages are rounded to nearest integer.

Table 2. Comparison of throughput on selected key links (veh/hr)

Period	Link	Field (veh)	CORSIM (veh)	
			Average	s.d.*
8-9am	SB LaSalle at Ohio	1651	1641	30.3
	EB Ohio at LaSalle	2790	2894	38.9
	SB Wells at Ohio	693	694	17.4
4-5pm	EB Ohio at Orleans	1948	1947	2.3
	NB Orleans at Ohio	1498	1489	25.0
	NB LaSalle at Ontario	1500	1478	28.0
5-6pm	EB Ohio at Orleans	1897	1896	2.5
	WB Grand at Wells	1204	1133	21.9
	NB LaSalle at Ontario	1636	1617	26.3

Note: 1) Field data obtained from video on May 25, 2000.
2) S.D. is the estimated (from 100 runs) standard deviation of a CORSIM run.
3) Averages are rounded to nearest integer.

Table 3. Comparison of stop rates on key links

Period	Link	Field (%)	CORSIM (%)	
			Average	s.d.*
8-9am	SB LaSalle at Ohio	50	30	1.8
	EB Ohio at LaSalle	56	35	3.9
	SB Wells at Ohio	99	94	1.2
4-5pm	EB Ohio at Orleans	50	59	1.0
	NB Orleans at Ohio	51	56	2.9
	NB LaSalle at Ontario	42	47	3.6
5-6pm	EB Ohio at Orleans	48	59	1.0
	WB Grand at Wells	55	53	3.3
	NB LaSalle at Ontario	78	62	3.4

Note: 1) Field data obtained from video on May 25, 2000.
2) S.D. is the estimated (from 100 runs) standard deviation of a CORSIM run.

Table 4. Comparison of key-links STVS (stop time per vehicle stopped)

Period	Link	Field (sec/veh)	CORSIM (sec/veh)	
			Average	s.d.*
8-9am	SB LaSalle at Ohio	27.8	32.3	1.8
	EB Ohio at LaSalle	15.4	18.6	0.8
	SB Wells at Ohio	33.1	39.6	0.4
4-5pm	EB Ohio at Orleans	18.4	18.7	0.3
	NB Orleans at Ohio	20.6	20.6	1.7
	NB LaSalle at Ontario	33.5	27.9	3.0
5-6pm	EB Ohio at Orleans	15.2	18.7	0.3
	WB Grand at Wells	8.3	10.5	2.1
	NB LaSalle at Ontario	33.4	34.2	2.9

Note: 1) Field data obtained from video on May 25, 2000.
2) S.D. is the estimated (from 100 runs) standard deviation of a CORSIM run.

Table 5. Field-measured throughput comparison at key links

Period	Link	May (veh)	Sep. (veh)
8-9am	SB LaSalle at Ohio	1650	1441
	EB Ohio at LaSalle	2790	2798
4:30-5:30pm	NB LaSalle at Ontario	1607	1696
	NB Orleans to Freeway	838	899
	NB Orleans at Ontario	1051	1107

Table 6. Comparison of stop rates on key-links

Period	Link	Field (%)	CORSIM (%)	
			Average	s.d.*
8-9am	SB LaSalle at Ohio	52	30	2.7
	EB Ohio at LaSalle	37	38	3.0
4:30-5:30pm	NB LaSalle at Ontario	36	51	4.2
	NB Orleans to Freeway	53	74	4.1
	NB Orleans at Ontario	47	43	2.0

Note: 1) Field data obtained from video on September 27, 2000.
2) S.D. is the estimated (from 100 runs) standard deviation of a CORSIM run.

Table 7. Comparison of STVS (stop time per vehicle stopped)

Period	Link	Field	CORSIM (sec/veh)	
		(sec/veh)	Average	s.d.*
8-9am	SB LaSalle at Ohio	16.9	22.0	2.0
	EB Ohio at LaSalle	15.2	21.6	1.2
4:30-5:30pm	NB LaSalle at Ontario	26.4	24.8	1.8
	NB Orleans to Freeway	31.4	51.9	7.6
	NB Orleans at Ontario	21.9	24.0	1.0

Note: 1) Field data obtained from video on September 27, 2000.
2) S.D. is the estimated (from 100 runs) standard deviation of a CORSIM run.

Table 8. Δ CORSIM vs. Δ Field

Link	ΔCORSIM	ΔReality
EB Ohio at LaSalle	0	3
SB LaSalle at Ohio	-11	-10
NB LaSalle at Ontario	-9	-5
NB Orleans to Freeway	13	15
NB Orleans at Ontario	1	-2

Note: Δ = STVS[September] – STVS[May]

Table 9 LOS Designation in the Highway Capacity Manual (1994)

Level of Service	Stopped Time per Vehicle (STV; sec/veh)
A	STV ≤ 5
B	5 < STV ≤ 15
C	15 < STV ≤ 25
D	25 < STV ≤ 40
E	40 < STV ≤ 60
F	STV ≥ 60

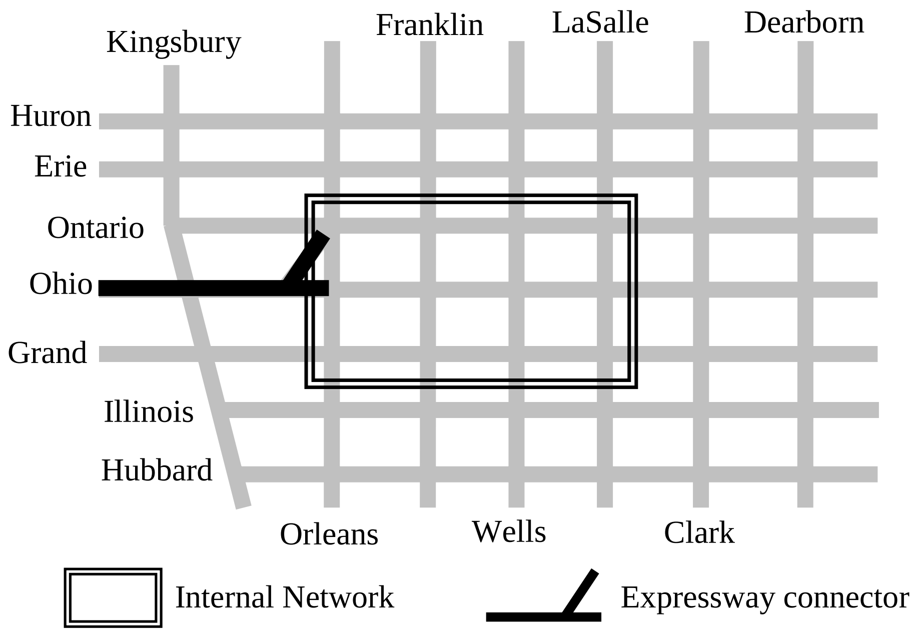

Figure 1. TEST-BED NETWORK

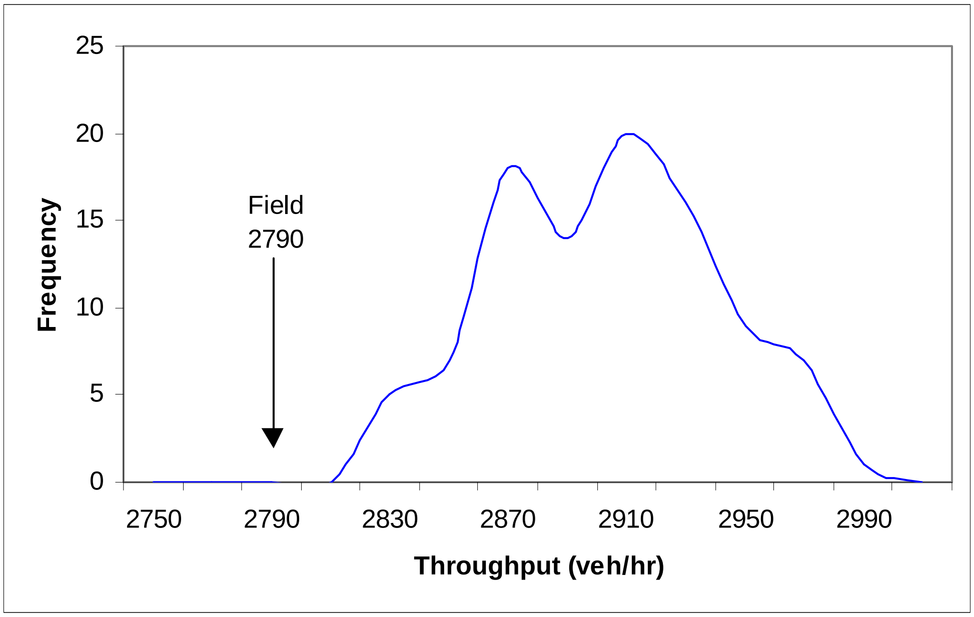

Figure 2. Link throughput at EB Ohio/LaSalle (8-9am)

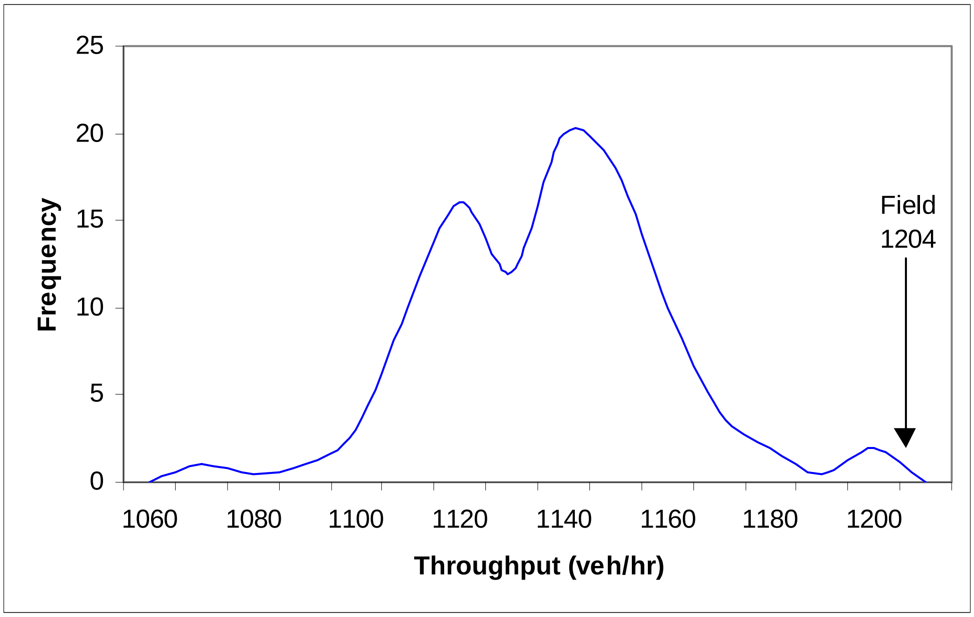

Figure 3. Link throughput at WB Grand/Wells (5-6pm)

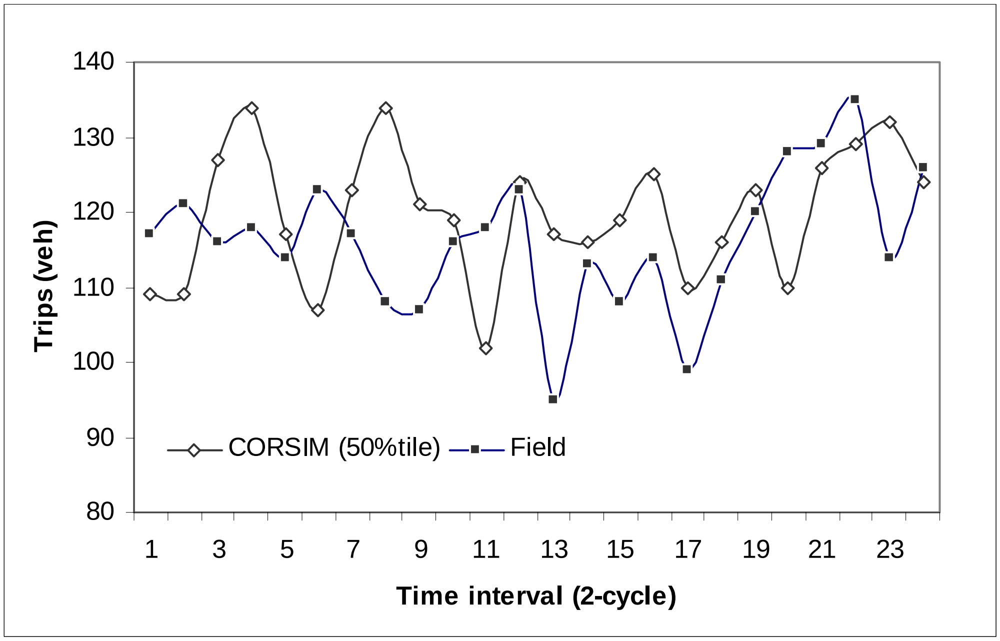

Figure 4. Comparison of CORSIM vs. Field variation (EB Ohio at LaSalle, AM Peak)

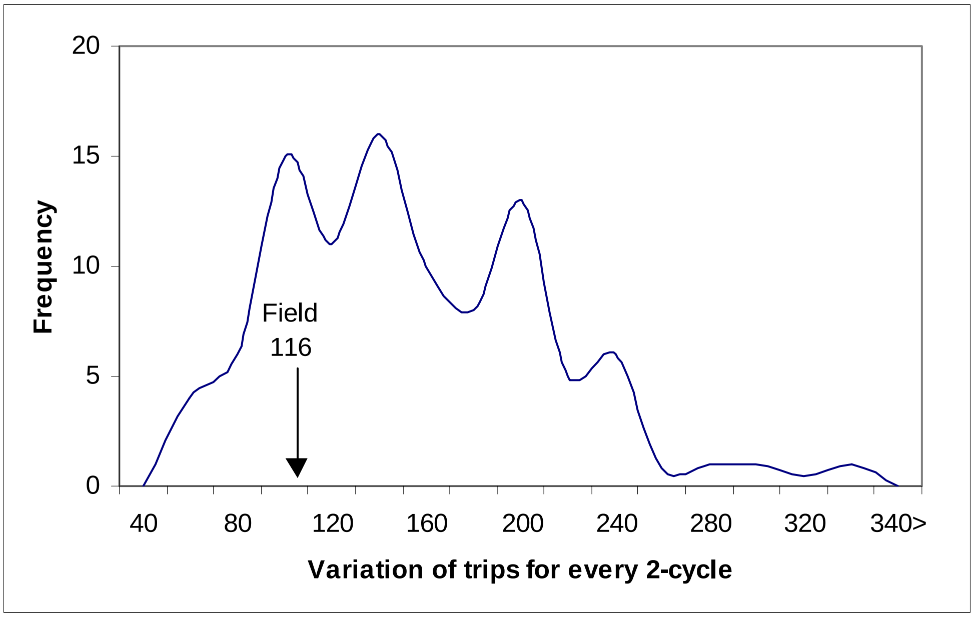

Figure 5. CORSIM variation vs. field (EB Ohio at LaSalle, AM Peak)

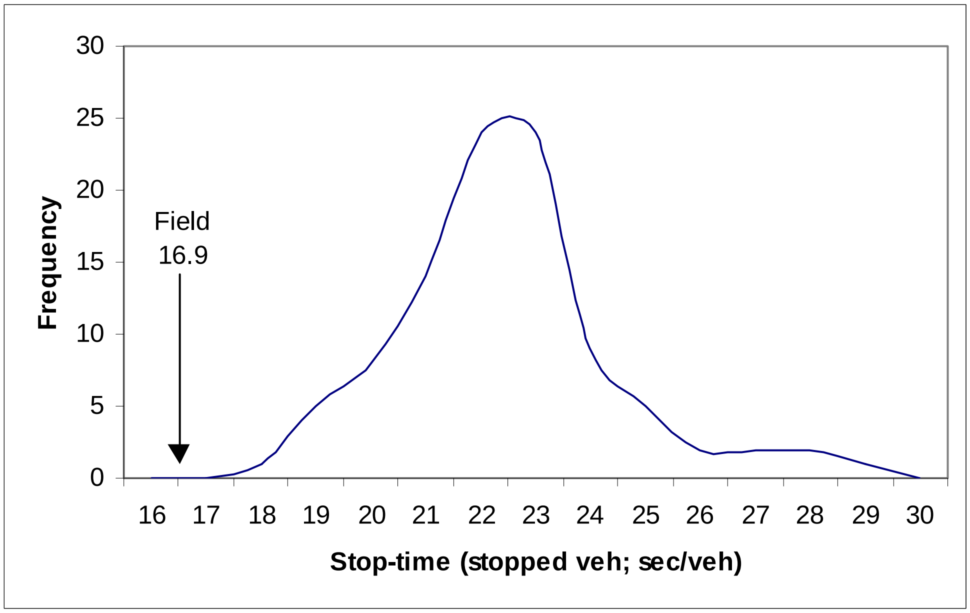

Figure 6. Link STVS at SB LaSalle/Ohio

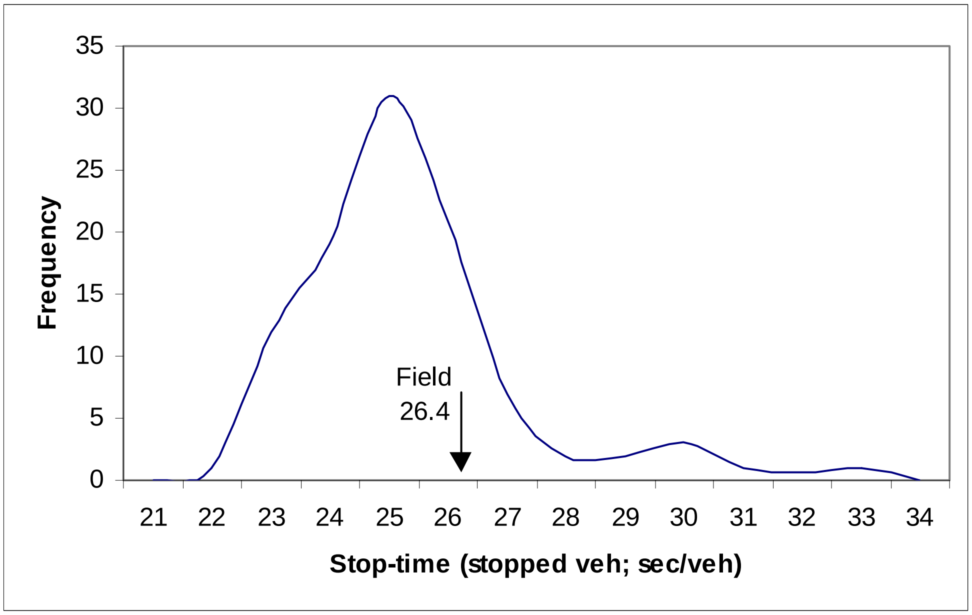

Figure 7. Link STVS at NB LaSalle/Ontario