Invited Contribution to
Decision Modelling and Information Systems
by Nikitas-Spiros Koutsoukis and Gautam Mitra

Prof. Arthur M. Geoffrion
James A. Collins Professor of Management
Decisions, Operations, and Technology Management
John E. Anderson Graduate School of Management
Box 951481, UCLA
Los Angeles, CA 90095-1481
(310) 825-1113 (office)
(310) 825-1581 (fax)
arthur.geoffrion@anderson.ucla.edu

Restoring Transparency to Computational Solutions

Abstract Computational tools in support of decision making have grown greatly in power during recent decades owing in large measure to Moore's Law. Nothing like this law operates in the realm of analytical thinking, which has led to an increasingly lopsided emphasis on one at the expense of the other among decision support system developers. The increased emphasis on computational tools is a mixed blessing, for these seldom excel at revealing why the solutions they yield are what they are. Yet in many situations, decision makers and policy makers need to understand the *why* behind these solutions in order to convince themselves and others of the need for action or to deepen their own understanding of the system under study. This chapter advocates and illustrates an approach to using conceptually simple models, arguments, exhibits and spreadsheets as adjuncts to complex computational models to help explain important aggregate properties of detailed computational solutions. This can improve the transparency, and hence the value, of such solutions.

Model-based analytical thinking in support of decision making first occurred on a significant scale in the late 1930s, about five years before the first digital computers. Computers were quickly recognized as a valuable tool for embodying and solving decision models, and contributed greatly to the post-war flowering of operations research and decision support systems (not called that at the time). One might say that the fields of decision modeling and computers were advancing with comparable speed until the early 1960s, when the first integrated circuits were produced. At that point, Moore's Law led to exponentially rapid advances in computers and communications that continue to this day. Decision modeling, on the other hand – being essentially intellectual in character – continued to accumulate its capital in the form of articles, books, and experienced know-how at what must surely be a low-order polynomial rate. This

difference in developmental rate has profoundly impacted the practice of decision modeling by markedly changing the most advantageous mix of decision support activities away from analytical thinking toward greater use of computer-based decision tools.

Thus, decision support is increasingly about applying decision analysis software, neural networks, OLAP, optimization software, simulation, spreadsheets and their add-ins, statistical analysis software, and other computer-based tools.

This ascendance of computational approaches is perfectly understandable in light of their increasing comparative advantage, but something tends to be lost as a result: decision support practitioners may become less able to explain the *why* of their results as their studies become more computational. This is because computational tools typically tell you *what* but not *why*. Why does a neural network flag some credit card transactions as likely to be fraudulent but not flag others? Why does a large mixed integer linear programming model yield a particular configuration of facility locations and transportation flows? Why does a service system simulation result in a particular average utilization for a certain category of equipment? The answers must be in the data, the model, and the computational method's structure, but it takes extra work to find those answers. Such work may be undertaken too seldom amid the distractions and pressures of today's business environment.

Sometimes "knowing why" matters. If a credit card transaction is flagged as possibly fraudulent and investigative action is to be undertaken, then the more transparent the reasons for flagging the easier it will be to conduct the investigation and the more readily will the credit card holder forgive an intrusive inquiry. If a supply chain needs to be reconfigured to reduce cost and delivery cycle times, then the more transparent the reasons for recommended changes the easier it will be for the top logistics executive to understand them and to convince others of the need for change. If adding a certain piece of new equipment to a service system is the best way to improve service levels, those who must authorize the new expenditure may ask why one kind of equipment rather than another ought to be added.

Unfortunately, few computational methods offer much transparency. They tend to be fully occupied with their main task of answering "what" questions like what credit card transactions are likely to be fraudulent, what the best configuration is for a supply chain, and what the average machine utilizations are in a service system under given conditions.

The main point of this chapter is to advocate using conceptually simple models, arguments, exhibits, and spreadsheets in tandem with full-scale computational models as a way to better understand why detailed results are what they are. Conceptual simplicity is important because "why" explanations must be made to decision and policy makers who have little time or appetite for complexity. Models are fine so long as they can be explained easily and quickly, preferably with graphical aids. Informal arguments involving elementary reasoning, common sense, and contextual understanding are also fine. They must avoid higher mathematics (inappropriate to the intended audience) but ideally they should be based on mathematically sound arguments.

Straightforward spreadsheets are also acceptable now that this technology is ubiquitous in business. They have many possible uses, including approximating or verifying quantities that properly require complex calculations, and exhibiting idealizations of system behavior otherwise buried inside a detailed computational model.

Such means are, of course, limited in what they can accomplish by comparison with means of unlimited conceptual complexity and scale of computation. Perhaps the main consequence is that one can only aspire to explain the results of full-scale computational models at a relatively aggregate level of detail. Generally one chooses selected aspects and important aggregate properties of detailed computational results and seeks to explain those. If these aspects and properties are well chosen, explanations transparently accessible to a non-technical audience will be possible and of considerable value to them. Fortunately, highly aggregate properties tend to be more interesting to decision and policy makers than very detailed properties because they give a "big picture" within which details can be assimilated as necessary, and they tend to be easier to understand using simple methods.

Our focus on conceptually simple models, arguments, exhibits and spreadsheets is not meant to deny that carefully designed suites of full-scale solver runs (scenario studies, parametric and sensitivity analysis, etc.) can help illuminate the "why" behind the "what" of the detailed computational results. That approach can be quite effective. Rather, the aim here is to highlight an oft-neglected but illuminating role for the small and simple as adjuncts to complex models and computations. This should lead to greater acceptance of detailed results, if not to their justified rejection, and hence to greater willingness to act on them. Either way, an organization's value from decision support systems would increase.

This approach is not well suited to every decision support application. We primarily have in mind applications where persons not expert in decision technology play a role in deciding what to do with computational "solutions". It is here that the insights offered by simple methods can be most helpful. Real-time and on-line applications that help mechanize business processes, where there is little or no human review of individual cases, offer little scope for even the best such methods to help after such applications are installed, although transparency still can be helpful during development.

Two examples from very different contextual and methodological domains illustrate the recommended approach: the first concerns facility location for a distribution system, and the second concerns choosing the number of patrol cars to assign to a police precinct. A conclusion summarizes and reflects on how these examples may help guide similar efforts for other kinds of applications.

1. FIRST EXAMPLE: DISTRIBUTION CENTER NETWORK DESIGN

Nearly all manufacturers with distribution centers (DCs) that are not co-located at a plant face a DC network design problem. For a given demand-cost-capacity scenario, in one of the simplest versions of this problem they typically want to know:

- How many DCs should there be?
- Where should these be located?
- What should the product flows be between plants and DCs?
- Which customers should get which products from which DCs?

Answering these questions requires the ability to configure an entire distribution system so as to meet all demand subject to all applicable constraints at minimum total annual cost.

Large-scale optimization software capable of computing answers to such questions for practical problems of almost any size has existed for more than two decades (Geoffrion and Powers 1995) and is widely used, but it does not directly yield explanations for why its answers are what they are. Such explanations require either a suitable series of optimization runs or an analysis of the type discussed below, or both.

It is best to begin by first seeking explanations of managerially interesting, very highly aggregate properties of an optimal solution, turning afterward to less aggregate properties as interest warrants. For example, the optimal number of DCs is almost always of great interest to senior logistics managers. Which particular candidate DCs should be open usually is of interest only after a reasonable level of comfort has been achieved concerning whether the current number of DCs is too high or low.

How, then, can one use conceptually simple means to devise transparently understandable explanations for senior logistics managers for important, highly aggregate properties of a detailed optimal solution? We focus on the optimal number of DCs, in many applications the most important aggregate property of all. More precisely, we focus on the shape of optimal total cost as a function of the number of open DCs in a detailed model, exclusive of cost components that do not depend significantly on the number of open DCs (those components do not impact how many DCs should be open). This optimal total cost function presumes that the detailed model makes the best possible use of each given number of DCs, which therefore requires performing a suite of optimization runs each with an equality constraint on the total number of open DCs.

The ability to predict and explain the relevant part of the optimal total cost function is much more useful than just predicting and explaining the total number of open DCs, since the behavior of optimal total cost for a varying number of open DCs is valuable for a variety of purposes as explained in section A of (Geoffrion 1979). Such a prediction is also far more convincing, since our audience might attribute success in predicting a single

number to simple good luck, whereas success in predicting a whole function is much less easily dismissed.

We now develop a very simple analytic model that accomplishes this task for a particular optimal DC location study carried out some years ago.

Since all demand must be met in this kind of study, knowing the optimal number of DCs is equivalent to knowing the optimal average DC size, where *size* is measured in terms of annual DC throughput volume in hundredweight (CWT), say. Optimal DC size has been much studied analytically (e.g., Bos 1965, Rutten 2001), and such analyses provide the foundation for constructing explanations of the type desired.

At bottom, we would tell our intended audience of senior logistics managers, what determines the optimal size of a DC is a single cost trade-off involving its annual fixed cost and the total annual cost of delivering to its customers. The goal is to find the size for which the sum of these two costs is minimal on a unit ($/CWT) basis, where fixed cost is allocated equally to each throughput unit. See *Figure 1*. If a DC is too small, then the unit delivery cost will be small because just nearby customers will be served, but the unit fixed cost will be excessive because there isn't enough volume to spread it over. If a DC is too large, then the unit fixed cost will be small but the unit delivery cost will be excessive because many customers will be far from the DC. A DC will be just the right size when these two effects exactly balance.

Among the simplest analytic models that captures this trade-off is one that assumes demand to be uniformly distributed on the plane, delivery cost to be strictly proportional to Euclidean distance and amount delivered, fixed cost to be independent of size, and the shape of the DC's service area to be circular with the DC at the center:

A	Circular DC service area in mi^2
ρ	Density of demand in CWT/mi^2
f	Fixed cost of a DC in $
t	Delivery freight rate in $/CWT-mi.

Under these assumptions, the unit fixed cost as a function of A is

(1) $f/\rho A$ $/CWT

and the unit delivery cost is

(2) $\frac{2}{3}\,(A/\pi)^{1/2}\,t$ $/CWT.

Figure 1 is a graph of these functions and their sum using values for ρ, f, and t from an actual consumer products manufacturer (Geoffrion 1976). The minimum over A of (1) plus (2) is at

(3) $A^* = (9\pi)^{1/3}\,(f/\rho t)^{2/3}$ mi^2.

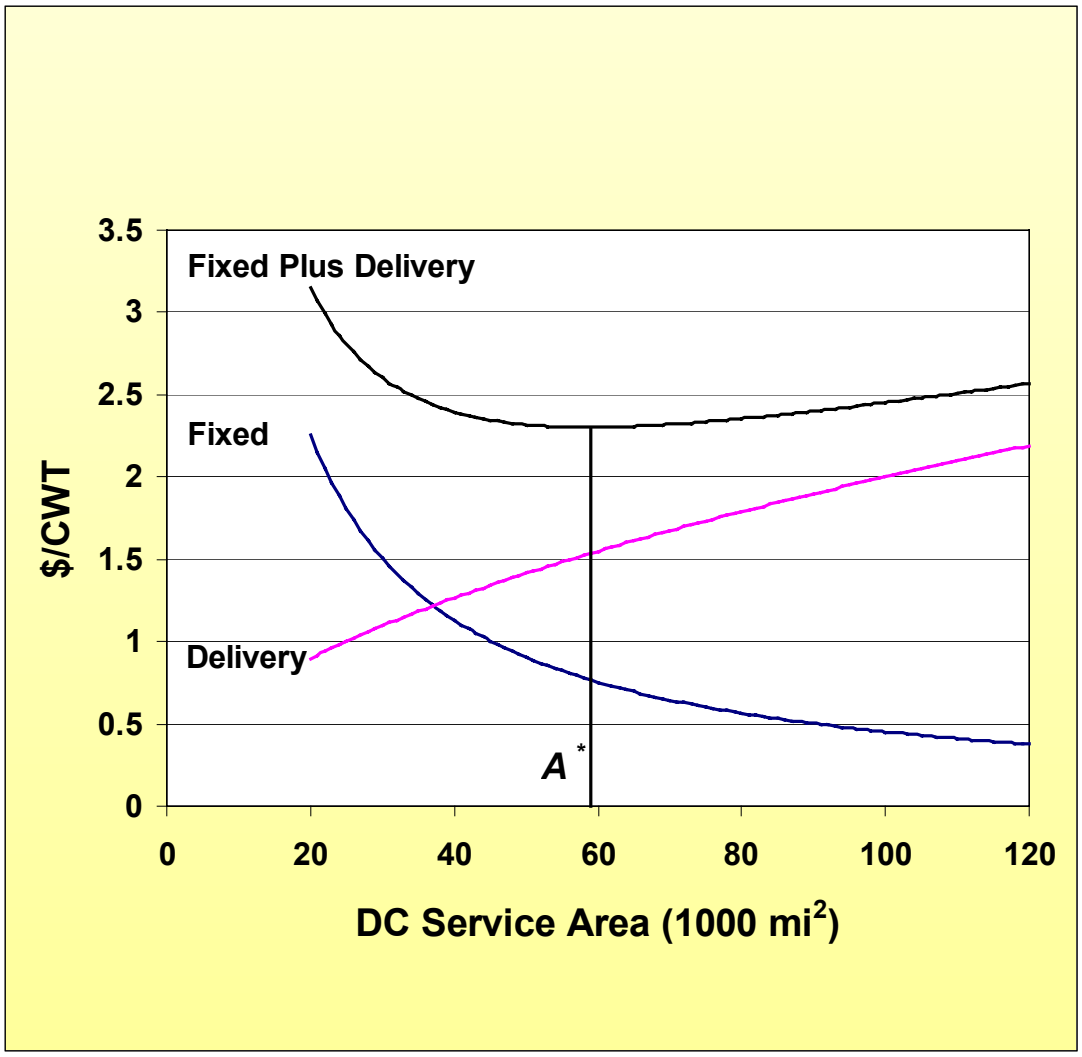


Figure 1. **Unit DC Fixed and Delivery Cost Trade-off Determines the Optimal Service Area (data for a consumer products manufacturer)**

The fixed cost function (1) is transparent, since ρA is just the total demand in a service area of A mi^2, but the delivery cost function (2) contains a mysterious factor and square root that are not transparent without further explanation. Taking the integral that proves (2) or the derivative that proves (3) would be inconsistent with the goal of managerial transparency.

How to explain where (2) and (3) come from? First consider (2). Clearly, unit delivery cost in \$/CWT must equal t times the average delivery distance traveled in miles. This distance must be some fraction of the radius r of a circle whose area is A; since $A = \pi r^2$, as almost every manager knows, r must solve that simple equation and so $r = (A/\pi)^{1/2}$. Thus the average delivery distance traveled must be some fraction of $(A/\pi)^{1/2}$, which explains the mysterious square root in (2).

Formula (2) says that this fraction is ⅔. This number has face validity – obviously the average distance traveled is a bit more than ½ r – and the justification of (2) using

only elementary arguments probably should stop here. If necessary, however, a simple argument for the value ⅔ could be provided.

The simplest argument takes the form of a simple physical demonstration. By symmetry, it suffices to examine any very narrow "pizza slice" taken from the service area. If one makes a lamina in the shape of an isosceles triangle with a very narrow vertex angle and then balances it on a knife edge, the lamina will balance at a point ⅔ of the way from its vertex to its base because that is the location of its center of mass about the median falling from the vertex. (Average delivery distance from the vertex and this coordinate of the center of mass are mathematically identical.) Another derivation of the value ⅔ relies on brute numerical integration implemented in a spreadsheet, which can be set up to be much more transparent than the corresponding manipulation in calculus. The easiest design divides the circular service area into many narrow annuli of equal width. Nothing more than the formula for the circumference of a circle is needed to set up a spreadsheet whose estimate of average delivery distance converges to ⅔ as the width of the annuli approaches 0.

Numerical evaluation of (3), although not the formula itself, is easy once (2) is accepted because it is obvious from the total unit cost plot of *Figure 1* what the optimal service area is and why: it is the point at which the incremental changes in the two costs being traded off exactly cancel as the DC service area changes. There is no need to take a derivative of (1) plus (2) and set it equal to 0.

Determining A^* from a graph like *Figure 1* leads to the optimal number of DCs, for if n is the number of DCs and A^+ is the total demand area to be served in square miles, then obviously minimizing total cost requires all DCs to have area A^* and the analytic model predicts

(4) $\quad n^* = A^+/A^*.$

As discussed earlier, it is not so much the ability to transparently predict the number of DCs open in an optimal solution of a detailed, large-scale optimization model that is of interest to decision makers as is the ability to predict the optimal total cost function produced by parametrically varying the number of DCs that must be open. Call this function $TC^*(n)$, where cost components that do not depend significantly on the number of open DCs are excluded. We focus on the *shape* of this function, best captured by normalizing both its domain and range scales. We normalize in two steps for the analytic model: first for the n scale, then for the total-cost scale.

To write down $TC^*(n)$ for the analytic model, one must be able to specify the service areas of all n DCs so that they sum to A^+ and minimize total cost. Lagrange multipliers can easily be used to show that all service areas must be the same size in an optimal solution, but this would destroy managerial transparency. Instead, this result can be seen easily from the shape (convexity) of (1) plus (2) in *Figure 1*, the essential observation being this: for any two points on this curve, the average of their values is greater than the value for the average of the two service areas. Since all pairs of DCs need to have the

same-sized service areas to minimize their joint costs, all n DCs need to have the same-sized service areas to minimize total costs. Thus A must equal A^+/n in (1) and (2) in order for total costs to be minimized, and multiplying the sum of these unit costs by total demand (ρA^+) yields

$$(5) \qquad TC^*(n) = (fn/\rho A^+)(\rho A^+) + \tfrac{2}{3}(A^+/n\pi)^{1/2} t (\rho A^+).$$

To normalize the n scale around n^*, write $TC^*(n)$ in the form $TC^*((n/n^*) n^*)$ and collect terms in (n/n^*). After some simple algebra and using (4) to rewrite n^* and (3) to rewrite A^*, one obtains

$$(6) \qquad TC^*(n) = [\tfrac{1}{3}(n/n^*) + \tfrac{2}{3}(n/n^*)^{-1/2}] (fA^+/(9\pi)^{1/3} (f/\rho t)^{2/3}).$$

Normalizing the cost scale around $TC^*(n^*)$ yields

$$(7) \qquad \frac{TC^*(n)}{TC^*(n^*)} = \tfrac{1}{3}(n/n^*) + \tfrac{2}{3}(n/n^*)^{-1/2}$$

which, surprisingly, does not depend on any of the problem data parameters (ρ, f, t, A^+). It implies, for example, that if $n/n^* = 1.4$ in the analytic model, then total fixed plus delivery costs exceed the minimum attainable value by 3% no matter what the problem data are. Formula (7) springs entirely from the geometry of the plane.

Figure 2 graphs (7). It also plots optimal values produced by a DC location study for a century-old mining company. That study had 23 products produced at 12 plants, 51 candidate public DC locations, and 110 customer zones. 7 optimization runs using Benders decomposition, each constraining the number of open DCs to a specific number, were executed to obtain the 7 plotted points.

The agreement evident in *Figure 2* is quite satisfying, and shows that even such a simple analytic model can produce a good prediction for the behavior of optimal distribution cost from a real large-scale optimization model.

In other real-life DC location studies, (7) may not agree so well with the results of mixed integer linear programming optimizations. Then it will be necessary to refine the analytic model in quest of an improved normalized formula that agrees more closely. There are many ways to undertake this. One would be to consider DC service areas other than circular, e.g., square or hexagonal. But from results in (Bos 1965), it is easy to show that this has a negligible effect on the unit delivery cost formula (2) and on all subsequent results. Another refinement would be to take into account the effect of DC replenishment costs modeled, like delivery costs, as constant per CWT-mile. This is done in the appendix of (Geoffrion 1979) where, surprisingly, it is shown that – to good approximation for realistic data values – (7) still holds without change in the case of a single plant. Additional possible refinements include delivery freight rates that diminish on a per-mile basis with increasing distance from the DC and other tractable variants studied in (Rutten et al. 2001). Some of these lead to changes in (7).

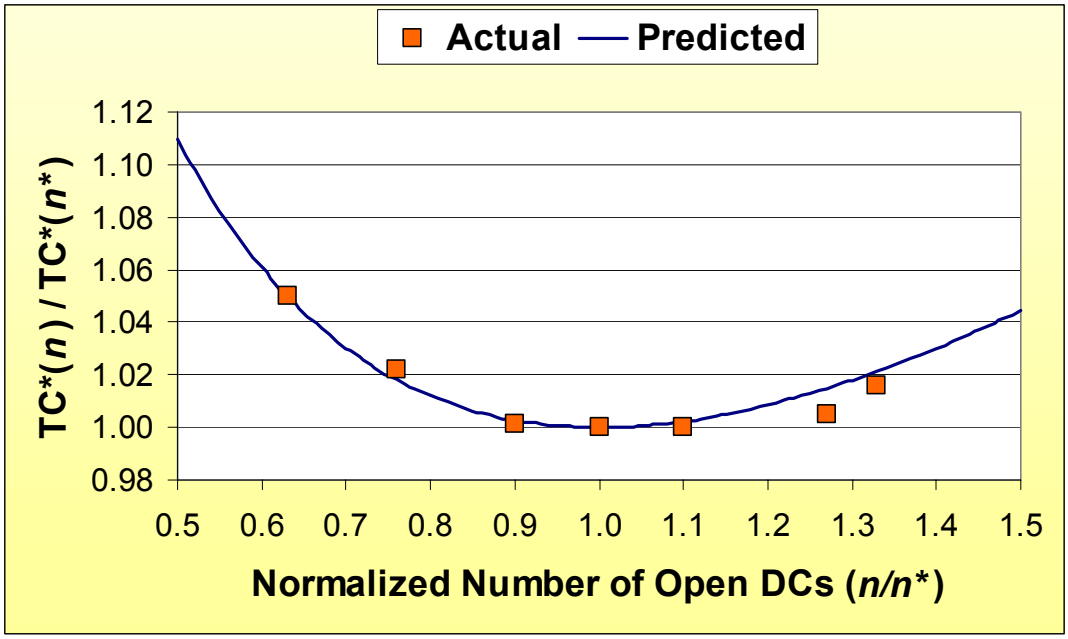


***Figure 2*. Predicted Minimum Total DC Fixed Plus Delivery Cost as a Function of the Number of Open DCs, Doubly Normalized; Actual Minimum Values for a Mining Company.**

To summarize, *Figure 2* shows that we have a good understanding of why the detailed optimization results are what they are, at the level of the shape of the optimal total cost function associated with parametrically varying the number of DCs that must be open. This shape derives from the cost trade-off illustrated in *Figure 1*. We took pains to derive (7) in a way that is managerially transparent – successfully replacing both differential and integral calculus operations and elementary optimization theory by simple arguments – with just one exception unmentioned until now: we did not give a transparent justification for the formula (3) needed to derive (6) and hence (7).

It turns out that, although it is easy to evaluate A^* from a graph like *Figure 1* for any particular values for f, ρ, and t, a sufficiently transparent derivation of the closed form (3) is elusive. Fortunately, there is a workaround, and a very insightful one at that. There is a managerially transparent way to graph (7) for any values for f, ρ, and t. (Of course, we know from (7) that these values do not really matter.) The key observation is that

$$\mathrm{NTC}^*(n/n^*) \equiv \mathrm{TC}^*(n)/\mathrm{TC}^*(n^*) \qquad \text{(NTC is mnemonic for normalized total cost)}$$

and

$$\mathrm{NUC}(A/A^*) \equiv \mathrm{UC}(A)/\mathrm{UC}(A^*) \qquad \text{(NUC is mnemonic for normalized unit cost)}$$

are very close relatives, where UC(A) is defined to be (1) + (2). Namely,

(8) $NTC^*(n/n^*) = NUC(n^*/n)$.

This relation follows easily from

(9a) $TC^*(n) = n\ UC(A)\ \rho A$ where $A = A^+/n$

and

(9b) $TC^*(n^*) = n^*\ UC(A^*)\ \rho A^*$ where $A^* = A^+/n^*$,

both of which are managerially transparent from observations made earlier. Using (8), it is a simple matter to graph NTC* once the easily-graphed function NUC is drawn. Now the successful prediction shown in *Figure 2* can be explained in a totally managerially transparent way.

2. SECOND EXAMPLE: ALLOCATING POLICE PATROL CARS TO PRECINCTS

The New York City-Rand Institute did extensive consulting work for the New York City Fire and Police Departments during the 1970s. One project centered on how many patrol cars to allocate to a police precinct to service calls requiring patrol car dispatch (Ignall et al. 1978). Both a simplified analytic queuing model and a detailed queuing simulation program were developed with the aim of investigating whether the former could be used for decision-making purposes rather than the latter, which was more cumbersome. Note that these roles for simple and detailed models, while appropriate to this application, are very different from what this chapter advocates: using a simple model to *explain* a complex model's results in an elementary way, not to *replace* it.

The basic situation is that incoming calls arrive randomly at a police precinct, where they are prioritized into five categories and assigned to patrol cars for servicing. There are N patrol cars, which amount to mobile servers, with random service times. (Ignall et al. 1978) identifies the most important congestion measures to be P_w, the long-term chance that an arriving call finds all N cars busy and therefore has to wait before dispatch, and W_q, the long-term average time that an arriving call spends in the queue awaiting a patrol car to become available for dispatch.

The consultants chose the M/M/N priority queuing model (Cobham 1954) even though its assumptions were known to be incorrect in several respects. For example, the arrival process was not stationary, service times were not exponential, it takes time to travel to the scene of a call, cars were not always in service, and multiple cars were dispatched for some calls. They used computerized records maintained by NYPD to estimate the model's parameters and to specify corresponding runs of a custom discrete event simulation program that aimed for far greater realism than the analytic model permitted. That program – 700 lines of SIMSCRIPT II.5 not counting the geometric specifications of the precinct under study or the historical "job stream" of calls that drives

the simulation – built in the five complexities just mentioned, among others, and was duly validated against historical data.

Upon comparing results as N varied over its relevant range, they found that the queuing model gave estimates of P_w and W_q quite good enough by comparison with the detailed simulation estimates to convince NYPD that the queuing model could be used for decision-making purposes in place of the detailed simulation. Some of these results are reproduced in *Figures 3A* and *3B*.

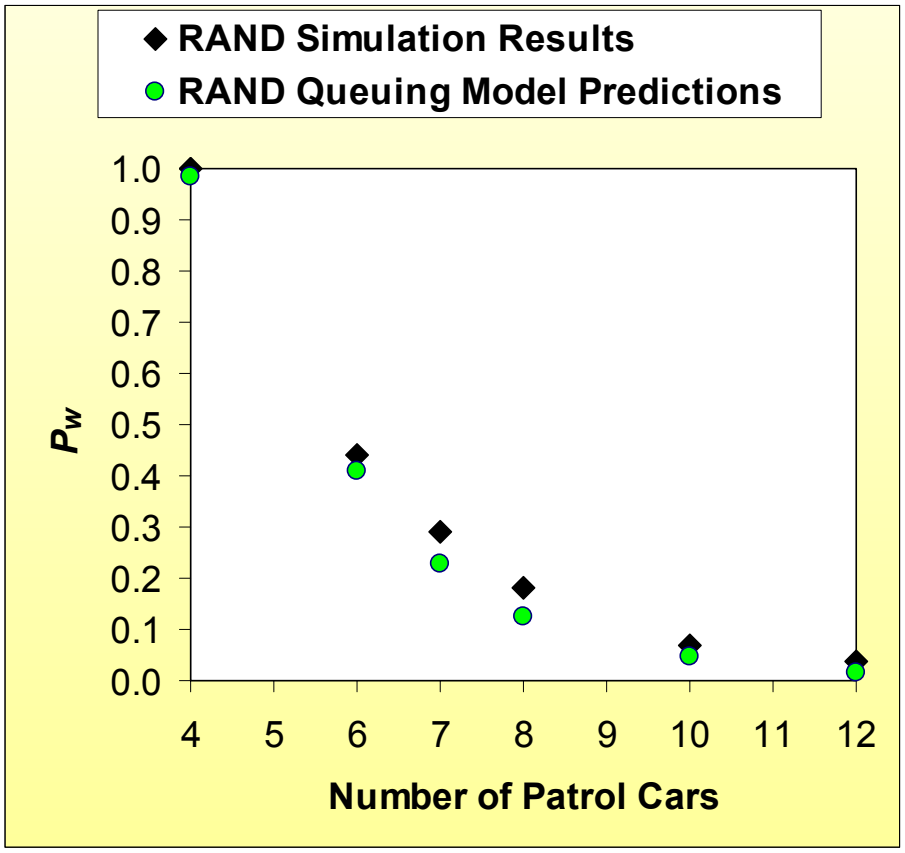


***Figure 3A*. Unavailability of Patrol Cars versus Number of Cars Assigned to Precinct**
Recreated from Figure 1 of (Ignall et al., 1978) (4 PM to Midnight Tour)

The main discrepancy, under-estimation of congestion, was anticipated because the simulation model permitted multi-car dispatch.

Plots like these, which show the trade-off between number of cars and surrogate measures for public satisfaction, marked the end of the formal analysis. It was up to NYPD officials to decide where to operate on these trade-off curves, that is, to decide how many patrol cars to allocate to a given precinct.

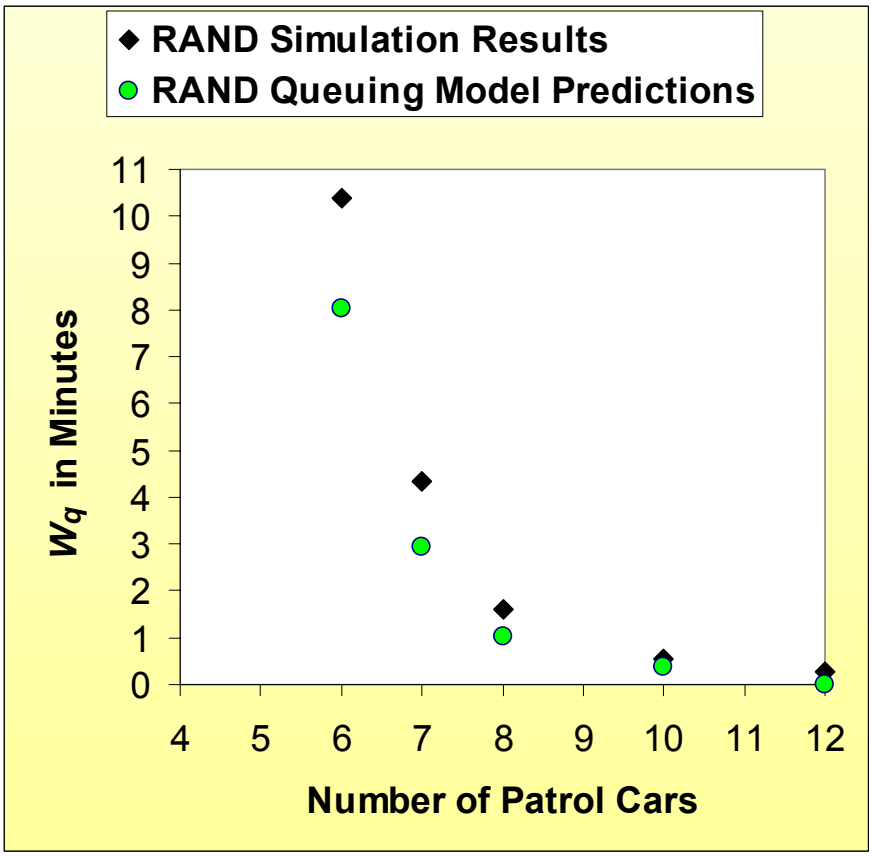


***Figure 3B*. Average Delay in Queue versus Number of Patrol Cars Assigned to Precinct**
Recreated from Figure 2 of (Ignall et al., 1978) (4 PM to Midnight Tour)

Having established that a highly simplified analytic queuing model captures the essence of the detailed patrol car dispatch simulator (perhaps after empirical adjustment for the mentioned cause of under-estimation), we turn now to the central question: Can this simple model be used to construct insightful explanations for mathematically naïve decision makers of why P_w and W_q vary with N as they do? The regularity evident in *Figure 3* begs for transparent explanation.

The M/M/N priority model is well understood mathematically. Unfortunately, this understanding is not accessible to our intended audience. Yet it is possible to answer the central question in a way that appeals to intuition rather than higher mathematics.

Below we assume a first-come-first-served queuing discipline rather than a priority discipline where calls fall into priority classes that determine which waiting call is served next, because there is no indication in (Ignall et al., 1978) or in the supporting RAND reports that service times differ for different priority classes. When service time distributions are the same for all classes, the M/M/N priority model reduces to standard M/M/N so far as P_w and W_q are concerned. A tiny amount of standard queuing notation will be useful, which should be bowdlerized when speaking to a non-technical audience:

λ for the call arrival rate and μ for the service rate of each patrol car server. Let the unit of time be one minute.

It would be nice if there were an intuitively accessible derivation of the general formula for P_w for the M/M/N model, both as an end in itself and because from it one can pass with intuitive ease to a general formula for W_q (see below). Then everything needed to explain the theoretical curves for P_w and W_q in *Figure 3* would be at hand. Unfortunately, we know of no such derivation. One can, however, develop some useful insights and provide a simple tool for estimating P_w as accurately as desired.

A good starting point is this simple conservation law: in the long term, what goes into the system must equal what comes out (non-technical people will likely accept that the system does not "blow up", but if not then the spreadsheet simulation introduced later gives evidence that it does not). What goes in over a very long period of T minutes is close to λT calls demanding service. What comes out over the same period is close to

[long-term fraction of all servers that are busy]$N\mu$T

served calls. Setting these quantities equal to one another yields

> **Observation 1**: If M/M/N total service capacity is sufficient to keep up with demand for service over the long term ($N\mu > \lambda$), then the long-term fraction of all servers that are busy is $\lambda / N\mu$.

Next comes a nearly trivial but still fundamental fact about what is really going on in M/M/N and most other queuing systems:

> **Observation 2**: If $N\mu > \lambda$ for M/M/N, then the queuing congestion that arrivals experience can be viewed as occurring because demands for service are not coordinated with the availability of service capacity.

To see this, notice from Observation 1 that not all servers will be busy all the time. Each server becomes idle infinitely often over an infinite time span because otherwise the conclusion of Observation 1 could not hold. In fact, by randomness, *all* servers will simultaneously become idle infinitely often over an infinite time span. Consider any time segment that begins and ends with times at which all servers are idle. Clearly all demand for service in such a segment, as measured by the total actual service time, could have been met without any delay whatever if the arrival times could have been adjusted to coincide with a server being free, or if the servers could have adjusted their service times to assure that no arrival had to wait for service, or by some combination of these two kinds of adjustments. For example, the partial sequence of arrival times {… 112.4, 115.5, 123.0, …} might need adjustment to {… 112.4, 117.7, 123.0, …} in order to preclude waiting by the second of these arrivals.

It is important to notice that such adjustments would not require changing either the total number of arrivals or the total time that the servers work during the considered time

segment. In other words, it is control over the individual inter-arrival times and/or of service times that is needed for perfect coordination, rather than control over their *averages*. The averages can stay the same. Total control of either would suffice, as would enough partial control over one or both. Notice that this control does not necessarily mean diminishing variability in one of the usual statistical senses; sometimes an inter-arrival or service time needs to depart more rather than less from the mean of its statistical distribution to reduce queuing congestion.

Thus

> **Observation 2A**: If $N\mu > \lambda$ for M/M/N, then perfect coordination of demands for service with the availability of service capacity would render it unnecessary for an arrival ever to encounter all servers busy or to wait in a queue. "Perfect coordination" means exercising sufficient control over the inter-arrival times and/or of the service times to avoid all service conflicts.
> *NB*: Coordination of any type would mean that either the arrival process or the service process or both would no longer obey the original assumptions.

As noted, the kinds of adjustments needed are not permitted in the classical M/M/N queuing system and certainly are difficult to imagine in the context of patrol car dispatching. But such adjustments are not at all difficult to imagine in other contexts, especially in the future. A momentary digression on this topic will help establish the feasibility and potential importance of service supply-demand coordination for at least some queuing systems.

Imagine a non-emergency medical clinic serving a population of people with wireless messaging devices. A patient who decides to visit the clinic could message it to obtain a number giving his place in a virtual line awaiting service. The clinic would keep the patient advised of when a doctor would likely be available, and the patient could continue going productively about his activities and coordinate his movements so that he arrives at the clinic very close to the time when a doctor becomes available to see him. There would be essentially no *physical waiting*, although *virtual waiting* could still be in accord with classical waiting line theory. If the cost of virtual waiting is negligible by comparison with the cost of physical waiting (with its unproductive dead time), the patient would experience essentially no costly queuing congestion at all. That would be his reward for coordinating his demand for service with service availability.

Some amusement parks and restaurants in pedestrian districts are already using similar means to reduce physical waiting, and it is by now widely recognized that wireless communications, especially with location-aware devices, open up important new possibilities for reducing queuing congestion. For example, one of the predictions in (Penzias 1997) is called "The End of Lines – GPS Becomes Indispensable". The vision is that "Networked alternatives to congestion rationing will extend the just-in-time concept to consumer services" (p. 163). This is one of the ways in which the coordination mentioned in Observation 2A could occur.

Observation 2A shows that coordination could reduce P_w and W_q to essentially zero so long as $N\mu > \lambda$, but otherwise all bets are off. The conservation law behind Observation 1 shows that the backlog of unserved calls would build up without limit over time if $N\mu < \lambda$, so P_w would tend to 1 and W_q can be thought of as tending to infinity. What about the case $N\mu = \lambda$? It turns out that, with perfect coordination, P_w and W_q would still remain at 0, while with some types of imperfect coordination they could be positive but less than 1 and infinity respectively. With other types of imperfect coordination, they would rise to 1 and infinity. We shall not consider the case $N\mu \leq \lambda$ further.

Observation 3: If $N\mu > \lambda$ for M/M/N and there is no coordination, then $P_w \leq \lambda / N\mu$ and equality holds when $N = 1$.

Like Observation 1, this follows from conservation – in the long term, what goes into the system must equal what comes out – but requires a more detailed version of the conservation equation. What goes in over a very long period of T minutes is close to λT calls. What comes out over the same period is the sum of two parts: what comes out while all N servers are busy and what comes out while fewer than all N servers are busy. The first part is close to

(10) $F^b\mathrm{T}(N\mu)$ calls,

where F^b is the fraction of T during which all servers are busy (the superscript ^{b} is mnemonic for "busy", not an exponent), and the second part is close to

(11) $(1\text{-}F^b)\mathrm{T}(N^{nb}\mu)$ calls,

where N^{nb} is the average number of servers busy when not all servers are busy (^{nb} for "not all busy"). Notice that all T minutes are accounted for. Notice also that, by chance, P_w – the long-term fraction of arrivals coming at a time when all servers are busy – will nearly equal F^b for sufficiently large T. Setting the system input, λT, equal to the sum of these two output components yields the conservation equation

(12) $\lambda\mathrm{T} = P_w\mathrm{T}(N\mu) + (1\text{-}\ P_w)\mathrm{T}(N^{nb}\mu)$

or, solving for P_w,

(13) $P_w = (\lambda - N^{nb}\mu)/(N\mu - N^{nb}\mu)$.

It is clear from its definition that N^{nb} must equal 0 for the case $N = 1$. This gives the desired result

(14) $P_w = \lambda / N\mu$ for $N = 1$.

A formula for N^{nb} is well known for general N, but is messy and not transparently accessible to non-technical people. This development must, therefore, get along without it. The inequality of Observation 3 is easily obtained by dropping the (necessarily nonnegative) term involving N^{nb} in (12). This concludes the justification of Observation 3 using only elementary arguments.

Finally, we need a formula for W_q, the long-term average time that an arriving call spends in the queue waiting for a patrol car to become available for dispatch in the absence of any coordination. We now develop such a formula, again using only elementary arguments.

An arriving call will either find all patrol cars busy (the long-term frequency of this is P_w) or not. Call the long-term average waiting time $W_q^{\,b}$ (the superscript ^{b} is for "busy", not an exponent) in the first case and $W_q^{\,nb}$ (^{nb} for "not all busy") in the second. Clearly $W_q^{\,nb}$ is zero, since there cannot be an idle car unless there is no queue, so

(15) $W_q = P_w\, W_q^{\,b} + (1- P_w)\, 0.$

Consider now what has to take place during time $W_q^{\,b}$: $L_q^{\,b}+1$ services must occur, where $L_q^{\,b}$ is the long-term average number of calls in queue when all cars are busy. Why the extra service? Because after $L_q^{\,b}$ services, the arriving call will be at the head of the line but one more service must occur in order for there to be a car available for dispatch. These $L_q^{\,b}+1$ services will be accomplished by a 100%-busy fleet that serves at the rate $N\mu$ calls per minute and hence with an average service time of $1/N\mu$ minutes per call. Therefore,

(16) $W_q^{\,b} = (L_q^{\,b}+1)/N\mu.$

Since Little's Law (which can be explained easily to non-technical people) applies during the times when all patrol cars are busy, $L_q^{\,b}$ can be written $\lambda\, W_q^{\,b}$ in this expression. Solving the result

(17) $W_q^{\,b} = (\lambda\, W_q^{\,b}+1)/N\mu$

for $W_q^{\,b}$ yields $W_q^{\,b} = 1/(N\mu - \lambda)$. Using this in (15) yields the desired formula for W_q:

Observation 4: If $N\mu > \lambda$ for M/M/N and there is no coordination, then $W_q = P_w/(N\mu - \lambda)$.

Figures 4 and *5* show the price of failing to coordinate according to the results of Observations 2A, 3 and 4 applied to M/M/1. They plot P_w and W_q against the ratio of service rate to arrival rate because this ratio is what determines system performance. Notice that the value of cooperation increases as the service and arrival rates approach one another, because arrivals are then increasingly likely to come when the patrol car is

out on another call. For sufficiently large values of service rate relative to arrival rate, both P_w and W_q approach their perfect-coordination values of 0 because services are more spread out in time and arrivals less likely to come when the patrol car is busy.

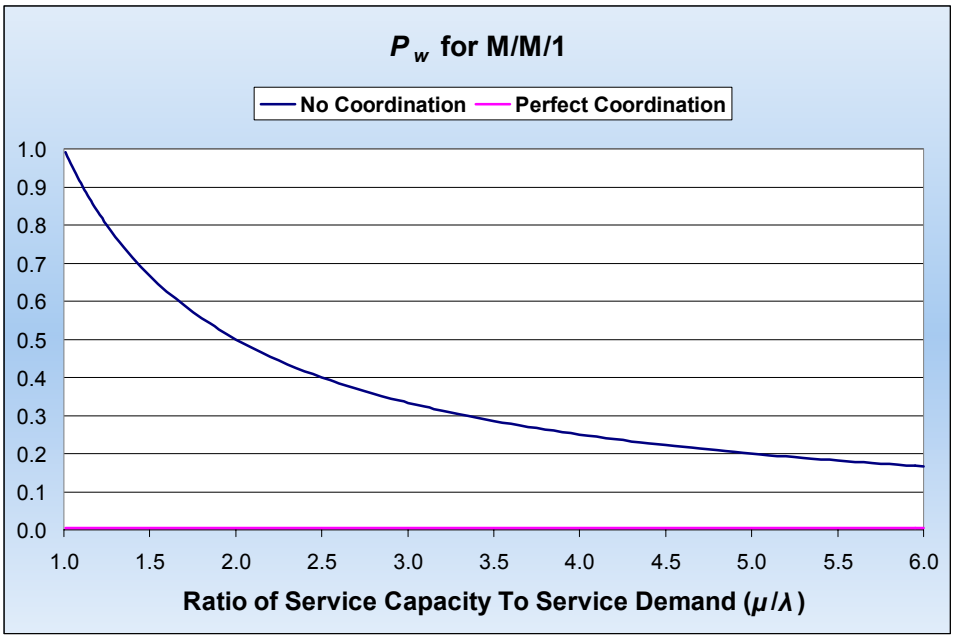


Figure 4. **P_w Can Be Reduced to Zero If Perfect Coordination Can Be Achieved Between Demands for Service and the Availability of Service Capacity**

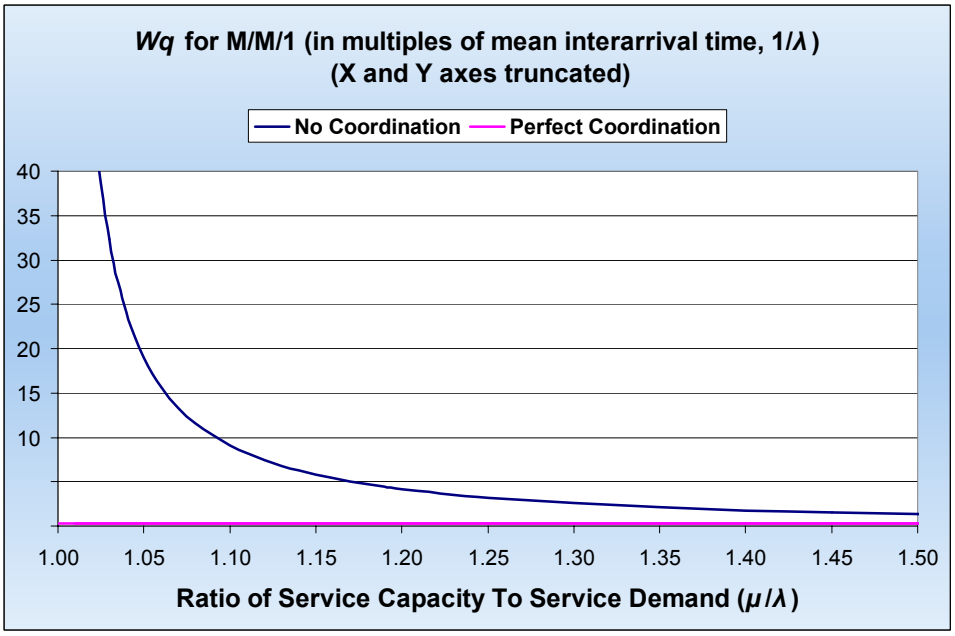


Figure 5. **W_q Can Be Reduced to Zero If Perfect Coordination Can Be Achieved Between Demands for Service and the Availability of Service Capacity**

It would be nice to exhibit similar graphs of P_w and W_q for general M/M/N, but Observation 3 supplies only a rather poor bound for P_w when $N > 1$ and hence neither P_w nor W_q can be graphed exactly without recourse to well known theoretical results inaccessible to a non-technical audience.

We shall resort to a numerical approach to estimate P_w, but first interject a useful managerial insight about what happens as N increases. As before, consider system performance as a function of the ratio of total service capacity to service demand. The appropriate ratio is now $N\mu/\lambda$, since N servers each able to service μ calls per minute give a total capacity to service $N\mu$ calls per minute if fully utilized. Notice that one can increase total service capacity by increasing N, or by increasing μ, or both. For given $N\mu$, would one prefer a larger N with a smaller μ or the reverse? The answer is that, if P_w and W_q are the only performance measures of interest, one prefers the first option.

The reason is that more (but proportionally slower) servers give lower values of P_w. Roughly, this is so because service capacity is broken into smaller chunks and thus there is a greater chance that at least one of these chunks will be free when a call arrives. In the extreme case of very many, very slow severs, an arriving call is almost certain to find a free server (i.e., P_w will be very small). The catch, of course, is that improved service availability is achieved at the cost of slower service times, but that is of no concern to P_w. The effect is quite strong, and is exhibited in *Figure 6* for several values of N. Lower values of P_w mean lower values of W_q also, by the formula in Observation 4.

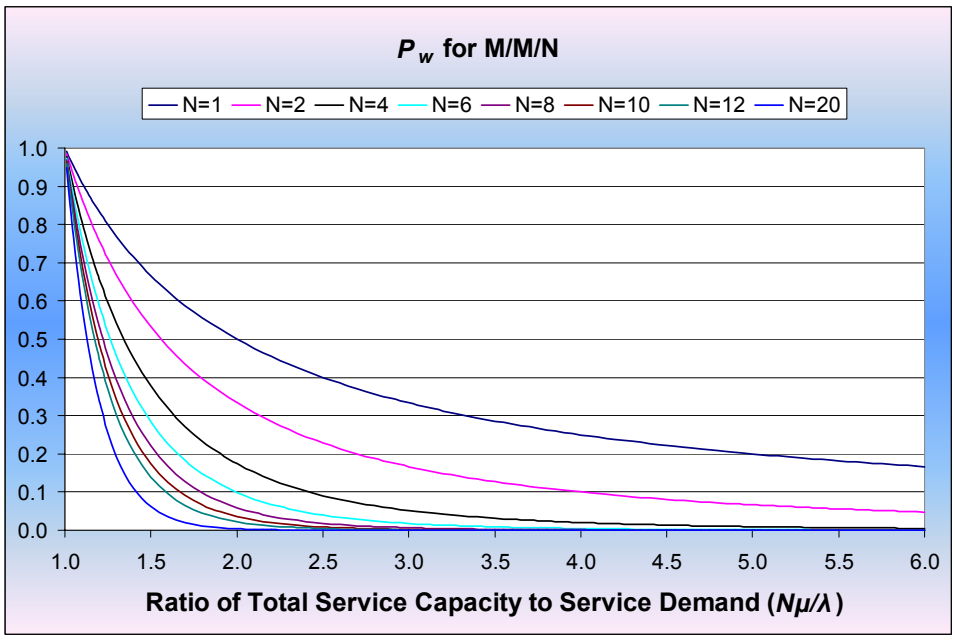


***Figure 6*. For a Given Total Service Capacity $N\mu$, Increasing N Reduces P_w (P_w values for $N > 1$ from a well-known theoretical formula)**

Note that *Figure 6* shows exactly how poor the bounds of Observation 3 are for $N > 1$: just look at the gap between the $N = 1$ curve (which has a secondary interpretation as graphing the bound $\lambda / N\mu$ for any N) and any other curve.

As mentioned earlier, the formulas behind this graph are not accessible to a non-technical audience for $N > 1$, so next we explain how to develop numerical estimates of P_w in an accessible manner. In this way, one can generate graphs like *Figure 6* for any values of N, μ, and λ .

In the absence of a transparent derivation of P_w for $N > 1$, one can estimate it using a conceptually simple spreadsheet that mimics the operation of the patrol-car dispatch system. Almost any of the many M/M/N spreadsheet simulators intended for educational purposes would suffice, possibly after cosmetic work to make it more suitable for viewing by police officials. Here is one possible design.

We aim at simplicity and the grossest level of detail that still yields an estimate of P_w that converges to the correct value if the simulation runs long enough. One reasonable choice is to simulate the birth-death process associated with the classical M/M/N queuing model, where the process starts in a known state and the simulation clock advances in small increments. As usual in this context, *state* means the number of calls in service or waiting for service at a given moment. Such a spreadsheet can capture and graph the complete state history, and decision makers can inspect this as a way to improve their intuition about the system's dynamic behavior as modeled. Notice that the usual steady-state equations are not addressed directly; P_w is estimated from simulated long-term behavior.

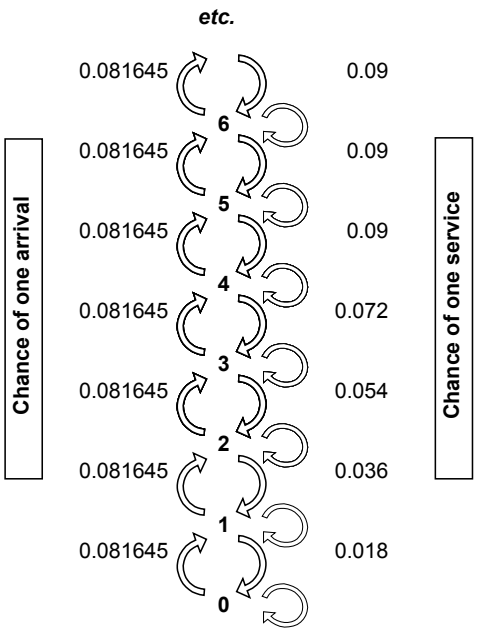
Figure 7 shows one way to set up such a simulation, which ignores inter-arrival and service time details and focuses only on state transitions. For sufficiently small time slices, the M/M/N assumptions imply that transitions to non-neighboring states occur negligibly often and the birth-death process for system state will look like a random walk with steps of length one. The diagram in this figure shows the rules for this walk: each arrow indicates a possible step, and the probability of each step up (in the event of an arrival) and step down (in the event of a service) is annotated on the diagram. Self-loops mean "no step" and have the obvious complementary probability for each state.

Using the data summarized at the top, from which the rest of the spreadsheet is calculated dynamically, the bottom part records the detailed history of the system for any desired number of time slices, in this case 10,000 (more can easily be added by a single copy operation). It is easy to explain how the spreadsheet works by narrating a few steps and following along in the state history log at the bottom. Finally, the desired estimate of P_w (.772 in the particular simulation shown) comes from tallying the fraction of time slices for which the state is 5 or higher.

Data Values

Lambda	**0.16329**	calls per minute	**Initial State of System**		**4**
Mu	**0.036**	calls per minute			
Slice Size	**0.5**	minutes	(10,000 slices =	83.33	hours)
N	**5**				

State Diagram



Estimate of P_w from Simulation **0.772**

Time	< ---------- History ---------- >		
Slice No.	**State**	**Srvrs Busy**	**No. Waiting**
1	**4**	4	0
2	4	4	0
3	3	3	0
4	3	3	0
5	3	3	0
rows omitted			
79	5	5	0
80	6	5	1
81	6	5	1
82	7	5	2
83	7	5	2
84	7	5	2
rows omitted			
7,768	24	5	19
7,769	24	5	19
7,770	25	5	20
7,771	25	5	20
7,772	25	5	20
7,773	26	5	21
rows omitted			
9,996	4	4	0
9,997	4	4	0
9,998	4	4	0
9,999	4	4	0
10,000	5	5	0

***Figure 7*. Spreadsheet for Simulating M/M/5 Queuing System**

Figure 8 graphs the state history for all 83⅓ hours of the simulation run of *Figure 7*. Managers should be interested to see the periodic excursions above and in the no-waiting zone (state < 5). This gives realistic meaning to the simulation's estimate that an arriving call will have to wait 77.2% of the time (the true steady-state probability turns out to be 77.9%). Another feature of this particular history is that all cars were idle twice, once early and once very late in the simulation.

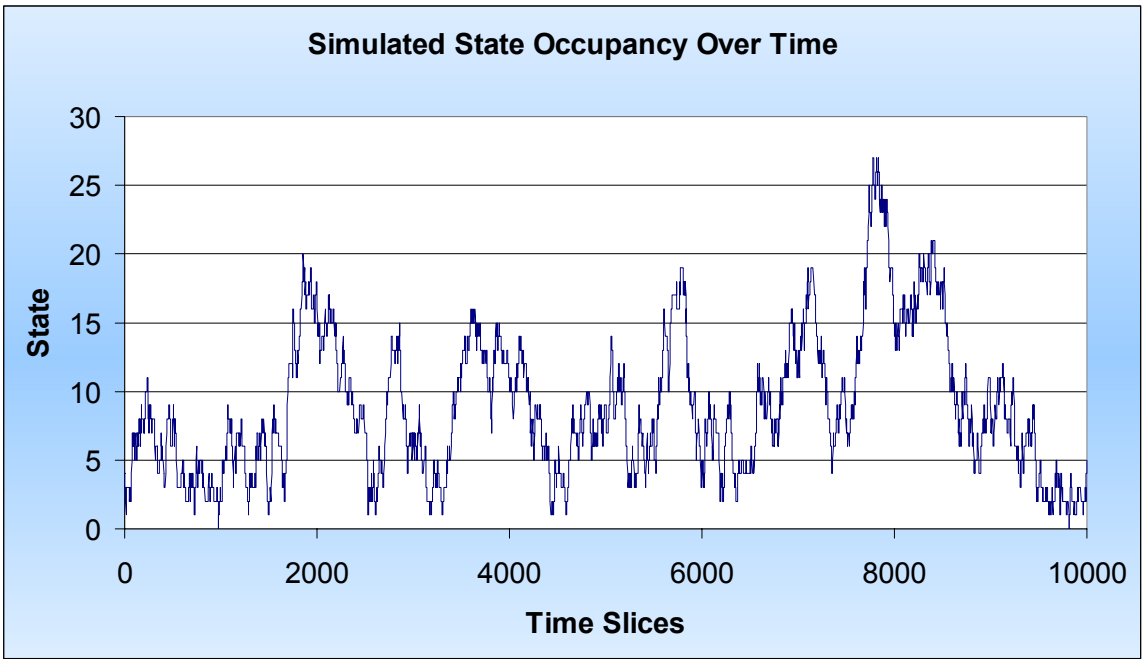


***Figure 8*. Spreadsheet Simulation of 83⅓ Hours of Operation for 5 Patrol Cars ("State" Equals Number of Calls Being Served or Waiting to Be Served)**

This spreadsheet enables non-technical people to estimate the value of P_w for any data values, including those plotted in *Figure 6*, without resorting to non-intuitive arguments. *Figure 8* also strengthens intuition concerning the dynamic behavior of patrol car utilization and the call queue.

With P_w and *Figure 6* verifiable in this way, it is time to introduce *Figure 6*'s companion for W_q using the formula given in Observation 4. See *Figure 9*.

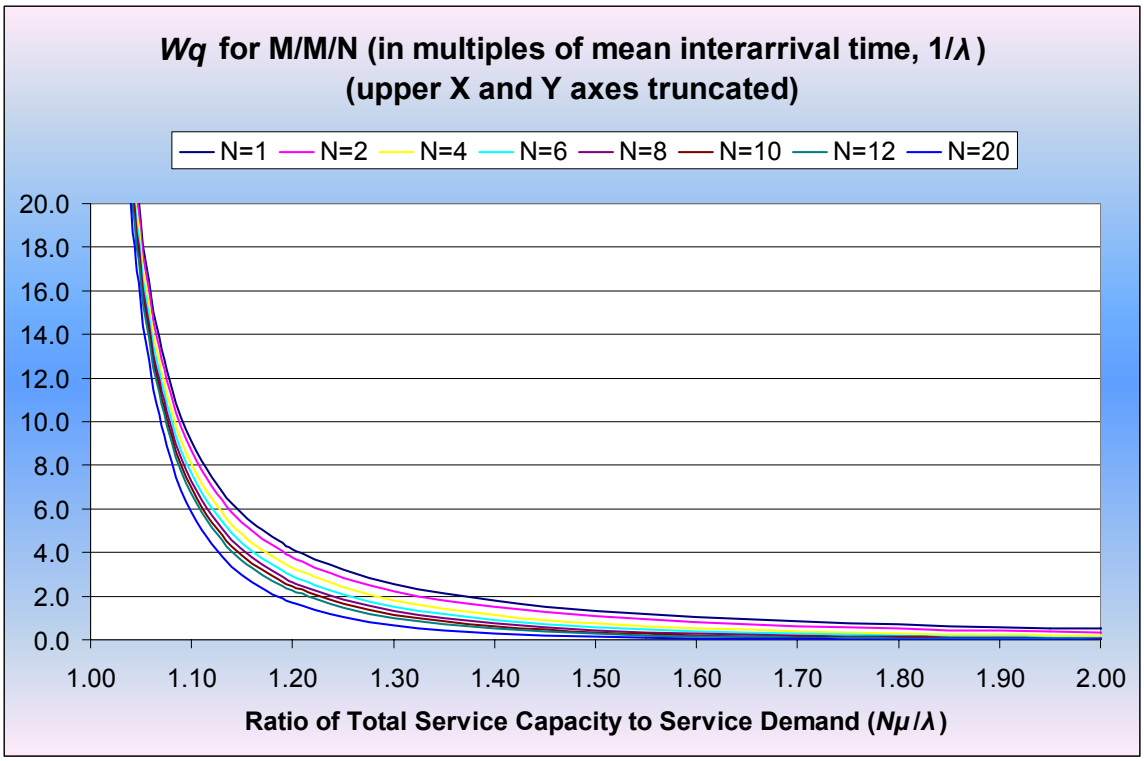


***Figure 9*. For a Given Total Service Capacity $N\mu$, Increasing N Reduces W_q (W_q values from P_w and the formula of Observation 4)**

This has been a fairly lengthy development. Here is a summary of the story that could be told to non-technical managers concerning why the detailed simulation results are what they are in *Figures 3A* and *3B*.

1. Changing the number N of patrol cars allocated to a precinct changes the precinct's service capacity to respond to calls. If calls arrive randomly at rate λ per minute and a single patrol car can serve them at the average rate μ per minute (that is, with average service time $1/\mu$ minutes), then the (maximum) precinct service capacity is $N\mu$ calls per minute on average. This capacity must be at least as large as λ , or obviously the patrol cars will be unable to keep up with demand and the number of queued calls will grow without limit. In other words, the service supply-demand ratio $N\mu/\lambda$ must be at least 1. This ratio should be small for reasons of economy, but larger values lead to better public service.

2. Two measures of service from the public viewpoint are P_w, the long-term chance that an arriving call finds all N patrol cars busy and therefore has to wait before dispatch, and W_q, the long-term average time that an arriving call spends in the queue awaiting a patrol car to become available for dispatch.

3. If it were possible to perfectly coordinate service demand and availability, then both P_w and W_q could be maintained at 0 even if $N\mu/\lambda$ were reduced all the way to its natural barrier value of 1 (Observation 2A). Such coordination requires

sufficient control over either inter-arrival times or service times, or over both, so that no arriving call need ever encounter all cars busy. The exercise of such control need not alter the average inter-arrival or service time. Such control is impractical in the present context, but this observation does reveal that the lack of it is solely responsible for positive values of P_w and W_q when $N\mu/\lambda > 1$.

4. In the absence of any coordination, both P_w and W_q deteriorate increasingly rapidly as the service supply-demand ratio $N\mu/\lambda$ approaches its natural barrier value of 1. For the special case of one patrol car, *Figure 4* shows what happens to P_w (Observation 3) and *Figure 5* shows what happens to W_q (Observation 4). For the general case of N patrol cars, graphing W_q ′s behavior is easy (Observation 4) once P_w′s behavior is known, but unfortunately P_w′s behavior is not elementary.

5. *Figure 6* shows P_w′s behavior based on a classical queuing theoretic result. It is elementary that increasing the number of patrol cars lowers the corresponding P_w-curves, since more cars for a given service supply-demand ratio means more flexible use of idle service capacity. But by exactly how much the curves drop is not elementary. Fortunately, simple spreadsheet calculations can verify the P_w-curves of *Figure 6* and, more generally, estimate P_w for any choice of N and $N\mu/\lambda$.

6. The spreadsheet shown in *Figure 7*, which mimics a random walk with a natural interpretation, is easy to understand and estimates P_w for any data values. Moreover, each spreadsheet recalculation yields a new realization for such a walk and plots it in a chart like *Figure 8* that helps build intuition about the modeled dynamic behavior of the patrol car fleet.

7. Using the spreadsheet-verified P_w values of *Figure 6* and Observation 4 to convert these to values of W_q, one obtains *Figure 9*, a natural companion to *Figure 6*. *Figure 9* shows how, for any fixed service supply-demand ratio, having more patrol cars reduces the inefficiencies resulting from an inability to coordinate calls with patrol car availability.

8. The analytic queuing model predictions plotted in *Figure 3* are but an application of *Figures 6* and *9* for specific data values. These predictions are remarkably close (especially after adjustment for systematic bias) to the detailed simulation results also plotted in *Figure 3* (Ignall et al. 1978). Therefore, the analytic model, which has been developed in a managerially transparent way, provides a satisfactory “why” explanation for the detailed simulation model results.

3. CONCLUSION

Two examples have shown how computational results from detailed models can be explained at an aggregate level in ways that mathematically naïve managers can

understand. Each explanation addresses key results of managerial interest with the help of a *vital principle* of the system at hand that determines "what is really going on" behind those results.

In the first example, which requires large-scale optimization to properly determine optimal DC locations and associated decisions, the explanation addresses the aggregate issue often of greatest managerial interest: how total cost increases with departure from the optimal number of DCs assuming that the system is otherwise fully optimized. The underlying vital principle is the simple unit cost trade-off shown in *Figure 1*.

In the second example, which potentially requires discrete event simulation of a queuing system to properly determine how much congestion results from having a given number of patrol cars on duty at a police precinct, the explanation addresses two key aggregate operating characteristics as a function of the number of patrol cars: P_w and W_q. The development adopts the underlying vital principle that arriving calls experience congestion only because there is no coordination between demands for service and the availability of service capacity; service supply and demand are blind to one another. This choice of vital principle deliberately favors managerial insight over mathematical insight. Other choices are possible.

It is dangerous to generalize from just two examples, but these are still suggestive. For instance, they suggest some of the ways in which transparency can be lost. The first illustrative application, which falls within the mathematical programming paradigm, loses transparency for at least three reasons: the sheer magnitude of the model data required, the technical nature of the algorithmic calculations required, and the scale of the computations necessary for true optimization. The second illustrative application, which falls within a very different paradigm (queuing), loses transparency for quite different reasons: the sheer mass of system operating history when logged over an extended period to get at long-term behavior, and the mathematical sophistication needed to derive system operating characteristics for predicting either transient or long-term behavior (such arcana as differential equations, generating functions, and Laplace-Stieltjes transforms). Very likely, most mathematical programming and queuing applications lose transparency for similar reasons.

Another way in which the two examples may be suggestive is that their vital principles may generalize to other applications within the same general paradigms. Cost trade-offs are at the heart of most mathematical programming models that seek to minimize some notion of total cost, although the particular driving trade-offs vary from model to model. There are trade-offs among costs or other measures of merit associated with most queuing models also, but those may or may not yield the vital principle(s) responsible for the aggregate system behavior(s) of greatest interest to management. Lack of coordination between service demand and capacity owing to randomness, lack of communication, and other reasons may be a useful default vital principle for other queuing applications because it points to possible ways to reduce congestion.

Finally, the general approach taken above in these two examples is essentially the same even though mathematical programming and queuing applications may seem to have almost nothing in common. This suggests that this general approach may be useful for other applications. It comprises the following 5 steps.

1. Choose aggregate system characteristics of great managerial interest in the context of the practical problem at hand.

   Example 1: Choose the optimal number of DCs and how total cost increases with departure from this number.

   Example 2: Choose P_w and W_q as functions of N.

2. Identify the vital principle (or principles) largely responsible for the chosen system characteristics.

   Example 1: The trade-off between unit DC fixed cost and unit DC delivery cost.

   Example 2: The lack of coordination between service supply and demand in a multi-server queuing system.

3. Formulate a conceptually simple and tractable model in accord with the vital principle(s) that can predict the chosen system characteristics.

   Example 1: See (1) and (2) (illustrated by *Figure 1*) and their associated assumptions. Equations (3) – (7) give consequent system characteristics.

   Example 2: The classical M/M/N priority queuing model was ready-made for this purpose.

4. Instantiate the simple model's predictions for the chosen system characteristics and verify that they agree reasonably well with the detailed computational results for the numerical case at hand.

   Example 1: See *Figure 2*.

   Example 2: See *Figure 3*.

5. Assemble these steps into a story that managers can understand and that lays bare for them, with the help of the vital principle(s), why the detailed computational results are what they are for the chosen aggregate system characteristics. This requires rendering managerially transparent the simple model and its predictions through conceptually simple arguments, pictures, and spreadsheets.

   Example 1: Transparent explanations were given for all essential non-elementary results, namely (2) (which ordinarily requires taking an integral) and (7) (which ordinarily requires elementary optimization theory and taking derivatives).

   Example 2: Most of Section 2 is devoted to developing transparent

explanations that circumvent the higher mathematics ordinarily used to derive the functions $P_w(N)$ and $W_q(N)$.

The aim of Steps 1-4 is for an analyst to understand the *why* behind the *what* of detailed computational results. Any and all analytical methods and tools are admissible. The aim of Step 5 is to transfer this understanding to managers who may have little or no analytical training; consequently, only the most elementary methods and tools are admissible.

A vital principle introduced at Step 2 need not be a key axiom or lemma or theorem, or a deep mathematical insight. Its role is not so much analytical as it is to provide a key thread in the story of Step 5. We favor aiming for a "story" in Step 5, rather than a "simple derivation", in the belief that this mindset will lead to better communication.

Carrying out this general approach has its challenges. Not the least of these is Step 5, where elementary means must be found to explain results that normally might require an arbitrary amount of mathematics, logic, and computational cycles. The constraints of managerial transparency are daunting, even with the focus on aggregate rather than detailed results. To the extent that transparency can be achieved for aggregate results of interest to decision and policy makers, the odds should improve that decision support systems will find greater acceptance and use.

4. ACKNOWLEDGMENTS

My sincere thanks go to Max Moroz, who assisted in important ways, and to Aydin Alptekinoglu, who made valuable suggestions on the final manuscript.

5. REFERENCES

Bos, C., *Spatial Dispersion of Economic Activity*, Rotterdam University Press, 1965.

Cobham, A., "Priority Assignment in Waiting Line Problems," *Operations Research*, **2**:1 (February, 1954), pp. 70-76.

Geoffrion, A., "The Purpose of Mathematical Programming is Insight, Not Numbers," *Interfaces*, **7**:1 (November, 1976), pp. 81-92.

Geoffrion, A., "Making Better Use of Optimization Capability in Distribution System Planning," *AIIE Transactions*, **11**:2 (June, 1979), pp. 96-108.

Geoffrion, A. and R. Powers, "Twenty Years of Strategic Distribution System Design: An Evolutionary Perspective," *Interfaces*, **25**:5 (September-October, 1995), pp. 105-127.

Ignall, E., P. Kolesar and W. Walker, “Using Simulation to Develop and Validate Analytical Models,” *Operations Research*, **26**:2 (March-April, 1978), pp. 237-253.

Penzias, A., "The Next Fifty Years: Some Likely Impacts of Solid-State Technology," *Bell Labs Technical Journal*, **2**:4 (Autumn, 1997), pp. 155-168.

Rutten, W., P. van Laarhoven, and B. Vos, “An Extension of the GOMA Model for Determining the Optimal Number of Depots,” *IIE Transactions*, **33** (2001), pp. 1031-1036.