

Research Article

A Mathematical Modelling Approach for Systems Where the Servers Are Almost Always Busy

Christina Pagel,¹ David A. Richards,² and Martin Utley¹

¹ *Clinical Operational Research Unit, University College London, 4 Taviton Street, London WC1H 0BT, UK*

² *Psychology, College of Life and Environmental Sciences, University of Exeter, Perry Road, Exeter EX4 4QG, UK*

Correspondence should be addressed to Christina Pagel, c.pagel@ucl.ac.uk

Received 15 November 2011; Accepted 23 January 2012

Academic Editor: John Hotchkiss

Copyright © 2012 Christina Pagel et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The design and implementation of new configurations of mental health services to meet local needs is a challenging problem. In the UK, services for common mental health disorders such as anxiety and depression are an example of a system running near or at capacity, in that it is extremely rare for the queue size for any given mode of treatment to fall to zero. In this paper we describe a mathematical model that can be applied in such circumstances. The model provides a simple way of estimating the mean and variance of the number of patients that would be treated within a given period of time given a particular configuration of services as defined by the number of appointments allocated to different modes of treatment and the referral patterns to and between different modes of treatment. The model has been used by service planners to explore the impact of different options on throughput, clinical outcomes, queue sizes, and waiting times. We also discuss the potential for using the model in conjunction with optimisation techniques to inform service design and its applicability to other contexts.

1. Introduction

Health treatment activities where arriving patients might have to wait for treatment and where duration of treatment follows a certain probability distribution have often been modelled using queueing theory. A classic example is the study of accident and emergency departments in acute hospitals [1, 2]. However in situations where “treatment” consists of a set of distinct treatment types, with the possibility of queues at each treatment stage, and the possibility of receiving a given type of treatment more than once, the use of queueing theory becomes very complex [3, 4].

The provision of mental health care for depression and anxiety in the primary care system is one such complex system. A configuration for mental health care delivery called “stepped care” is advocated for patients with common mental health problems [5, 6] to replace traditional systems (see Figure 1). Stepped care [7] is based on two principles: (a) “least burden,” so that an intervention received by a patient should be effective and appropriate whilst burdening the patient and the health care system as little as possible

[8] and (b) “self-correction,” the provision of a system in place to detect lack of improvement, which in turn leads to alternative more intensive treatments being offered [9]. Thus, in a stepped care system, patients are typically first considered for low-intensity interventions such as guided self-help, group work, or a short course of individual therapy. High-intensity interventions typically involve many sessions with a highly trained professional, such as a cognitive behavioural therapist. There are many different types of both low-intensity and high-intensity interventions, and an individual patient can step both “up” to, or between, high intensity treatments and “down” to, or between, low-intensity treatments.

The introduction of stepped care within an existing mental health care framework is challenging. Planning the delivery of stepped care requires decisions concerning the treatments to be offered, the number and type of staff, the protocol for how patients transfer between treatments, and the balance of provision between low and high intensity treatments. The other key feature of mental health care systems other than their complexity is that they are often

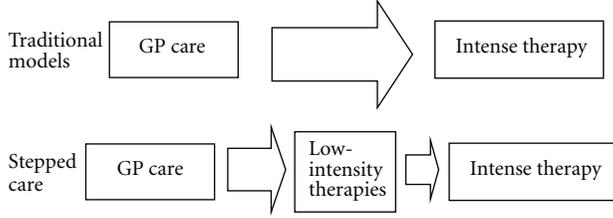


FIGURE 1: A diagram highlighting the difference between a traditional care model and the stepped care model. GP: General Practitioner.

operating at capacity. This is more feasible than in acute hospital environments since patients in the queue effectively wait “at home” using little resource.

In this paper we describe a mathematical model we have developed to help planners design a stepped care mental health system [10, 11], by providing rapid estimates of throughput and changes in waiting times for different potential configurations. This model was implemented within a software tool that was distributed to pilot primary mental health care providers [11]. We note that the mathematics presented here, although discussed in the specific context of mental health care delivery, is intended to be generic and is suitable for many other systems that meet certain assumptions.

2. Methods: Mathematical Model

The approach used is complementary to traditional queueing theory and is most suited to systems where traffic intensities are greater than or equal to one or to a system where the starting states have large queues. It complements recent work on queueing systems where some of the servers are always busy [12]. Additionally, this analysis is not dependent on the distributions of arrivals or duration of treatment. We note that where the traffic intensity is greater than one, there is no mathematical steady state to the system (since queues will increase indefinitely). However, the features of the system can still be explored within a specified time frame to understand better the distribution of demand between servers and the potential impact of increased or redistributed capacity.

2.1. A Single Treatment Slot

2.1.1. Assumptions. The unit of capacity we consider in this analysis is a time slot in a diary (e.g., a therapy session) and we assume that patients are treated in discrete sessions. A given time slot in a diary is assumed to be devoted to one, and only one, distinct treatment type. We further assume that a patient takes at least one session to be treated, and that at the start of the modeled period there is no patient currently undergoing treatment (i.e., at $t = 0$ the time slot is either empty or a patient has just started their treatment). We assume that durations of treatment of different patients are independent of one another.

2.1.2. Notation. For $x \geq 1$, let p_x denote the probability that a patient’s treatment time is exactly x time units. Define $p_0 = 0$.

For $x \geq 1$, let s_x denote the probability that a patient’s treatment time is strictly longer than x time units. Define $s_0 = 1$.

For $i \geq 1$, $t \geq 1$ let $r_{i,t}$ be the probability that exactly i patients have completed their treatment and that no other patient has started their treatment by time t . Define $r_{i,0} = 0$ for $i \geq 1$.

For $i \geq 1$, $T \geq 1$, let $f_{i,T}$ be the probability that at time T exactly i people have completed their treatment. Note that another patient may have started. Define $f_{i,0} = 0$ for $i \geq 1$.

2.1.3. The Distribution for the Number of People Who Have Completed Treatment by Time T . We begin by considering $r_{1,t}$, the probability that by time $t \geq 1$ exactly one person has arrived and left and no one else has yet started

$$r_{1,t} = p_t. \quad (1)$$

We can then define $r_{i,t}$ iteratively:

$$r_{i,t} = \sum_{k=1}^t r_{i-1,k} r_{1,t-k} = \sum_{k=1}^t r_{i-1,k} p_{t-k}. \quad (2)$$

Thus we have derived an expression for the probability that at some time t , i people have been treated and no one else has yet started. To relax this latter condition, we now consider $f_{i,T}$, the probability that at some time T , exactly i people have completed their treatment. We use $r_{i,k}$ to calculate this and for a given time T :

$$f_{i,T} = \sum_{k=1}^T r_{i,k} s_{T-k}. \quad (3)$$

2.2. A Network of Treatment Slots. We now extend the concept of a single treatment slot to a network of treatment slots that can be thought of as representing a given system.

2.2.1. Notation. Consider a treatment slot of type i , for $i = 1 \cdots L$ and absorbing exit states of type i , for $i = L+1 \cdots M$. There are a constant number, N_i , of each type of treatment slot or exit state for $i = 1 \cdots M$ where $N_i = 1$ for $i = L+1 \cdots M$.

Let α_{ij} be the probability that a person leaving a slot of type i , for $i = 1 \cdots M$ goes to a treatment slot or exit state of type j for $j = 1 \cdots M$. Define $\alpha_{ij} = 0$, for all $i = L+1 \cdots M$ and $j = 1 \cdots M$ (i.e., no one leaves an exit state).

Define the random variable X_i as the number of people who have left a single treatment slot of type i , in a given time period for $i = 1 \cdots L$. The expectation and variance of X_i are denoted $E(X_i)$ and $\text{Var}(X_i)$, and we assume that these quantities are well defined.

Let λ_i be the Poisson arrival rate from outside the system to a slot of type i . Define $\lambda_i = 0$, for all $i = L+1 \cdots M$ (i.e., no arrivals to exit states from outside the system).

Define the random variable W_{ij} as the number of people who have arrived at the queue for any slot of type j from

a single treatment slot i , in a given time period for $j = 1 \cdots M$ and $i = 1 \cdots L$.

Define the random variable Y_{ij} as the number of people who have arrived at the queue for any slot of type j from all N_i treatment slots i , in a given time period for $j = 1 \cdots M$ and $i = 1 \cdots L$.

Define the random variable Y_j as the number of people who have arrived at the queue for any slot of type j in a given time period for $j = 1 \cdots M$.

Define $B(p)$ as the Bernoulli distribution with parameter p , where $0 \leq p \leq 1$.

Define $G_Y(s)$ as the generating function associated with any given probability distribution Y .

2.2.2. General Results from Probability Theory. For a constant number N_i of independent random variables X_i ,

$$\begin{aligned} E\left(\sum_{k=1}^{N_i} X_i\right) &= \sum_{k=1}^{N_i} E(X_i) = N_i E(X_i), \\ \text{Var}\left(\sum_{k=1}^{N_i} X_i\right) &= \sum_{k=1}^{N_i} \text{Var}(X_i) = N_i \text{Var}(X_i). \end{aligned} \quad (4)$$

If a positive integer-valued distribution Z has generating function $G_Z(s)$ and a probability distribution Y has generating function $G_Y(s)$ then the distribution $W = \sum_{k=1}^Z Y$ has generating function:

$$G_W(s) = G_Z(G_Y(s)). \quad (5)$$

Additionally the expectation and variance of Y are given by

$$\begin{aligned} E(Y) &= G'_Y(1), \\ \text{Var}(Y) &= G''_Y(1) + E(Y) - E^2(Y). \end{aligned} \quad (6)$$

Proof of these results can be found in Grimmett and Stirzaker [13].

2.3. Flows through a Treatment in a General Network. Consider a treatment in a general network as shown in Figure 2. Here we consider flows in and out of a constant number, N_j , of units of capacity of type j . In this “always full” system, people arriving at a treatment slot of type j will first enter a queue of unlimited size. The number of people in a queue waiting to enter any treatment slot of type j is denoted Q_j . In what follows, we assume that there is no balking, but balking could be added to the system by specifying a maximum queue size.

As shown in Figure 2, flows into the queue for any slot of type j can come from either other units of capacity of type $i \neq j$ or from outside the system for $j = 1 \cdots L$. In the special case of an exit state, there are no external arrivals and only one state of each type.

2.3.1. Inputs from a Treatment Slot of Type i . For each person leaving a particular treatment slot of type i , for $i = 1 \cdots L$,

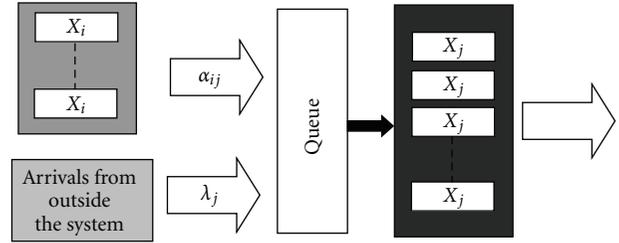


FIGURE 2: Flows in and out of treatment slots of type j .

we can consider their destination as a Bernoulli trial, where they will enter the queue for a slot of type j with probability α_{ij} . Over a given time period, we thus have a random number X_i of Bernoulli trials. Then

$$W_{ij} = \sum_{k=1}^{X_i} B_k(\alpha_{ij}). \quad (7)$$

From (5) and (6) we obtain the following (a more detailed derivation is given in the appendix):

$$\begin{aligned} E(W_{ij}) &= \alpha_{ij} E(X_i), \\ \text{Var}(W_{ij}) &= \alpha_{ij}^2 \text{Var}(X_i) + \alpha_{ij}(1 - \alpha_{ij}) E(X_i). \end{aligned} \quad (8)$$

Equation (8) give the expectation and variance for the total number of people who have arrived at the queue for a treatment slot of type j from a particular treatment slot of type i over the given time period. However, we have a block of N_i units of capacity of type i so we use (4) to derive the total number of people, Y_{ij} , who arrive at the queue for any treatment slot of type j from the block of units of capacity of type i over the given time period:

$$\begin{aligned} E(Y_{ij}) &= N_i \alpha_{ij} E(X_i), \\ \text{Var}(Y_{ij}) &= N_i (\alpha_{ij}^2 \text{Var}(X_i) + \alpha_{ij}(1 - \alpha_{ij}) E(X_i)). \end{aligned} \quad (9)$$

Thus

$$\begin{aligned} E(Y_j) &= \lambda_j + \sum_{i \neq j} \alpha_{ij} N_i E(X_i), \\ \text{Var}(Y_j) &= \lambda_j + \sum_{i \neq j} (\alpha_{ij}^2 N_i \text{Var}(X_i) + \alpha_{ij}(1 - \alpha_{ij}) N_i E(X_i)), \end{aligned} \quad (10)$$

where the sum is over all different types of treatment slot i , $i = 1 \cdots L$. Note that α_{ij} and λ_j can equal zero.

In circumstances where the network is always full, the output from units of capacity of type j for $j = 1 \cdots L$ is not dependent on the input into the queue. Thus the expected output from units of capacity of type j for $j = 1 \cdots L$ is

$$\begin{aligned} E(\text{out}) &= N_j E(X_j), \\ \text{Var}(\text{out}) &= N_j \text{Var}(X_j). \end{aligned} \quad (11)$$

TABLE 1: Flows of patients between different types of appointment and two endpoints of the stepped care system.

From\To	Assessment	Low intensity	High intensity	Completed treatment	Dropped out of treatment
Assessment	0%	40%	20%	10%	30%
Low intensity	0%	0%	20%	50%	30%
High intensity	0%	0%	0%	80%	20%

We are now in a position to consider the expectation and variance of the change in the queue size Q_j :

$$\begin{aligned}
E(\Delta Q_j) &= \sum_{i \neq j} \alpha_{ij} N_i E(X_i) + \lambda_j - N_j E(X_j), \\
\text{Var}(\Delta Q_j) &= \sum_{i \neq j} (\alpha_{ij}^2 N_i \text{Var}(X_i) + \alpha_{ij} (1 - \alpha_{ij}) N_i E(X_i)) \\
&\quad + \lambda_j + N_j \text{Var}(X_j).
\end{aligned} \tag{12}$$

Note that the variance of the change in queue size can be very large if there are a large number of potential inputs for that particular type.

For a patient waiting to receive treatment for a mental health problem, waiting time in a queue is more likely to be of concern than the actual number of people waiting. Let $1/\mu_j$ represent the mean number of sessions required to treat a patient in a unit of capacity of type j for $j = 1 \dots L$. We can estimate the change in waiting time, ΔP_j for an individual arriving in the queue Q_j as

$$\Delta P_j = \frac{E(\Delta Q_j)}{\mu_j N_j}, \quad j = 1 \dots L. \tag{13}$$

2.4. Potential for Optimisation. The linear nature of the equations above in terms of N_j suggests the possibility that linear programming techniques might be used to optimise the configuration of the stepped care system according to some relevant criteria. An illustrative optimisation is given below where there are a set of treatment types j and a desired outcome k (for instance successful discharge) and the intention is to find the allocation of sessions to types of treatments that maximises the number of people achieving the desired outcome. We note that different objective functions can be defined and that optimisation functionality was not included in the software tool [11] produced as part of this project.

2.4.1. Objective Function. Maximise:

$$Z = \sum_{j \neq k} \alpha_{jk} N_j E(X_j), \quad j = 1 \dots L, \quad L < k \leq M. \tag{14}$$

2.4.2. Constraints.

- (1) N_j are positive integers, $j = 1 \dots L$
- (2) Total number of therapy sessions per week: $\sum_{j=1}^L N_j \leq S$ for some integer S .

TABLE 2: Arrivals to the system and allocation of resources within the system.

Appointment type	Average number of new, external arrivals every week	Weekly appointment slots allocated
Assessment	20	30
Low intensity	10	40
High intensity	0	30

- (3) Total number of sessions for a treatment of type j is dependent on number of therapists qualified to deliver that type of treatment and thus there are capacity constraints for each treatment type: $N_j \leq S_j$, where S_j are positive integers, $j = 1 \dots L$.

- (4) Specify a maximum increase in waiting time of W weeks at each step:

$$E(\Delta Q_j) \leq W N_j \mu_j, \quad j = 1 \dots L. \tag{15}$$

3. Results: Illustrative Example

This mathematical model has been implemented as part of a project examining the implementation of stepped care systems [11]. Here we give an illustrative example of its use on a hypothetical mental health care system and the subsequent potential for optimisation.

In this system there are three types of appointments available: an initial screening appointment, a low-intensity therapy appointment, and a high-intensity therapy appointment. People are either referred by their GP (General Practitioner) into the system in which case they begin with an assessment appointment or they can self-refer directly to low-intensity therapy treatment.

The proportion of patients moving between different types of appointment is shown in Table 1.

The health system managers have 100 weekly appointments available to cover all types of appointment. For the purposes of this example, we assume that the number of weekly booked sessions a patient uses for each type of treatment is exponential with means of 1, 3, and 6 for assessments, low-intensity and high-intensity sessions respectively. The arrivals and capacity allocation for the current system are given in Table 2.

Although the waiting times for screening and high-intensity treatments are acceptable, the increase in waiting time for a low-intensity appointment over 6 months is unacceptably long (12 weeks) (see columns 2 and 3 of Table 3). The manager wishes to optimise the allocation of

TABLE 3: Example use of optimisation to allocate available treatment slots to treatment types.

Appointment type/End point	Current weekly appointment slots allocated	Average increase in waiting time (weeks)	Suggested weekly appointment slots allocated	Average increase in waiting time (weeks)
Assessment	30	1.3	26	4
Low intensity	40	11.7	45	6
High intensity	30	7	29	7
<i>Dropped out</i>	257	—	248	—
Completed	292	—	301	—

appointment slots to appointment types to maximise the number of patients who successfully complete treatment in a 26-week period, according to the following constraints.

- (1) There can be a maximum of 100 total allocated appointments.
- (2) The maximum increase in average waiting time for an assessment is 4 weeks.
- (3) The maximum increase in average waiting time for either low or high intensity treatment is 8 weeks.
- (4) There must be at least 20 assessment sessions, 30 low-intensity, and 20 high-intensity sessions every week.

We ran this optimisation problem using Microsoft Excel Solver (version 2003). We note that Microsoft Solver is a standard add-in to Microsoft Excel and there exist several resources on its use within Excel (e.g., [14]). The new allocation and the output parameters calculated using the model are given in columns 4 and 5 of Table 3.

The suggested appointment schedule has resulted in a more even distribution of the expected waits for each type of treatment and increased the expected total number of people completing treatment over the 6-month time frame.

4. Limitations

A clear limitation is the assumption that the system is always busy. However, application of the model to any given system would still provide the maximum possible throughput of the system over a given period of time. In the context of a mental health system, other limitations apply. Firstly, all patients are considered to be homogeneous and no allowance is made for patients with different characteristics (for instance presenting problem) having different duration of stay distributions or different pathways through the system. Secondly, in this analysis, time is considered to be defined by the number of treatment slots and thus application to a system where sessions are not regularly spaced in time is more complicated. Finally, “holding” or “blocking back” behaviour is not accounted for in the model, in that both the duration of treatment and the destination of patients from each treatment are assumed independent of the state of the system.

5. Discussion

This mathematical model was developed in response to a specific problem within the configuration of mental health care services [10]. As part of that project, the model (without the optimisation aspect) has been implemented within a software tool developed to help planners explore the consequences of different configurations for a given mental health service. Details of the software tool and its use in designing stepped care systems can be found in the project final report [11].

We have described a simple way of analysing throughput and flows for a networked system in the situation where a system is always busy or where this is a reasonable approximation. We note that this is not a steady-state model and instead considers changes in mean output, queue sizes and waiting times over a relatively short (6 months) time period. We have also shown how optimisation techniques might be applied to the subsequent design of a network in the context of a mental health system.

This approach could be useful in other health systems where “servers” (whether beds, clinicians or other resources are always busy) but a key assumption that needs to be met is that there will always be a queue. In practice this assumption is less likely to be valid for systems where queues involve people waiting in a physical allocated space (for instance, in an emergency department) than where people can “virtually” wait at home. Nonetheless, considering such busy systems over a short amount of time using these sorts of models can provide insight into the allocation of resources and management of arrivals to complement standard steady state queuing theory.

Appendix

Remember that the random variable W_{ij} is defined as the number of people who have arrived at the queue for any slot of type j from a single treatment slot i , in a given time period for $j = 1 \cdots M$ and $i = 1 \cdots L$.

For each person leaving a particular treatment slot of type i , for $i = 1 \cdots L$, we can consider their destination as a Bernoulli trial, where they will enter the queue for a slot of type j with probability α_{ij} . Over a given time period, we thus have a random number X_i of Bernoulli trials. Then

$$W_{ij} = \sum_{k=1}^{X_i} B_k(\alpha_{ij}). \quad (\text{A.1})$$

From (5) we know that

$$G_{X_{ij}}(s) = G_{X_i}(G_B(s)). \quad (\text{A.2})$$

and thus

$$G'_{X_{ij}}(s) = G'_{X_i}(G_B(s))G'_B(s), \quad (\text{A.3})$$

$$G''_{X_{ij}}(s) = G''_{X_i}(G_B(s))(G'_B(s))^2 + G'_{X_i}(G_B(s))G''_B(s).$$

It is a standard result that $G_B(s) = (1 - \alpha_{ij}) + \alpha_{ij}s$, and so using this and (6) we obtain the following:

$$E(W_{ij}) = \alpha_{ij}E(X_i), \quad (\text{A.4})$$

$$\text{Var}(W_{ij}) = \alpha_{ij}^2 \text{Var}(X_i) + \alpha_{ij}(1 - \alpha_{ij})E(X_i).$$

Acknowledgments

This project was funded by the National Institute for Health Research Service Delivery and Organisation programme (Project no. 08/1504/109). The views expressed in this publication/presentation are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. The NIHR SDO programme is funded by the Department of Health.

References

- [1] M. L. McManus, M. C. Long, A. Cooper, and E. Litvak, "Queuing theory accurately models the need for critical care resources," *Anesthesiology*, vol. 100, no. 5, pp. 1271–1276, 2004.
- [2] J. K. Cochran and K. T. Roche, "A multi-class queuing network analysis methodology for improving hospital emergency department performance," *Computers and Operations Research*, vol. 36, no. 5, pp. 1497–1512, 2009.
- [3] X. Song and M. Mehmet Ali, "A performance analysis of discrete-time tandem queues with Markovian sources," *Performance Evaluation*, vol. 66, no. 9-10, pp. 524–543, 2009.
- [4] H. Daduna and R. Szekli, "On the correlation of sojourn times in open networks of exponential multiserver queues," *Queueing Systems*, vol. 34, no. 1-4, pp. 169–181, 2000.
- [5] National Institute for Health and Clinical Excellence, "Depression. management of depression in primary and secondary care," National Clinical Practice Guideline number 23, NICE, London, UK, 2004.
- [6] National Institute for Health and Clinical Excellence, "Anxiety: management of anxiety (panic disorder, with or without agoraphobia, and generalised anxiety disorder) in adults in primary, secondary and community care," National Clinical Practice Guideline number 22, NICE, London, UK, 2004.
- [7] D. A. Richards, P. Bower, and S. M. Gilbody, "Improving access to psychological therapies," in *Handbook of Primary Care Mental Health*, L. Gask, H. Lester, R. Peveler, and A. Kendrick, Eds., Gaskell, London, UK, 2009.
- [8] M. B. Sobell and L. C. Sobell, "Stepped care as a heuristic approach to the treatment of alcohol problems," *Journal of Consulting and Clinical Psychology*, vol. 68, no. 4, pp. 573–579, 2000.
- [9] M. G. Newman, "Recommendations for a cost-offset model of psychotherapy allocation using generalized anxiety disorder as an example," *Journal of Consulting and Clinical Psychology*, vol. 68, no. 4, pp. 549–555, 2000.
- [10] S. Gallivan, M. Utley, C. Pagel, and D. Richards, "Applying operational research techniques to the reorganisation of mental health care in the UK," in *Operational Research Applied to Health Services in Action*, M. Lubicz, Ed., pp. 155–162, 2009.
- [11] D. A. Richards, A. Weaver, M. Utley et al., "Developing evidence-based and acceptable stepped care systems in mental health care: an operational research project," Final Report, NIHR Service Delivery and Organisation Programme, 2010, <http://www.sdo.nihr.ac.uk/projdetails.php?ref=08-1504-109>.
- [12] G. Weiss, "Jackson networks with unlimited supply of work," *Journal of Applied Probability*, vol. 42, no. 3, pp. 879–882, 2005.
- [13] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, Clarendon Press, Oxford, UK, 1992.
- [14] "Microsoft webpage," <http://office.microsoft.com/en-us/excel-help/introduction-to-optimization-with-the-excel-solver-tool-HA001124595.aspx>.



Hindawi
Submit your manuscripts at
<http://www.hindawi.com>

