Computer Simulations as Experiments: Using Program Evaluation Tools to Assess the Validity of Interventions in Virtual Worlds

Richard Berk
Department of Statistics
Jason Bond
Department of Statistics
Rong Lu
Department of Atmospheric Sciences
Richard Turco
Department of Atmospheric Sciences
Robert Weiss
Department of Biostatistcs

UCLA

February 15, 1999

1 Introduction

The concepts and tools developed by Donald Campbell and his colleagues have quite properly dominated the field of evaluation research. Internal validity, external validity, and later, construct validity and statistical conclusion validity, for instance, are central concepts on which a variety of evaluation research methods depend (Rossi and Freeman, 1993). However, these methods have been developed to understand phenomena that may be empirically observed. In this chapter, we broaden their applicability by considering evaluations undertaken in the virtual worlds manufactured by computer simulations.

For the past year, a research project undertaken jointly by members of UCLA's Departments of Statistics, Biostatistics and Atmospheric Sciences has been developing statistical tools to evaluate a large scale computer model of meteorology and air quality in the Los Angeles Basin. The computer model is meant to contribute to basic research undertaken within the discipline of atmospheric science. More important for our purposes, it is also a policy tool with which one can undertake virtual experiments. The basic idea is to run the computer model under control conditions (usually the extant situation) and then under an experimental condition. Impacts of the intervention are represented in the differences between the two sets of computer output. For example, one could simulate the impact on air quality in an urban area of planting a very large number of trees.[1] One computer run might seek to represent smog concentrations under current land use and another computer run might seek to represent smog concentrations if a million trees were planted. Differences in smog concentration between the two sets of output could then be used to evaluate the impact of the tree planting.

But like real experiments designed to evaluate policy interventions, there is always the key question of how valid the virtual experimental results really are. All of the usual kinds of validity concerns are relevant: the quality of causal inference, generalizability, proper consideration of uncertainty, and the accuracy of the data brought to bear. But because it is a virtual world rather than an observed world that is at issue, the approach to evaluation is necessarily somewhat different. Indeed, at first blush it is not at all apparent how to assess the validity of virtual experiments. To set the stage for this discussion, we first consider the nature of the virtual worlds, constructed to serve as scientific and program evaluation tools.

2 Virtual Worlds

In the conventional view of program evaluation, one can use statistical tools to provide an empirically-based model of a program and its setting. From such models, it is then possible to summarize program process and impact.

[1] Trees can improve air quality by removing from the air some smog precursors. Through shading and evapotranspiration, they can also constrain the relevant atmospheric chemistry by reducing ambient temperatures.

And ideally, a variety of diagnostic procedures are brought to bear on the statistical model to help insure that it usefully represents key features of the real world. Thus, one might undertake an observational study of the impact of a health care program using propensity scores to address selection into experimental and comparison groups. These scores could serve as matching variables to help make the experimental and comparison groups comparable. Tests between weighted means might follow. The robustness of the findings could then be considered by applying sensitivity tests making a variety of assumptions about the selection process (Rosenbaum, 1995).

For much work in the physical and life sciences, however, well-developed and widely-accepted theory allows the construction of virtual worlds capturing key features of the phenomena in question.[2] These virtual worlds are built from computer code meant to simulate the real world. Thus, there are virtual climate systems, virtual epidemics, virtual ecosystems, and virtual communications networks. For example, General Circulation Models (GCMs) are meant to represent dynamic variation in global climate from which the role of anthropogenic (i.e., human-produced) greenhouse gases can be studied (Barron, 1995). These models are used for many purposes including conducting virtual experiments about the impact on climate should humans double atmospheric carbon dioxide concentrations over the next century.

In principle, working in virtual worlds has a number of scientific benefits. First, virtual worlds can be quite simple, or at least far more simple than the empirical world. Simplicity is one key to understanding insofar as key features of the relevant empirical world are captured nevertheless.

Second, virtual worlds are fully manipulable. Consequently, one has complete control to construct the particular worlds of one's choosing. This includes control over what to manipulate and what to hold constant that no real-world laboratory, let alone field setting, could every match. Indeed, this control allows for causal inferences that in principle exceed even the gold standard of randomized experiments. For example, if the study units are grid points on a map, the grid points for the experimental run will differ from the same grid points for the control run *only* in ways that are determined by the researcher.

Third, realizations of virtual worlds are replicable. That is, the computer code can be run over and over under varying conditions to see what the

[2] Economists also commonly build computer models, but there is far less consensus about the underlying theory. Economic models will not be considered in this paper.

implications are for the output. Among other things, this allows for a number of different experiments that vary in known ways. In effect, one can undertake replications to determine how well a particular set of results generalizes.

Finally, virtual worlds are often deterministic: given the model, input fully determines output.[3] This is useful, because uncertainty (broadly defined) is reduced as a result. That is, the signal can be more easily separated from the noise.

In practice, however, some of these useful characteristics of virtual worlds are compromised. While simplicity is desirable, there is often pressure to trade simplicity for a more complete rendering of the empirical world. A model with five equations, for instance, can soon become a model with one hundred equations. This, in turn, undermines control because feedbacks and other non-linear relationships make it very hard to know how changing one part of the computer code will affect the manner in which the rest of the code performs. Another price is replicability. At some point, the complexity of the numerical calculations make computer runs very lengthy and costly, and in some cases, totally impractical. Finally, complexity can make deterministic models behave like stochastic models. Very small differences in initial conditions can lead to large and unanticipated differences in output. In short, as virtual worlds are made more complex in an effort to better reproduce the empirical world, they can undermine the rationale for constructing virtual worlds to begin with. Thus, there always needs to be a balance between complexity and tractability.

3 Evaluation Research in Virtual Worlds

The scientists who construct these virtual worlds usually are well aware of potential imperfections. Some of these problems result from the need to simplify and some from errors in how the simplified world is represented. In either case, simulation output can be compared to empirical observations and to expectations from accepted theory. If the simulation output seriously contradicts either, model revisions often will follow. For example, the output of GCMs is often compared to data from ice cores that can be used to char-

[3] If, however, it is desirable to simulate stochastic processes, deterministic processes can be constructed that behave in nearly the same fashion as stochastic processes. Indeed, at some point the distinction between what is ”really” deterministic and what is ”really” stochastic can break down.

acterize climate variation over many thousands of years in the past. If well known cycles of warming and cooling are not reproduced, the GCM may be substantially altered (Barron, 1995). There sometimes also are forecasting tests, such as recently conducted for computer models of the El Niño phenomenon (Syu and Neelin, 1997). Weak forecasting performance can often mean problems with the computer model.[4] Finally, "experiments" commonly are undertaken with the simulations themselves in which input data are perturbed or in which small parts of the computer code are manipulated to see whether the model is overly sensitive to such changes. Such studies are common in regional scale modeling of climate impacts (Bonan, 1997).

These kinds of model assessments highlight a critical distinction between the relative importance of theory and data when working in a virtual world compared to working in the empirical world. In the observed world with which program evaluators are familiar, flesh and blood people introduce actual programs whose content and outcomes are literally measured. From these measurements come data summaries and data-informed models through which inferences are drawn. Theory, broadly construed, is a junior partner.

In the virtual world, theory transformed into computer code is used to simulate what might happen in the observed world. The outputs from the simulations are the "measurements" of interest from which conclusions are drawn. Actual measurements, in contrast, play a decidedly secondary role.

First, measurements may be used to "initialize" the simulation. If, for example, one has a computer model of the atmospheric chemistry that produces smog, at the very beginning of the simulation one might introduce actual readings of local pollutant concentrations. In effect, this primes the simulation pump.

Second, actual measurements are sometimes used to help "parameterize" the simulation model. When there is insufficient theory to fully specify the model, empirical values and relationships sometimes are inserted. These are seen as temporary fixes until the science needed to fill the gaps is completed.

Finally, as noted above, actual measurements can be used to help assess the quality of computer simulation models. In the most direct instance, the simulation output can be compared to "ground truth" represented by real data. This is not to say, however, that comparisons between data and

[4] For some real-world systems, however, highly non-linear relationships produce chaotic results that greatly limit how far any computer model can see into the future (Briggs and Peat, 1989).

computer output are easy to make. Typically, the data and the computer output are arrayed in both time and space. In effect, there are gridded maps for various moments in time, with data and output in each cell. The trick is to drawn conclusions not just about how well the model fits the data overall (e.g., using the mean of squared deviations between the model output and the data), but how well it fits in different time periods and in different locations. In effect, large matrices are compared so that the fit at different times and places can be judged.[5]

Such assessments of computer simulations apply with equal force when the computer simulations are used to conduct virtual experiments that are in effect, virtual program evaluations. A computer model that cannot reproduce a known world cannot be trusted to properly assess the impact of some intervention in the status quo. In other words, if the impact of the control condition is not captured with sufficient accuracy, all comparisons to the impact of the experimental condition are necessarily flawed. At the very least, the control condition is not properly represented.

3.1 Building Competitive Statistical Models

Another interesting, challenging and potentially instructive approach for computer model assessment is to build statistical models from real data and then compare the performance of the statistical models to the performance of the simulation models. We turn to that strategy now.

There are two broad approaches one may employ to bring empirically based statistical models to bear validity of such simulations. The first requires that one duplicate the functional relationships between simulation inputs and outputs. The second rests on developing a statistical model to ”compete” with the computer simulation model.

3.2 Duplicating the Functions in the Computer Model

For ease of exposition, assume a single input vector x and a single output vector y. The input vector is data, and the output vector is a product of the simulation. For simplicity, we will also assume they are related through a single differential equation.

[5] One of the products of our work will be several new statistical procedures to compare such matrices.

Now, suppose it were possible to develop a statistical model relating x to y so that the y constructed from the statistical model is identical to simulated y. In a very simple case, one might regress y on a relatively low order polynomial in x and obtain an R^2 of effectively 1.0. Then, $\hat{y}$ would be virtually the same as the y.

And now the point: if a computer simulation model and a statistical model have the same inputs and same outputs, they are by definition the *same* function. We are not saying that a simple polynomial and a differential equation are the same. What we are saying is that they transform x into the same y. Formally, therefore, the two functions are the same.

It then follows that one can use the statistical model *in place of* the computer simulation model for a wide variety of purposes. For example, it may be impossible to obtain sensible estimates of uncertainty from a computer simulation, in part because the model is deterministic. But estimates of uncertainty may perhaps be readily obtained from a statistical model and then applied to the simulation model. In the language of evaluation research, statistical conclusion validity for the simulation can now be far more effectively addressed.[6] And for a virtual evaluation, it would be addressed for the control run, the experimental run, and comparisons between the two sets of outputs. In other words, one might be able to apply a conventional t-test for the intervention effect using information from the statistical model applied to the output from the computer simulation model.

3.3 Developing a Competing Statistical Model

The second strategy is not to duplicate the simulation function, but to develop a statistical model to serve as a competitor. One now builds an empirically based model relating observed inputs to observed outputs (not the simulation outputs). Thus, the inputs may be different from those used in the computer simulation. For those inputs that are the same, their functional relationships to the empirical outputs may be different.

Broadly stated, the purpose of developing a competitive statistical model is to see if there is information in the empirical world that is not well represented in the computer simulation. For example, one might regress ozone

[6] One can also use this approach to link the output of the first model iteration to the output of the last model iteration, if the primary goal is to duplicate the function implied by the computer code.

concentrations on spline functions of temperature and wind speed.[7] The shape of the functional relationships would then be heavily determined by the data. And one might find a disparity between the statistical model and the computer model in the relative importance of these two inputs and in their functional relations with the output. It generally would not be clear which model was "right;" indeed, notions of "right" and "wrong" probably would make little sense. But there could well be useful clues about ways the computer model might be improved.

Improving the computer model could be very important if temperature, for instance, is a key policy variable. And it would be if the goal of the intervention was to reduce atmospheric ozone by reducing ambient temperature. As noted above, planting a large number of trees could well have this effect. Thus, a virtual evaluation of the impact of tree planting would depend fundamentally on getting the link between temperature and ozone production right.

There is not space in this chapter to illustrate both broad strategies: duplicating the function embodied in the computer model and competitive model building. Fortunately, either approach would suffice to make our general point about the broader application of concepts and tools from evaluation research. But because the strategy based on developing a competitive statistical model will seem less foreign, we will concentrate on that.

4 An Illustration: Building Statistical Models of Air Quality

We turn now to the research project aimed at improving regional meteorological and air quality computer models. The basic approach has been to develop descriptive statistical models for the spatial and temporal distribution of air pollutants that may be compared to structural models of the same phenomena, based on the underlying microphysics and chemistry.

The computer model we have been using as a benchmark is the SMOG integrated air pollution modeling system developed by Lu et al. (1997a). In effect, the model captures key features of the processes that manufacture and transport air pollutants in the Los Angeles basin. Some details are

[7] A spline function is a collection of pieces of smooth curves. The pieces are joined end to end at points called "knots," subject to smoothness constraints at the knots.

provided in the next two paragraphs, which can be skipped if the terms are too unfamiliar.

4.1 The SMOG Computer Model

The modeling system consists of four major components that govern the regional meteorology, pollutant transport and dispersion, chemical and microphysical transformations, and solar and terrestrial radiation transfer. The meteorological component solves a set of partial differential equations for fluid dynamics and predicts the wind, temperature, humidity, and turbulence fields over the model domain. These variables are used to calculate the movement of pollutants across model grids by solving dispersion equations for each pollutant represented in the model, including pollutant emissions from the surface as well as physical removal processes. The chemistry and aerosol component computes the rates of chemical and microphysical transformation of the pollutants within the model grid cells. The radiation transfer component determines the solar and terrestrial rediative heating rates to force the meteorology prediction and the photodissociation rates for chemistry calculation.

The computer model characterizes meteorology and air quality for the Los Angeles basin using over 200 differential equations. The relationships represented are highly non-linear. Model calculation starts from the initial values for weather conditions and atmospheric pollutant concentrations. Marching forward in time, the model is forced by time-dependent solar and terrestrial radiative heating, anthropogenic and natural emissions of air pollutants, and lateral boundary conditions of the model domain. The predictions of local meteorology and air pollutant concentrations are produced for grid cells in three spatial dimensions and one temporal dimension. The model is run in temporal steps of a few seconds for meteorology and five minutes for chemistry covering a typical air quality episode of three days to a week.

To evaluate model predictions, outputs from the computer model are usually compared with observed values measured at monitoring stations, when they are available. In the Los Angeles basin, pollutants are routinely measured at about 37 monitoring stations operated by the South Coast Air Quality Management District (SCAQMD). The pollutants that have been measured include ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), and nitric oxide (NO). Lu et al. (1997b), for example, used similar data to compare the SMOG model predictions to observations collected during the

Southern California Air Quality Study (SCAQS, 1987).

4.2 Competing Statistical Models

Once again, the purpose of the statistical approach was to develop a competing model. The competing model, in turn, would be used to help assess the quality of the computer simulation model. If the computer simulation model could not generate an accurate description of ozone concentration over time in the Los Angeles basin, it would be potentially misleading as the control run in any virtual evaluations of interventions meant to reduce ozone concentrations. Massive tree planting is one illustrative intervention. New controls on tail pipe emissions on cars and trucks is another.

The statistical models we will consider are based on data from the 37 monitoring stations, or more typically, a subset of them for which missing data are not a serious problem.[8] These are the cross-sectional units. The data are aggregated to provide hourly averages; hours are the temporal units commonly employed in this field.

A variety of statistical tools, described later, were applied to arrive at satisfactory descriptive models. Here, we report results for each of the four air quality measures. Explanatory variables include lagged values of the pollutant in question[9], wind speed, wind direction, temperature, and various regions defined as having similar air quality characteristics.

We will provide not only descriptive output from the statistical models, but information that can be used for statistical inference. However, the data are not a probability sample of any specified population. Consequently, for conventional (i.e., non-Bayesian) inference, a super population must be specified so that the data are sensibly conceptualized as a random sample or "realization" (Berk et al., 1995). Since regional meteorology and air quality are highly replicable phenomena, a formulation of this sort is not an enormous stretch.[10]

[8] It is common, for example, for temperature measurements, which are one of our key explanatory variables, to be missing.

[9] A large fraction of an air pollutant present one hour remains for the next. In addition, the amount of pollutant often affects the subsequent atmospheric chemistry. Given these two processes and hours as our temporal units, we used a one hour lag.

[10] In this case, the super population would be full set of samples that could have been produced in principle by the underlying processes affecting the regional meteorology and atmospheric chemistry. The data are then treated as a random draw from that set.

4.3 Results for Ozone

Much of our initial statistical work was done using three days of data from August 26 through August 28, 1987. One of the key reasons for choosing this period is that it is also the time interval for which there are results from the meteorological model, later used for purposes of comparison. All of the analyses were done using daylight hours (i.e., from 6 AM to 6 PM), because sunlight is a key driver in the chemistry of ozone production.

Consider the results for the 27th, although the story is much the same if other days are used. Ozone concentrations range from (effectively) zero parts per hundred million (pphm) at sunrise to a high of 24 pphm in the middle of the afternoon. The mean value is 7.0, and the standard deviation 6.0. Since the median is 5.5, it is clear that the ozone distribution is skewed to the right.

We began by exploring the possible relationships between our "input" (explanatory) and "output" (response) variables using non-parametric nonlinear multivariate procedures with both single response variables (Young et al., 1976) and multiple response variables (van der Burg and de Leeuw, 1983), and based on b-splines (Gifi, 1990: 365-370).[11] The results strongly favored linearity as a good first approximation. There was also evidence from scatterplot matrices (Cook and Weisberg, 1994: 47-53) that the relationships between the explanatory variables were approximately linear.[12] We proceeded, therefore, with linear least squares, buttressed by various diagnostics and alternative procedures where necessary. This enormously simplified the computational task and interpretation of the results.

Table 1 shows the results from a linear regression analysis (ordinary least squares) in which ozone concentrations for each of 12 hours at 20 monitoring sites is regressed on the ozone concentrations one hour earlier (i.e., lagged by one hour), temperature (centigrade), and the wind in meters per second in the eastward and northward direction.[13] The fit is quite good. The R^2 is

[11] For the multiple outputs, we used all of our air quality constituents at once. For the single output, we used only ozone concentrations. The inputs were temperature, wind speed and direction, for some models, values of the outputs in question one hour earlier. The basic goals of this analysis were exploratory and in effect, we let the data tell us what functional forms fit best.

[12] It is too often overlooked that linear regression requires that the regressors be related linearly to one another (Cook and Weisberg, 1994).

[13] Wind speed and direction are represented by orthogonal gradients. Zero degrees is due north. We computed the gradients by multiplying the wind speed by the cosine of the

Variable	Coefficient	Stand. Error	t-value
Constant	-8.60	0.69	-12.43
Ozone Lagged	0.66	0.02	27.45
Temperature	0.43	0.03	14.58
Wind East	-0.15	0.14	-1.06
Wind North	-0.27	0.07	-3.93

Table 1: Linear Regression of Ozone Concentations on Lagged Ozone, Temperature, and Wind (N=240)

.91, and the estimated standard deviation of the residuals is 1.73.[14]

Equally important, the coefficients seem to make sense. Perhaps most striking is the lagged impact of ozone. For example, 10 pphm at any given hour contributes nearly 7 pphm to the next hour's concentration. One logical explanation is that ozone is relatively stable over time and space. Another possible explanation is that existing ozone provides a positive feedback to later ozone production. But as a substantive matter, the lagged value was clearly a vital part of the story and not a mere statistical convenience.

On the average, with every degree increase in temperature (centigrade), ozone concentrations increase by .43 pphm. This too makes sense since sunlight and temperature are positively related to ozone production. And the size of the effect has deceptively important policy implications. If, for example, planting a large number of trees could reduce ambient temperature about 5 degrees, ozone concentrations would decline by about 2 pphm on any give hour, given the amount of ozone "carried over" from the previous hour. This may not seem like a large effect, but about 70% of that is in addition removed from the next hour. In other words, the reductions compound.[15]

Increasing wind velocity in both eastward and northward directions seems to reduce ozone concentrations. Each meter per second increase in wind speed in an eastward direction reduces ozone by about 15 pphm and in a northward direction by about 27pphm.[16] Yet, the wind effect in the eastward direction

wind direction angle (in radians) for the east/west direction, and by the sine of the wind direction angle (in radians) for the north/south direction.

[14] We will consider the possible spatial and temporal dependence later.

[15] For example, $-2 + .70(-2) + .70 .70(-2) + \ldots$

[16] To help put these results in proper context, temperature ranges from 17 to 38 degrees C, the eastward wind gradient ranges from -2.7 to 2.7 meters per second, and the northward

is small relative to its standard error.

4.3.1 Model Diagnostics Based on Fit

The model's residuals look promising. Graph 1 shows a histogram of the residuals with a kernel smoother overlaid. [17] The distribution is nearly symmetric, with no evidence of outliers. Graph 2 shows a plot of the residuals against the fitted values suggesting that there are no obvious problems with outliers or leverage points and that, therefore, the linear fit looks reasonable. Clearly these are encouraging diagnostics. On the other hand, Graph 2 also shows that the variance of the residuals increases somewhat with the fitted values suggesting that the model is less successful for the highest ozone concentrations. We confirmed this conclusion by examining a nonconstant variance plot in which it was apparent that the residual variance increased as an approximately linear function of the fitted values.[18]

Graph 1: Residuals for Ozone Analysis

wind gradient ranges from -6.1 to 3.0 meters per second.

[17] Here, the kernel smoother produces an estimate the residual distribution underlying the histogram.

[18] A nonconstant variance plot is a scatter plot with the square root of the absolute values of the residuals on the vertical axis and the fitted values on the horizontal axis.

Graph 2: Scatter Plot of the Residuals Against Fitted Values

To consider possible improvement in the model fit, we examined added variable plots for each of the explanatory variables. These plots are literally the residualized data from which the regression coefficients can be computed (Cook and Weisberg, 191-195).

In each graph, a linear regression line is overlaid, which is the regression line implied by the model in Table 1. The linear fit is generally reasonable, and there is no clear evidence that some nonlinear form is required. There is also no evidence of important leverage points or outliers affecting the linear fit. Yet to the eye, it might seem as if a quadratic term could improve the fit a bit for a number of added variables plots. In Graph 3, for example, the relationship with the lagged value of ozone seems to flatten out at higher ozone concentrations. However, the improvement when a quadratic term is added to the statistical model is very small and a consequence of a few leverage points. The case for a quadratic rather than a linear fit is, therefore, not very compelling. We had similar experiences with the other added variable plots; there were hints of reduced relationships at higher ozone values, but we could not definitively determine if they should be taken seriously. Nevertheless, we will return to possible nonlinear effects shortly.

Graph 3: Added Variable Plot for Lagged Ozone

Graph 4: Added Variable Plot for Temperature

Graph 5: Added Variable Plot for Wind in Northward Direction

Graph 6: Added Variable Plot for Wind in Eastward Direction

We further explored the credibility of the model by applying sliced inverse regression (Li, 1991). The goal was to see if a second (orthogonal) linear combination of the regressors might improve the fit. It was clear that the first dimension was necessary and that if a non-parametric fit was employed, the R^2 could be increased from .91 to .99. However, working with non-parametric fits for the added variable plots suggested that the non-linearities were significantly driven by a few leverage points and probably not worth taking seriously. There was also a hint that a second dimension could be exploited in the analysis, but again, it seemed driven by a few leverage points. In particular, while there was a strong, positive linear relationship in this second dimension between temperature and ozone overall, the strength of the relationship seemed to fall off at the highest temperatures when the relevant plots were examined. Indeed, there seemed to be almost no relationship between temperature and ozone at these temperatures (about 10 observations). Whether this apparent pattern is real or an artifact of the data is difficult to determine with so few extreme observations.

Nevertheless, we further experimented with several alterations in the model. For example, we tested quadratic and cubic terms for temperature and the lagged value of ozone. The fit improved, but not very dramatically. Similarly, we added three binary variables for three of four geographical regions in the Los Angeles air basin that shared similar air quality characteristics, but there was no substantial improvement in the model. In short, for ozone at least, a very simple linear formulation with four regressors seems to fit the data quite well.

Finally we examined the quality of the fit at different times and at different places. We found that the statistical model fit the data a bit better than the simulation model earlier in the day. Perhaps earlier in the day when the temperature is relatively low, the atmospheric chemistry represented in the simulation model is not as relevant as later. We also found that both models did not fit sites in the north east area of the basin as well as in other places. We suspect that both models fail to account for the transportion of air pollutants over the spatial boundaries where there are passes between the coastal mountain ranges.

4.3.2 Model Diagnostics Based on Forecasts

Another test of the model is to determine how well it forecasts. If nothing else, forecasting skill can be a good reality check, especially when there is the substantial possibility of overfitting. Such is the case here because of the highly empirical way in which the final model was developed.

We used the fitted values from the model for ozone concentrations on August 27th to forecast ozone concentrations on August 28th (called “lead ozone”). That is, the forecast is for 24 hours later. This is not the way one would forecast if the primary goal was to accurately anticipate the future. In that case, one would forecast one or more periods beyond which one had data, but in the same temporal units with which one were working. For our model, that would mean forecasting one hour ahead. But given the very strong relationship between ozone concentrations one hour and the next, we already knew that such forecasts would be highly accurate. A better antidote to overfitting would be to try to forecast one day ahead using the model's fitted values.

When the observed values for the 28th were regressed on the fitted values for the 27th, the R^2 was .88, and the standard deviation of the residuals was 2.45. For comparison purposes, we regressed the observed values for the 28th on the output from the meteorological computer model for the 27th. The R^2 was .83, and the standard deviation of the residuals was 2.66. On the average, the simple statistical model was able to "forecast" one day ahead as well, or perhaps a bit better, than the structural model.

We also examined in detail plots of the observed ozone concentrations on the 28th and forecasts of those values from both the statistical and structural model. Graphs 7 and 8 show the results with a least squares line overlaid on each. Note that the scales of the two graphs are a little different because of a few statistical predictions slightly less than zero, and a few relatively large predicted values from the meteorological model.

Overall, both sets of forecasts do quite well. Both also stumble a bit when they predict a few very high ozone readings when in fact, the ozone readings are in the middle ranges.[19] In short, there is no strong evidence of differential forecasting ability. For example, both models forecast the large ozone values with about the same level of accuracy. Perhaps the main concern with the statistical model is the few forecasts of small negative values. Exactly how best to handle this is unclear. When transformations were applied to eliminate negative forecasts, the overall fit declined. And the post hoc "solution" of simply declaring all negative forecast to be forecasts of zero, is at least inelegant.

What about any remaining dependence in the residuals? It would seem that the lagged value of ozone effectively addresses the temporal correlation and the spatial correlations, since ozone concentrations are strongly related at any given monitoring station over time. Efforts to improve the fit by adding regressors to address any remaining dependence were not very successful. Recall, for example, that including some binary variables for location did

[19] For what it may be worth, these largest errors were primarily for monitoring stations near San Bernardino.

not improve the fit much.

Finally, we were curious how well a statistical model would perform when we did not include lagged ozone values. In fact, the model fits rather well. The R^2 is .63, and the standard deviation of the residuals is 3.54. The pattern of results parallels those shown in Table 1, but the coefficients for temperature and wind in the northward direction approximately doubled. There is again a hint that these relationships flatten out at the very higest ozone concentrations, but a few leverage points are primarily responsible.

Graph 7: Ozone Predictions from Statistical Model

Graph 8: Ozone Predictions from Meteorological Model

4.3.3 Summary of Findings for Ozone

The findings from the statistical model are easily stated. Ozone concentrations increase with earlier increases in ozone concentrations and with increases in temperature. Ozone concentrations decrease with increases in wind speed in easterly and especially northward directions. The overall message is that in these variables, or what they represent, there is information that the current meteorological model may not be fully exploiting. At the same time, both the meteorological model and the statistical model do a good job of describing temporal and spatial variation in ozone concentration. Both models also forecast well 24 hours ahead. What may be surprising is that such a simple statistical sufficed.

From an program evaluation point of view, these conclusions have several important implications. First, it is clear that since the statistical model accounts for most of the variation in ozone, so does the computer model. The good fit for the computer model is one indication that one may well be able to take the computer simulations for ozone seriously. Second, the good forecasting ability of both the statistical model and the computer model is, likewise, an encouraging sign. Third, we confirm from the statistical model that temperature is one of the key drivers of ozone concentrations. This is vital for interventions such as tree planting, whose effects on ozone are through changing ambient temperature. In other words, we have not only a good descriptive model in the computer simulations, but also a model that can represent one of several key policy-relevant variables. Fourth, from the statistical model we have information that can be used to gain insight about the impact on the simulation results of sampling error. One cannot simply apply the statistical model's standard errors to the computer simulation output. But, insofar as it may turn out that such high correlations are found for other model comparisons using ozone at the outcome, one can apply the statistical model's standard errors with a "fudge factor" to the output from the computer simulations. That "fudge factor" would inflate the standard errors when they are applied to the simulation output.[20] In short, it appears that the computer simulation may be a good tool for studying the impact on ozone of various interventions.

4.4 Results for other Air Quality Constituents

There is not space here to review similar exercises undertaken for three other air quality constituents: carbon monoxide, nitrogen dioxide, and nitric oxide. Suffice it to say that in all three cases, both the statistical model and the computer model did not reproduce the observed data nearly as well as was done for ozone. Moreover, in each case, the statistical model suggested that the computer simulation may not be exploiting all of the relevant information available. It may be that in particular, temporal and spatial stability and/or positive feedbacks in air pollution production could be represented more effectively. Thus, one would proceed with virtual experiments for these air quality constituents at considerable risk.

5 Conclusions

What have we learned about the validity of virtual experiments that might be undertaken with the SMOG model? First, validity appears to depend on the particular air quality constituent under examination. The simulations are far more credible for ozone production than for the other three air quality constituents. Second, we have learned that validity seems to be better after the early morning hours and for sites not in the north eastern portion of the

[20] This approach was suggested by our UCLA colleague Wing Wong.

LA area. Third, insofar as reducing ambient temperature is an intermediate step in changing ozone concentrations, the statistical model seem to have the impact of variation in temperature on variation in ozone well represented. Fourth, it is apparent that because of the relatively long residence times of ozone, the effect of changing ozone concentrations earlier in the day have cumulative effects later in the day. Thus, small reductions in the morning hours may well translate into much larger reductions in the afternoon. The bang for this buck earns compound interest. Finally, we are on the way to possibly being able to put the rough equivalent of standard errors on outputs from the computer simulation. But everything will depend on continuing to get very high correlations between the output of the statistical model and output from the computer simulation.

More generally, it should be apparent that considering the validity of virtual experiments raises the same broad issues as considering the validity of real experiments. But of necessity, the tools brought to bear will be different. Work on the validity of virtual experiments is many years behind the work on validity in real experiments. We are perhaps at a point with virtual experiments where Donald Campbell and his colleagues were with real experiments over a generation ago. I am optimistic, however, that the increasing use of computer simulations for so many different scientific and policy applications will lead to very rapid growth in the sophistication with which virtual experiments are assessed.

6 References

Barron, E.J., 1995. "Climate Models: How Reliable are their Predictions?" *Consequences: The Nature and Implications of Environmental Change*(1),3: 17-27).

Berk, R.A., R.E. Weiss, and B. Western. 1995. "Statistical Inference for Apparent Populations" (with discussion). *Sociological Methodology, 1995.* P. Marsden (ed.). New York: Blackwell.

Bonan, G.B. 1997. "Effects of Land Ise on the Climate of the United States."*Climate Change* (37), 3: 449-486.

Briggs, J., and F.D. Peat. 1989. *Turbulent Mirror.* New York: Harper and Row.

Cook, R.D., and S. Weisberg. 1994. *An Introduction to Regression Graphics.* New York: John Wiley.

Gifi, A. 1990. *Non-Linear Multivariate Analysis.* New York: John Wiley.

Li, K.C. 1991. "Sliced Inverse Regression for Dimension Reduction (with discussion)." *Journal of the American Statistical Association* (86): 316-342.

Lu, R., R. P. Turco, and M. Z. Jacobson. 1997a. "An Integrated Air Pollution Modeling System for Urban Regional Scales: Part I: Structure and Performance." *Journal of Geophysical Research* 102: 6063-6079.

Lu, R., R. P. Turco, and M. Z. Jacobson. 1997b. "An Integrated Air Pollution Modeling System for Urban Regional Scales: Part II: Simulations for SCAQS 1987." *Journal of Geophysical Research* 102: 6081-6098.

Rosenbaum, P.R. 1995. *Observational Studies.* New York: Springer-Verlag.

Rossi, P.H., and H. E. Freeman. 1993. *Evaluation: A Systematic Approach*, (fifth edition). Newebury Park: Sage Publications.

Syu, H.-H, and J.D. Neelin. 1998. "ENSO in a Hybrid Coupled Model:Sensitivity to Physical Parameterizations and Predictions with Piggyback Data Assimilation." *Climate Dynamics*, submitted.

Van der Burg, E., and J. De Leeuw. 1983. "Non-Linear Canonical Correlation." *British Journal of Mathematical and Statistical Psychology*(36): 54-80.

Young, F. W., J. De Leeuw, and Y. Takane. 1976. "Regression with Qualitative and Quantitive Variables: An Alternating Least Squares Approach with Optimal Scaling features." *Psychometrica*(41): 505-529.