# Tariff-Rate Quotas

## Difficult to model or plain simple?

**Paper presented at the annual conference of the New Zealand Agricultural and Resource Economics Society**

**Working Paper 2001/7**

**ISSN 1170 2583**

July 6-7 2001

**NZ INSTITUTE OF ECONOMIC RESEARCH (INC.)**

8 Halswell St. Thorndon

P O BOX 3479 WELLINGTON

Tel: (04) 472 1880

Fax: (04) 472 1211

## Preface

The New Zealand Institute of Economic Research (NZIER), based in Wellington, was founded in 1958 as a non-profit making trust to provide economic research and consultancy services. Best known for its long-established *Quarterly Survey of Business Opinion* and forecasting publications, *Quarterly Predictions* and the annual *Industry Outlook* with five-yearly projections for 25 sectors, the Institute also undertakes a wide range of consultancy activities for government and private organisations. It obtains most of its income from research contracts obtained in a competitive market and trades on its reputation for delivering quality analysis in the right form, and at the right time, for its clients. Quality assurance is provided on the Institute's work:

- by the interaction of team members on individual projects;
- by exposure of the team's work to the critical review of a broader range of Institute staff members at internal seminars;
- by providing for peer review at various stages through a project by a senior staff member otherwise disinterested in the project;
- and sometimes by external peer reviewers at the request of a client, although this usually entails additional cost.

## Authorship

This paper has been prepared by Phillip M. Bishop, Charles F. Nicholson, James E. Pratt, and Andrew M. Novakovic.

Bishop is Senior Economist at NZIER (phil.bishop@nzier.co.nz); Nicholson and Pratt are Senior Research Associates in the Department of Applied Economics and Management, Cornell University; and Novakovic is the E.V. Baker Professor of Agricultural Economics in the Department of Applied Economics and Management, Cornell University, Ithaca NY 14853-7801.

# ABSTRACT

The "difficulty" of reliably and accurately incorporating tariff-rate quotas (TRQs) into trade models has received a lot of attention in recent years. As a result of the Uruguay Round of GATT negotiations, TRQs replaced an assortment of tariff and non-tariff instruments in an effort to standardise trade barriers, and facilitate their future liberalisation. Understanding the nuances of TRQs is now particularly crucial for New Zealand because of the preferential access arrangements that New Zealand has for a number of products in highly protected markets such as the European Union, Japan, and the United States.

It has been argued that TRQs are complex instruments and are difficult to model because for any trade flow between two countries, one of three regimes may be applicable:

1. The import quota may not be binding and the within-quota tariff applies;
2. The quota may be binding, the within-quota tariff applies, and a "quota rent" is created; or
3. Trade occurs over and above the quota, in which case an over-quota tariff applies (although, even in this regime, someone is still able to collect the quota rent on within-quota trade).

But even this characterisation, which many claim is too complex to model, is a major simplification of reality. Bilateral preferences are ubiquitous, and such preferences are usually included in the determination of multilateral market access quotas. It is usual, therefore, that the TRQ instrument has several tiers to the quota schedule, plus a number of within- and over-quota tariff rates applicable on either a bilateral or a multilateral basis.

Further trade liberalisation creates something of a dilemma for New Zealand. Any decrease in over-quota tariffs and/or increase in quota levels potentially reduces the value of quota rents, many of which accrue to New Zealand due to the bilateral preferences. It is important, therefore, that New Zealand trade negotiators understand how much additional trade is required to offset the loss of New Zealand's quota rents. Modelling trade in the presence of TRQs is the only way to ascertain this knowledge.

The purpose of this paper is to show that complex TRQs can be modelled very easily and precisely. The only catch is that the model must be formulated as a *complementarity* problem rather than the more conventional linear or nonlinear optimisation problem. The concept will be demonstrated using a simple 3-region, single commodity spatial price equilibrium model of trade.

Keywords

Tariff-rate quota, trade modelling, mathematical programming, complementarity.

# CONTENTS



# FIGURES

# 1. INTRODUCTION

The use of tariff-rate quotas (TRQs) is widespread, especially in agricultural and other primary sector trade policy settings. Because a significant share of New Zealand's exports derive from such sectors, understanding the nuances of TRQs as a trade policy instrument is of critical importance. For example, one question of interest at present is how further liberalisation of trade will impact New Zealand when the volume of that trade is constrained by TRQs. Analysing the trade-off between diminished quota rents and increased trade volumes is, potentially, a non-trivial undertaking. What's more, the problem is made even more difficult because the design and operation of most TRQs is not as straightforward as simple textbook illustrations would imply.

Fortunately, many of the extant trade models can be reformulated as *mixed complementarity problems* (MCPs). They are then capable of being used to analyse even the most complex of TRQ instruments.

It is the purpose of this paper to demonstrate how complex TRQs can be effortlessly analysed within the context of a model formulated as an MCP. There is little that is original in this paper. Indeed, it draws upon a number of sources, in particular the material found in the first 8 or 9 pages of Ferris and Munson (2000), and Rutherford (1995). We hope the contribution of this paper is that some rather demanding material is presented in a format that enables it to be accessed by an audience it would not otherwise reach.

While the primary objective of the paper is to show how TRQs can be analysed, it also has a secondary objective. That is, to explain some of the jargon associated with optimisation models, mathematical programming, and complementarity, and to reveal the role in the economist's analytical toolbox of models that employ such concepts.

The paper is organised as follows. We begin with what we call a 10-minute tour through optimisation theory. This section culminates with a brief discussion of the Kuhn-Tucker conditions, which enables us to then motivate the complementarity problem. We then change tack and present a series of small models. We start with a simple transportation problem and work up to a spatial price equilibrium model formulated as a mixed complementarity problem, which is coded using the GAMS software. The next section discusses the tariff-rate quota as a policy instrument. We conclude the paper by bringing it all together, incorporating tariff-rate quotas into a simple 3-region, single commodity trade model.

We make extensive use of snippets of GAMS code to demonstrate the models presented in this paper (Brooke et al., 1998). The models presented herein are well within the dimension limitations of the free version of the GAMS software that can be downloaded from www.gams.com.

Finally, we are aware that readers of this paper might be encountering some of this material for the first time. Hence, we tend to repeat ourselves in order to continually reiterate the key points. We hope that this is helpful rather than annoying.

# 2. A 10-MINUTE TOUR OF OPTIMISATION

## 2.1 Introduction

Fundamental to most economic analysis is the notion of optimisation; consumers make choices so as to maximise utility, firms seek to maximise profits, nations search for ways to maximise GDP growth, and so on. Moreover, such optimisation problems usually have constraints associated with them. For example, consumers maximise utility subject to a budget constraint. In this paper we focus on models that find optimal solutions to economic problems. The key tools employed to do this fall under the rubric of *mathematical programming*. To be quite clear at the outset, we are not talking about statistical, or econometric, models, and nor are we talking about a set of simple accounting or arithmetic relationships that might be found in a spreadsheet application.

The purpose of this section is to very quickly introduce a few of the key concepts found in mathematical programming.[1] Later in the paper, we will use examples of simple but realistic models to further scrutinise these concepts. For now we are simply interested in establishing a convenient starting point, and in introducing some of the jargon.

## 2.2 Unconstrained optimisation

The simplest type of optimisation problem is one where we wish to find the extreme value (either a minimum or a maximum) of some function. In actual fact, a problem this simple does not even require the tools of mathematical programming; the classical techniques of differential calculus are all that is required.

Consider the following long-run average cost function:

$$AC = f(Q) = Q^2 - 5Q + 8 \tag{1}$$

Here we have a quadratic function, which when plotted will reveal a U-shaped curve with its lowest point occurring when Q = 2.5. In other words, when Q = 2.5, AC will equal 1.75, representing the lowest value that this long-run average cost function can obtain. In this particular problem, we say that AC = *f(Q)* is the objective function and Q is the choice variable, or decision variable. There are no constraints to this problem, which implies we have an unconstrained optimisation problem. If our interest is in minimising the long-run average cost, then our task is to *choose* the appropriate level of the decision variable, Q, that is consistent with the objective function yielding its lowest possible value. We would follow the same procedure if there were more than one decision variable or if the problem was one to be maximised.

As an aside, one may restrict the domain of Q to be non-negative as it would make no sense to contemplate producing a negative amount of Q. However, we don't have to think of such a restriction as a constraint.

The standard approach to solving an unconstrained optimisation problem is to use the tools of differential calculus. In the case of our long-run average cost function, it will be recalled from high school calculus that we simply take the first derivative, set it equal

[1] There are many good textbooks offering a comprehensive treatment of mathematical programming and its underlying theory. A gentle introduction is: Chiang, A. (1984). *Fundamental methods of mathematical economics* (3rd edition), McGraw-Hill. Indeed, much of the material in this section on optimisation theory is drawn directly from Chiang.

to zero, and solve for the level of Q. For example, letting *f′(Q)* denote the first derivative of *f*, we have:[2]

$$f'(Q) = 2Q - 5 \tag{2}$$

which, when set equal to zero, yields Q = 2.5.

This problem is easy because the function is quadratic, which means that it only has one turning point. A slightly more complicated problem is a cubic function, which has two turning points. For example,

$$y = f(x) = x^3 - 12x^2 + 36x - 8 \tag{3}$$

Taking the first derivative and setting it equal to zero yields two "roots", or levels at which *f′(x)* = 0; namely x = 2 and x = 6. But now we need to determine which of these roots is associated with the maximum point and which is the minimum – we know that a cubic function will have one of each. We do this by taking the second derivatives, i.e. the first derivative of the first derivative. Armed with this information, we can then determine which way the function, *f(x)*, is turning as *x* approaches 2 and 6. In other words, is the function concave or convex at these points? We will cease this exposition at this point as it can be reviewed in any elementary calculus textbook. Suffice it to say that the theory of unconstrained optimisation generalises to *n*th-degree polynomial functions.

Before turning to constrained optimisation though, it is worth noting a couple of final points. First, one needs to be careful about referring to extreme points as maximum or minimum points of the primitive function. They should more properly be referred to as *relative* (or *local*) extrema, as a function may have several extreme points, some of which may be maxima while others are minima. Only one of each can be the *global* (in contrast to the local) maximum or minimum. And some points at which the first derivative equals zero may fall into a third category, points of inflection.

Second, differential calculus has been used to derive a universal set of necessary and sufficient conditions. These conditions, or rules, can easily be applied to any function to determine the status of all stationary points. That is, all points where the first derivative or the slope of the function is zero. The necessary conditions are based on the first derivatives and are therefore referred to as the *first-order conditions*. Likewise, the sufficient conditions are based on the second derivatives, and are referred to as *second-order conditions*.

## 2.3 Optimisation with equality constraints

It is nearly always the case in economics that some limiting factor(s) impinges upon the choices we are able to make when trying to solve any given optimisation problem. An obvious example is when firms set out to maximise profits; they are constrained by the need to employ the available technology. Hence it is necessary to have a technique available for solving *constrained optimisation* problems. The standard method is that of *Lagrange multipliers*. The essence of the Lagrange multiplier method is to convert the constrained optimisation problem into a form where the first-order conditions of the unconstrained problem can still be applied.

By way of example, consider the utility function

$$U = x_1 x_2 + 2x_1 \tag{4}$$

[2] Whereas *f′* is termed the derivative function, the original function, *f*, is sometimes referred to as the primitive function.

and the budget constraint given by

$$4x_1 + 2x_2 = 60 \tag{5}$$

The first step in employing the method of Lagrange multipliers is to write down what is termed the *Lagrangian function*:

$$Z = x_1x_2 + 2x_1 + \lambda(60 - 4x_1 - 2x_2) \tag{6}$$

This is nothing more than a modified version of the primitive objective function that incorporates the constraint (or constraints if there is more than one).[3] Z is now the objective function value, i.e. the value we are trying to optimise (maximise in this case). The symbol λ is called the Lagrange multiplier and represents an as yet undetermined number. Notice how the constraint was rearranged when we wrote out the Lagrangian function. If we can be assured that the constraint will be satisfied, then the expression inside the brackets in our Lagrangian function goes to zero, and the entire last term of equation (6) vanishes, regardless of the value of λ. With the constraint thus dispensed with, we could then seek the *free* (unconstrained) value of Z in lieu of the constrained value of U. The question, then, is how do we make the term in parentheses in equation (6) vanish?

The tactic employed by the method of Lagrange multipliers is to treat λ as an additional variable in (6). That is, consider Z to be a function not only of $x_1$ and $x_2$, but also of λ, e.g. $Z = Z(\lambda, x_1, x_2)$. The first-order conditions for the free extremum (i.e. maximum in this case) will now consist of a set of simultaneous equations:

$$\begin{aligned} Z_\lambda &\equiv \partial Z/\partial\lambda = 60 - 4x_1 - 2x_2 = 0 \\ Z_1 &\equiv \partial Z/\partial x_1 = x_2 + 2 - 4\lambda = 0 \\ Z_2 &\equiv \partial Z/\partial x_2 = x_1 - 2\lambda = 0 \end{aligned} \tag{FOC 1}$$

So, by taking the first derivative of Z with respect to each of the three variables, and by setting each derivative to be equal to zero, we have generated the first-order conditions. They are nothing more than a set of 3 unknowns and 3 equations. But significantly, the first of the three equations in (FOC 1), i.e. $Z_\lambda$, gives us the assurance that the last term in equation (6) will vanish. It is a simple matter to solve this system and find that $\lambda = 4$, $x_1 = 8$, and $x_2 = 14$.[4] Putting these values into equations (6) and (4) yields, respectively, Z = 128 and U = 128.

For our present purposes, we need not take this discussion any further; we are now ready to consider problems with inequality constraints. Chapter 12 of Chiang (1984) covers optimisation with equality constraints in much greater detail and is worth reviewing. For example, it covers such topics as the interpretation of the Lagrange multiplier, the *n*-variable and *m*-constraint case, second-order conditions, and the significance of testing for concavity and convexity.

## 2.4 Optimisation with inequality constraints

Thus far in our quick tour through optimisation theory we have been able to confine ourselves to using the methods of classical optimisation, i.e. techniques based on

[3] When there is more than one constraint, the Lagrangian function would have a distinct λ associated with each one.

[4] For example, rearrange $Z_2$ to yield $\lambda = 0.5x_1$. Then substitute this expression for λ into $Z_1$ yielding $x_2 + 2 - 4(0.5x_1) = 0$. Rearranging this gives $x_2 = 2x_1 - 2$, which we can then substitute into $Z_\lambda$. Rearranging the result of that gives us $x_1 = 8$. If $x_1 = 8$, then $Z_2$ must imply that $\lambda = 4$, and finally, putting $\lambda = 4$ into $Z_1$ reveals that $x_2 = 14$.

differential calculus. In order to tackle problems with inequality constraints, and for some other reasons that will become obvious shortly, we now need to move into the realm of *non*classical methods. Mathematical programming, which includes (among other topics) linear programming, nonlinear programming, and complementarity methods, is the name given to this collection of *non*classical solution techniques. Clearly the introduction of inequality constraints enables more interesting and realistic problems to be contemplated.

The specific models we get to later in the paper deal with problems containing inequality constraints. Hence we will not belabour their presentation at this juncture. What we wish to do in this section is introduce the *Kuhn-Tucker conditions*, and then provide a brief introduction to the mixed complementarity problem.

The types of problems encountered in *classical* optimisation have three key characteristics:

- They contain no explicit restrictions on the sign of the choice variables;
- They contain no inequality constraints; and
- The first-order conditions for a relative or local extremum is simply that the first partial derivatives of the Lagrangian function with respect to all choice variables and the Lagrange multipliers be zero. In other words, we are restricted to situations where there are no boundary or corner solutions. Stated differently, we have only interior solutions.

Mathematical programming methods enable us to find solutions to problems where one or all of these characteristics are not present.

### 2.4.1 The Kuhn-Tucker conditions

The single most important result in nonlinear programming is the Kuhn-Tucker conditions (Kuhn and Tucker, 1951). These conditions can be thought of as the nonlinear programming equivalent of the classical first-order conditions. As we shall see, the Kuhn-Tucker method generalises the first-order conditions *for an equilibrium* to a set of boundary conditions *for finding an equilibrium*. But unlike those classical conditions, the Kuhn-Tucker conditions cannot be accorded the status of necessary conditions, unless a certain proviso is satisfied.[5] However, in specific circumstances, the Kuhn-Tucker conditions turn out to be sufficient conditions, or even necessary and sufficient conditions as well.

At this point we are going to do little more than present the derivation of the Kuhn-Tucker conditions. Their relevance will become clearer later in the paper. Our primary reason for presenting them at all is because they provide a bridge from the theory of constrained optimisation to the mixed complementarity problem.

Consider the following generic optimisation problem, which incorporates two constraint functions and explicit nonnegativity conditions:

[5] That proviso is called the *constraint qualification*. It imposes a certain restriction on the constraint functions of a nonlinear program so that irregularities on the boundary of the feasible set don't invalidate the Kuhn-Tucker conditions, should the optimal solution occur at that boundary. While the constraint qualification is important to the theory of nonlinear programming, and to the Kuhn-Tucker sufficiency theorem, we are not going to discuss it in this paper. A nonlinear programming text should be consulted for further information.

$$\begin{aligned} &\textit{Maximise} && \pi = f(x_1, x_2, x_3) \\ &\textit{subject to} && g^1(x_1, x_2, x_3) \le r_1 \\ & && g^2(x_1, x_2, x_3) \le r_2 \\ & && x_1, x_2, x_3 \ge 0 \end{aligned} \qquad \text{(NLP 1)}$$

We can imagine that f and $g^1$, say, are nonlinear functions. Hence we call this problem (NLP 1) for NonLinear Program. As before, we can write the Lagrangian function for this problem:

$$Z = f(x_1, x_2, x_3) + \lambda^1\left(r_1 - g^1(x_1, x_2, x_3)\right) + \lambda^2\left(r_2 - g^2(x_1, x_2, x_3)\right) \qquad (7)$$

The resulting Kuhn-Tucker conditions can be stated compactly as follows (or more accurately, one version of the Kuhn-Tucker conditions, expressed in terms of the Lagrangian function Z, can be stated this way):

$$\begin{aligned} &\frac{\partial Z}{\partial x_j} \le 0, \quad x_j \ge 0, \quad \text{and} \quad x_j \frac{\partial Z}{\partial x_j} = 0 \qquad \forall \text{ j} = 1, 2, 3 \\ &\frac{\partial Z}{\partial \lambda^i} \ge 0, \quad \lambda^i \ge 0, \quad \text{and} \quad \lambda^i \frac{\partial Z}{\partial \lambda^i} = 0 \qquad \forall \text{ i} = 1, 2 \end{aligned} \qquad \text{(KT 1)}$$

What do these six expressions comprising the Kuhn-Tucker conditions tell us?

First of all, two of the six expressions are nothing more than a restatement of parts of the original problem, i.e. $x_j \ge 0$ simply restates the nonnegativity conditions for the three *primal* variables. Similarly, $\partial Z/\partial \lambda^i \ge 0$ reiterates the constraints, i.e. one such condition for each of the two constraints. But notice that associated with each variable type, i.e. the choice variables ($x_j$) and the Lagrange multipliers ($\lambda^i$), there is a corresponding marginal condition that must be satisfied by the optimal solution.[6]

Finally, there is what is known as the *complementary slackness* conditions, i.e. the last two expressions on each line of (KT 1), which simply state that the product of two terms must equate to zero. In other words, each variable is characterised by complementary slackness in relation to a particular partial derivative of the Lagrangian function Z. And what does this mean? It means that for each $x_j$, we must find in the optimal solution that either:

- The marginal condition holds with a strict equality (as in the classical context); or
- The choice variable in question must take on a zero value; or
- Both of the above.

Analogously, for each $\lambda^i$, we must find in the optimal solution that either the associated marginal condition holds as an equality – meaning that the $i^{th}$ constraint is satisfied exactly – or the Lagrange multiplier vanishes, i.e. becomes zero, or both.

It is by exploiting the complementary slackness conditions that it becomes possible to find corner or boundary solutions. Of significance in the case of models that explicitly incorporate tariff-rate quotas, it is the property that enables endogenous regime switching to occur. In fact, short of exploiting complementary slackness, there exists no

[6] In case it is not yet apparent, it is worth noting at this point that the first-order conditions, of both constrained and unconstrained optimisation problems, describe a set of equations that must hold true at the equilibrium or optimal solution. They are sometimes thus referred to as the *first-order equilibrium conditions*. So, we have two ways to describe the problem; the underlying optimisation problem, and a set of conditions that characterise (i.e. must hold true) a solution to the underlying optimisation problem.

other way to explicitly model such behaviour.[7] We will return to this point later.

Before we turn to the mixed complementarity problem, we would point out that the Kuhn-Tucker conditions give rise to a natural economic interpretation. The Lagrange multipliers can be regarded as shadow prices. Thus, the Kuhn-Tucker conditions tell us that, in an optimal solution, when a constraint holds with a strict inequality, then by complementary slackness, the associated shadow price must be zero. Similarly, if an activity level (e.g. a primal variable) is strictly greater than zero, then the associated marginal condition must hold with a strict equality. We will return to this in some detail shortly.

### 2.4.2 The Mixed Complementarity Problem (MCP)

Now that we have understood the Kuhn-Tucker conditions, we are ready to examine complementarity. A nonlinear complementarity problem consists of a system of simultaneous (linear or nonlinear) equations that are written as inequalities and are linked to bounded variables in a manner that encapsulates complementary slackness relationships. In a mixed complementarity problem (MCP), the equations may be a mixture of inequalities and strict equalities. For more rigorous details and a mathematical definition, see Rutherford (1995).

It is possible to rewrite (KT 1) in accordance with this definition, and thereby transform (NLP 1) into an equivalent MCP. However, the generic nature of the problem specified in NLP 1 means that the expression for the equivalent MCP may well be confusing for the novice. To avoid this possibility, we would prefer to derive an MCP using a specific example, and shall do so shortly.

Before leaving this section, however, we would make a few general comments relating to MCPs.

The underlying theory of complementarity, and the closely related *variational inequality* was developed in the 1960s, e.g. see Cottle et al., 1992; Lemke and Howson, 1964; Hartman and Stampacchia, 1966. For a thorough review, see Harker and Pang (1990). It was not until much later that commercially available algorithms capable of solving large scale complementarity problems became available, e.g. see Rutherford (1993) and Dirkse and Ferris (1993). The solver we use is called PATH (Dirkse and Ferris, 1993), which is seamlessly linked with the GAMS modelling software (Brooke et al., 1998).

As noted earlier, the equilibrium conditions that characterise an underlying optimisation problem are an alternative way of expressing that problem. But there are some problems that can be formulated as an MCP for which there is no equivalent underlying optimisation problem. Such problems arise frequently in economics. Hence, the MCP offers the modeller greater choice and flexibility than traditional NLP formulations. A pertinent case in point is the explicit treatment of tariff-rate quotas, which requires a modelling framework able to handle regime switching. For a description of the conditions that give rise to the lack of equivalence between NLPs and complementarity problems, and for some common examples, see Nicholson et al. (1994) and Nagurney et al. (1996a). For a similar discussion in the context of general equilibrium modelling, see Lofgren and Robinson (1999).

[7] There are, of course, techniques for *approximating* regime switching that don't require complementarity.

# 3. MULTI-REGION MODELS[1]

## 3.1 Introduction

Having reviewed, albeit very briefly, the theory underlying optimisation, we now turn to some specific examples. Our purpose here is twofold:

- To practically demonstrate what we've been talking about in the previous section; and
- To develop a realistic but simple model that incorporates tariff-rate quotas.

Whether a problem is formulated as a traditional NLP (of which linear problems are just a special case) or as an MCP, it is necessary to be able to then find a solution to the problem. Only in the most trivial of cases will the method of substitution and elimination, which we used earlier, be helpful. The most popular means of specifying and then solving the types of problems we are concerned with here is the GAMS modelling software. GAMS, which stands for General Algebraic Modelling System, employs a syntax that results in a model representation that is easily understood by humans, as well as computers. It also incorporates (is internally linked to) many well-known commercial solvers, which enables almost any problem type to be solved. Hence, the models we now present will be depicted in GAMS format.

## 3.2 A simple transportation problem

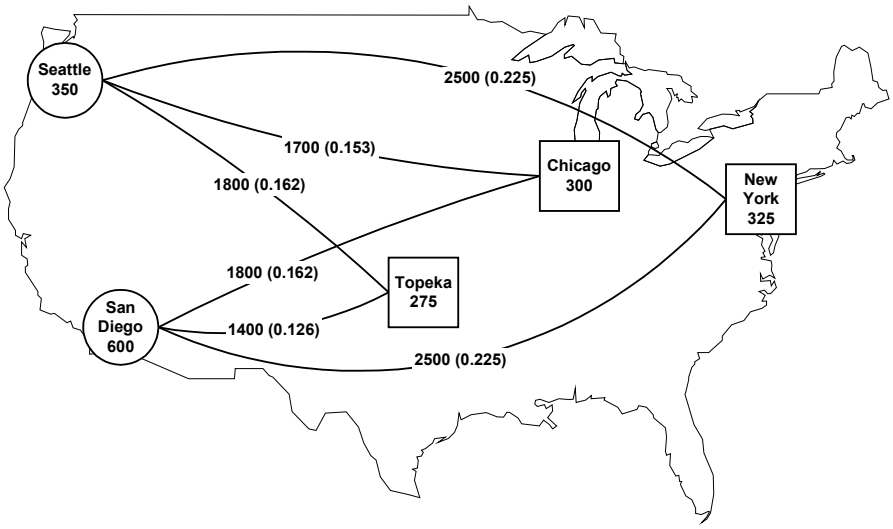
The transportation problem, formulated simultaneously and independently by Kantorovich (1939) and Hitchcock (1941), is the starting point for the partial equilibrium literature. The canonical problem of this genre, popularised by Dantzig (1963) and solved using linear programming (LP) algorithms, is depicted in Figure 1.

Here we have a problem where the objective is to minimise the transportation cost of shipping a single homogeneous good from one set of regions or cities (points in space) to a second set of regions. To make the problem very simple, supply and demand is fixed, i.e. it does not respond to price. In other words, the problem is to satisfy the fixed demands (of 275, 300, and 375 at Topeka, Chicago, and New York, respectively) from the fixed supplies (of 350 and 600 at Seattle and San Diego, respectively), at the least cost.

The transportation cost is specified to be a linear function of distance. Specifically, it is $90 per case per thousand miles. The values on the arcs in Figure 1 denote the distance in miles between each pair of cities, while the number in parentheses is the transportation cost in thousands of dollars per case. Notice that in total there is excess supply; there are 950 cases available for supply, while the total fixed demand is only 900 cases.

[1] We focus on multi-region models in this paper. It should be understood, however, that the same concepts apply to products (or sectors or markets) and time. In other words, the practical analysis of space, form, and time amounts to much the same thing.

***Figure 1 Dantzig's transportation problem***




Source: Dantzig (1963), NZIER.

In keeping with the style of the previous section, we can state this problem algebraically as follows:

$$
\begin{aligned}
&\textit{Minimise} && \sum_i \sum_j c_{ij} x_{ij} \\
&\textit{subject to} && \sum_j x_{ij} \le s_i \qquad \forall i \\
& && \sum_i x_{ij} \ge d_j \qquad \forall j \\
& && x_{ij} \ge 0
\end{aligned}
\qquad \text{(LP 1)}
$$

Alternatively, we can write the model using GAMS (see Figure 2).

Except for the line numbers, which have been added for expositional convenience, the code in Figure 2 is literally a GAMS program. Not only does it include the algebraic description of the model, lines 36 through 40, it also contains the data used to parameterise the model. Specifically, lines 1-3 specify the sets, or indices, on which the problem's parameters, variables, and equations are defined, i.e. set i denotes the supply points (Seattle and San Diego), while set j denotes the three demand points. Lines 5 through 13 declare the two parameters, $s_i$ and $d_j$, and also assigns values to these parameters. It ought to be apparent that $s_i$ denotes the supply quantities of 350 and 600 available at Seattle and San Diego, respectively. Likewise $d_j$ denotes the demand quantities. Lines 15-18 produce a table of distances that is subsequently used in line 23 to calculate the shipping cost associated with each of the six arcs or routes.

Lines 25-29 declare the variables. Note that z is just the objective function value. Also note the command in line 29; it is the GAMS-equivalent of specifying the non-negativity condition on the variable $x_{ij}$.

Lines 31 through 34 declare the equations contained in the model, while lines 36-40 specify the algebra of each equation. In GAMS, =e= means strictly equal to, =l= means less than or equal to, and =g= means greater than or equal to (the ≤, =, and ≥ signs are reserved for use in assignment statements). The command in line 42 simply says take

all of the equations that have been specified and use them to create a model called transport. Line 44 tells GAMS to solve the model called transport using the LP solver. In other words, the user is telling GAMS that this model is an LP problem.

### *Figure 2 GAMS code for Dantzig's transportation problem*

```
1   Sets
2     i     canning plants  / seattle,  san-diego /
3     j     markets         / new-york, chicago, topeka /;
4
5   Parameters
6     s(i)  capacity of plant i in cases
7           /seattle     350
8            san-diego   600 /
9
10    d(j)  demand at market j in cases
11          /new-york    325
12           chicago     300
13           topeka      275 /;
14
15  Table dist(i,j)  distance in thousands of miles
16                  new-york       chicago      topeka
17    seattle          2.5           1.7          1.8
18    san-diego        2.5           1.8          1.4 ;
19
20  Scalar f  freight in dollars per case per thousand miles  /90/;
21
22  Parameter c(i,j)  transport cost in thousands of dollars per case;
23  c(i,j) = f * dist(i,j)/1000;
24
25  Variables
26    x(i,j)  shipment quantities in cases
27    z       total transportation costs in thousands of dollars ;
28
29  Positive Variable x;
30
31  Equations
32    cost        define objective function
33    supply(i)   observe supply limit at plant i
34    demand(j)   satisfy demand at market j ;
35
36  cost..        z =e= sum((i,j), c(i,j)*x(i,j));
37
38  supply(i)..   sum(j, x(i,j)) =l= s(i);
39
40  demand(j)..   sum(i, x(i,j)) =g= d(j);
41
42  Model transport /all/;
43
44  Solve transport using lp minimizing z;
```

Source: GAMS model library (trnsport.gms), www.gams.com.

To recap, what we have specified here is a linear programming problem. An LP is a special case of the NLP class of problems where both the objective function and all of the constraints are linear. Incidentally, if the objective function was quadratic and the constraints were linear, then we would have a problem known as a quadratic programming (QP) problem. LP and QP problems are significant from an algorithmic point of view because specialised solvers are able to solve these problems much more quickly than general NLP solvers can. An NLP may have an objective function that contains nonlinearities of a higher order than quadratic. It may also have linear or nonlinear constraints, or a mixture of both.

Finally, for the sake of completeness, we should point out that a pure transportation problem such as (LP 1) need not be formulated as an LP problem in order to find a

solution; there is another whole class of problems called network problems, for which very fast solution algorithms have been designed.

A solution to this problem will generate the quantities shipped (defined by city of origin and destination). The objective, as already noted, is to minimise the total transportation cost. The only constraints to this simple problem are that shipments must be non-negative (they may be zero along a particular route); the supply cities cannot ship more than they have available to ship; and the total quantity of shipments into a demand city must be at least as great as the fixed amount demanded at that city. Given that each case of the good shipped incurs a positive cost, we would expect that this last constraint will be satisfied with a strict equality, i.e. the quantity shipped into a city will exactly equal the quantity demanded at that city.

The solution is as follows:

- Ship 300 cases from Seattle to Chicago;
- Ship 325 cases from San Diego to New York; and
- Ship 275 cases from San Diego to Topeka.

The total transportation cost, z, will be $153,675. This represents an optimal solution, which means that given the fixed supplies, it is not possible to satisfy the fixed demands at a lower total cost. Note that the excess supply of 50 cases remains at Seattle. This problem actually has more than one optimal solution. The per unit shipping cost is identical along both the Seattle-New York and the San Diego-New York routes. Hence, an equally optimal solution would have been the same as above except Seattle could have shipped 50 cases to New York while San Diego could have shipped 275 cases to New York, instead of 325. In this alternative optimal solution, the total transportation cost would still be $153,675 but the excess 50 cases would be located at San Diego.

Finally, formulating and solving this problem as an LP yields two other pieces of valuable information; the shadow price associated with each of the constraints and the marginal cost associated with each variable. While the solution technique enables this information to be generated, it is important to understand that these values, i.e. the shadow prices and the marginal costs, are not explicit *variables* in the problem being solved.

The shadow prices associated with the two supply constraints are zero. This should not be a surprise as there is excess supply in this problem. The shadow prices represent the value of relaxing the constraint. In other words, how much would the objective function value, z, change if we had one more case at either Seattle or San Diego? The answer is zero because demands are fixed and another case available for supply therefore has no value – anywhere. Recall that the shadow prices are akin to the Lagrange multipliers.

The shadow prices associated with the three demand constraints are 0.225, 0.153, and 0.126 at New York, Chicago, and Topeka, respectively. What does this mean? Consider the New York demand constraint. If New York required 326 cases, i.e. one more than is currently specified, then the objective function value would increase by 0.225 (or, equivalently, the total transportation cost would go up by $225). This, it can be seen, is nothing more than the cost of getting another case from Seattle, the point at which there exists excess supply.

The marginal costs associated with the six variables are zero, except for along the Seattle-Topeka arc where it is 0.036, and along the San Diego-Chicago arc where the

marginal cost is 0.009.[2] What does this mean? It means that it is optimal to send nothing from Seattle to Topeka and from San Diego to Chicago. But if you insisted on sending just one case along these arcs, it would add $36 and $9, respectively, to the optimal (minimum) total transportation cost.

## 3.3 The transportation problem as a Linear Complementarity Problem (LCP)[3]

There is no compelling reason why the transportation problem shown above should be formulated and solved as a complementarity problem. Nevertheless, we will do it so as to explicitly relate the complementarity problem back to a simple LP. We will then move on to a model that includes price responsive supply and demand functions.

Consider the simple transportation problem as specified in (LP 1). As we've just seen, this problem can be solved as a linear program. But let's go back to the theory we discussed earlier. The derivation of the optimality conditions for this problem begins by associating with each constraint a multiplier, alternatively termed a shadow price or dual variable. These multipliers represent the marginal price on changes to the corresponding constraint. Instead of the $\lambda$s we used earlier, let's label the prices on the supply constraint $p^s$ and those on the demand constraint $p^d$.

Intuitively, for each supply node i we have:

$$0 \le p_i^s \quad \text{and} \quad s_i \ge \sum_j x_{ij} \tag{8}$$

In other words, supply prices cannot be negative and the quantity available for supply at a particular location must be greater than or equal to the sum of all shipments out of that location.

Consider what happens when $s_i > \Sigma_j x_{ij}$, i.e. supply is strictly greater than the sum of the shipments. In a competitive setting, no rational person would be willing to pay for more supply at location i; it is already oversupplied. Therefore $p^s$ at that location would be zero. Alternatively, when $s_i = \Sigma_j x_{ij}$, that is the market clears, one might be willing to pay for some additional supply of the good. Therefore $p^s \ge 0$. We can write these two conditions succinctly as:

$$0 \le p_i^s \perp s_i \ge \sum_j x_{ij} \qquad \forall i \tag{9}$$

where the "$\perp$" symbol is understood to mean that at least one of the adjacent inequalities must be satisfied as a strict equality, i.e. either $p^s = 0$ or $s_i = \Sigma_j x_{ij}$. This is nothing more than a formal statement of the complementary slackness result that we saw earlier when presenting the Kuhn-Tucker conditions.

We can go through a similar logic with respect to the demand markets and derive the complementarity relationship:

$$0 \le p_j^d \perp \sum_i x_{ij} \ge d_j \qquad \forall j \tag{10}$$

In other words, if shipments into a demand location were to exceed the quantity demanded, then we'd expect the demand price to be driven down to zero.

Also, from basic intuition, we know that the supply price at i plus the transportation

[2] Not counting the objective function variable, z, whose marginal cost is zero, there are six variables in this model, i.e. i = 2 times j = 3 equals 6, and x is defined on i and j.

[3] This section draws heavily on Ferris and Munson, 2000.

cost $c_{ij}$ from i to j must exceed the market price at j:

$$p_i^s + c_{ij} \geq p_j^d \tag{11}$$

This, as we'll see later in the paper, is just a statement of a simple spatial price equilibrium (SPE) condition. If this condition was not true, then in a competitive market place, another producer could replicate supplier i and thereby increase the supply of the good, which in turn would drive down the market price. This process would continue until the inequality condition (11) was restored. Furthermore, if (11) held with a strict inequality, i.e. the cost of delivery (supply price plus the transportation cost) exceeded the market price, then nothing would be shipped from i to j because doing so would incur a loss. In such a circumstance, it is clear that $x_{ij} = 0$. Therefore,

$$0 \leq x_{ij} \perp p_i^s + c_{ij} \geq p_j^d \qquad \forall i, j \tag{12}$$

We can combine (9), (10), and (12) into a single problem:

$$\begin{aligned} &0 \leq p_i^s \perp s_i \geq \sum_j x_{ij} \qquad \forall i \\ &0 \leq p_j^d \perp \sum_i x_{ij} \geq d_j \qquad \forall j \\ &0 \leq x_{ij} \perp p_i^s + c_{ij} \geq p_j^d \qquad \forall i, j \end{aligned} \tag{LCP 1}$$

(LCP 1) defines a linear complementarity problem that should, by now, be easily recognised as the complementary slackness conditions associated with (LP 1). For linear programs, the complementary slackness conditions are both necessary and sufficient for **x** (a 6 by 1 vector) to be an optimal solution to (LP 1).[4]

Looking a little more carefully at (LCP 1) we can gain further insight into complementarity problems. A solution to (LCP 1) not only tells us how much to send along each route, it also specifies the routes to be used. This property represents the key contribution of a complementarity problem over a system of equations.[5] If we knew a priori which routes to use, we could solve a simple system of equations to find the quantity to ship along each route. However, the key to the modelling power of complementarity is that it chooses which inequalities to satisfy as equalities.

We can therefore exploit this property and generate models with different regimes and let the solution determine which ones are to be active. A frequently used example of this can be found in the economics literature relating to climate change and the atmospheric accumulation of greenhouse gas. It is common in that literature to see modelled a "backstop" technology such as windmills that becomes active once the price of traditional energy sources have reached a certain threshold level, following the introduction of carbon taxes.

In Figure 3 we present the GAMS code to specify and solve (LCP 1). Up until line 25, it is identical to the LP specification shown in Figure 2. The key differences between Figure 2 and Figure 3 are as follows.

---

[4] If the reader is confused at this point, we suggest returning to the section presenting the Kuhn-Tucker conditions. The Kuhn-Tucker conditions and the complementarity problem contain essentially the same information.

[5] As noted earlier, (LP 1), when solved as a linear programming problem, also tells us which routes to use. There is thus no compelling reason to use a complementarity formulation for such a simple problem. But there are many instances when economic problems can't be solved using traditional LP or NLP techniques. In such cases, the available options are to solve a *system of equations* representing the equilibrium conditions for the underlying optimisation problem, or to formulate and solve the problem as a complementarity problem.

### *Figure 3 GAMS code for a linear complementarity problem*

```
1   Sets
2     i     canning plants  / seattle,  san-diego /
3     j     markets         / new-york, chicago, topeka /;
4
5   Parameters
6     s(i)  capacity of plant i in cases
7           /seattle     350
8            san-diego   600 /
9
10    d(j)  demand at market j in cases
11          /new-york    325
12           chicago     300
13           topeka      275 /;
14
15  Table dist(i,j)  distance in thousands of miles
16                  new-york       chicago      topeka
17    seattle          2.5           1.7          1.8
18    san-diego        2.5           1.8          1.4 ;
19
20  Scalar f  freight in dollars per case per thousand miles  /90/;
21
22  Parameter c(i,j)  transport cost in thousands of dollars per case;
23  c(i,j) = f * dist(i,j)/1000;
24
25  Positive Variables
26    x(i,j)      shipment quantities in cases
27    p_supply(i) shadow price at market i
28    p_demand(j) shadow price at market j ;
29
30  Equations
31    supply(i)     observe supply limit at plant i
32    demand(j)     satisfy demand at market j
33    zprofit(i,j)  zero profit condition ;
34
35  supply(i)..     s(i) =g= sum(j, x(i,j));
36
37  demand(j)..     sum(i, x(i,j)) =g= d(j);
38
39  zprofit(i,j)..  p_supply(i) + c(i,j) =g= p_demand(j);
40
41  Model transport /zprofit.x, supply.p_supply, demand.p_demand/;
42
43  Solve transport using mcp;
```

Source: GAMS model library (transmcp.gms), www.gams.com.

The LCP has no objective function so there is no need for an objective function variable, i.e. the variable z in the LP specification. But the shadow prices seen in the solution to the LP model are explicitly included as variables in the LCP. Hence, the variables that we earlier called $p^s$ and $p^d$ are included in the LCP GAMS code as p_supply(i) and p_demand(j), respectively. The spatial price equilibrium condition, equation (11) above, is included in the LCP model as the equation called zprofit(i,j). As described earlier, this is the condition that, in a competitive setting, will cause profits over and above a firm's normal profit to be driven to zero. Hence the name "zero profit condition". One can think of this condition as an arbitrage condition, i.e. the potential for an agent to profitably exploit non-equilibrium situations is sufficient to drive the market back to an equilibrium.

Notice the "Model" statement in line 41. Whereas the LP model simply assigned "all" equations to the model called "Transport", the LCP model requires a different approach. Specifically, the "." takes the place of the "⊥" symbol, which we used earlier

in (LCP 1). So, line 41 specifies that the variable x is complementary to the equation called zprofit. Likewise, the variable called p_supply is complementary to the equation called supply and the variable called p_demand is specified to be complementary to the equation called demand. It is important to note that the modeller must explicitly specify the complementarity pairings in order for the solver to exploit this information. Simply formulating a model that contains arbitrage conditions, such as (11), in an NLP setting is not the same as exploiting complementarity.

Finally, the "Solve" statement in line 43 is differs in two ways from that in the LP model. First, we need to tell GAMS that this is an MCP class of model, and not an LP. And as a consequence of this, there is no need to specify an objective function value to be minimised (or maximised).

A solution to this problem will generate the quantities shipped (defined by location of origin and destination), and supply and demand prices, i.e. both prices *and* quantities are variables, unlike in (LP 1). Both price and quantity variables may be zero, i.e. the model endogenously selects the appropriate regime that satisfies the model and its constraints.

The economic question contained herein is what quantity should be shipped between each supply and demand point so as to minimise the overall transportation cost? The answer to this question describes the regime determined by the model's solution. For markets, i.e. combinations of supply and demand points, the regime is either one of full utilisation with a positive market clearing price, or excess supply with a zero price. For the arcs, i.e. the transportation flows, the regime is either one of active links associated with a positive shipment, or inactive links and a zero flow. The solution may be viewed as a market equilibrium, albeit subject to the restrictive assumption of fixed, i.e. non-price responsive, supply and demand quantities.

Not surprisingly, the solution to the LCP is identical to the solution of the LP problem. To reiterate though, the prices are explicit variables in the LCP formulation, whereas the LP model yields prices only implicitly. This points to some of the flexibility available from the LCP formulation compared to the LP model. For instance, in the LCP (or the MCP) setting, it is a straightforward matter to directly simulate policies that operate on prices, e.g. agricultural support prices.

## 3.4 Adding price responsive behaviour

The spatial price equilibrium (SPE) model is something of a workhorse in trade and interregional analysis. A simple formulation of an SPE model is just the transportation problem with its fixed supplies and demands replaced with price responsive supply and demand functions. In this section, we describe a little of the background to the SPE model, and then amend (LCP 1) so that it incorporates price responsive behaviour. To keep things uncluttered we will assume that quantity is a simple function of own price, i.e. all conceivable cross-price terms are zero. Once we have explored the specification of the SPE model formulated as a complementarity problem, we will be in a position to then add tariff-rate quotas to the model.

Enke (1951) and Samuelson (1952) were the first to extend the transportation model by introducing price responsive regional supply and demand functions. Samuelson's formulation shows that the problem of maximising "net social payoff" (the sum of consumers' and producers' surpluses in each region less transportation costs) subject to regional commodity balance equations generates a set of optimality conditions that

define an equilibrium in each regional market.[6] Given the significance of Samuelson's contribution, it is worth dwelling on this for a moment.

Imagine a three region model where each region both supplies and demands a single good. Further imagine that the cost structures and consumer preferences are sufficiently different in each region that they engage in trade in order to maximise welfare. In a graphical sense, it is easy enough to picture the area denoting producer and consumer surplus in each region. The transportation cost is simply the quantity traded between each pair of regions multiplied by the appropriate transportation cost.

Samuelson's innovation was that if we simply set out to maximise the sum, over all regions, of the producers' and consumers' surplus, less the total transportation cost, and observed the supply and demand constraints, then the resulting solution to such an optimisation problem would, in fact, be the equilibrium market solution. That is, the solution would yield the quantity that each region would supply and demand, the quantity that would be traded between regions, and the supply and demand prices in each region could be gleaned from the solution as the shadow prices associated with the supply and demand constraints.

Takayama and Judge (1964) operationalised Samuelson's approach by showing that if the supply and demand functions were linear, then the resulting optimisation problem was a quadratic programming problem (i.e. quadratic objective function with linear constraints), which could be solved quite readily with available QP solvers.[7] Takayama and Judge also extended the model to multiple products incorporating cross-price terms in the supply and demand functions. The work of Samuelson, and subsequently Takayama and Judge, spawned a great deal of empirical modelling, especially in the area of trade. Even today, many trade models are constructed in the tradition of the Samuelson-Takayama-Judge (STJ) genre. A very accessible and graphical exposition of the STJ model can be found in Martin (1981).

Most spatial equilibrium models of the STJ type are formulated in the quantity domain. This means that the supply and demand curves are inverted, the primal variables in the model are quantities, and prices are read from the solution as shadow prices, i.e. as we saw earlier. Alternatively, but less commonly in the case of trade models, the "dual" problem could be formulated and solved, whereby the problem is solved in the price domain. In other words, the variables in the model are prices and the shadow values are the quantities. Either way, LP, QP, and more general NLP solution techniques require that the demand functions be symmetric (see Nicholson et al., 1994). This shortcoming was overcome with linear complementarity techniques (see the primal-dual formulations in Takayama and Judge, 1971). However, up until the 1990s, when complementarity solvers became commercially available, such primal-dual problems were usually configured such that they could be "forced" into conventional NLP solvers. Hence, the objective function, even if it was vacuous, still had to observe the symmetry condition.[8]

---

[6] We hasten to point out that Samuelson warned of the problems associated with using his result to make inferences about welfare. Hence his term "net social payoff", which explicitly excludes a reference to welfare. The literature would seem to suggest, however, that Samuelson's cautionary note was almost immediately ignored. See his 1952 paper for further details.

[7] In fact, the objective function in the Takayama and Judge QP formulation embodies the integral functions of the inverse linear demand and supply functions. It calculates the area under the demand curve between the origin and the optimal demand quantity (a decision variable in the model), less the area under the supply curve between the origin and the optimal supply quantity, less the transportation costs. The result is the sum over all regions of producers' and consumer' surplus.

[8] Technically speaking, NLP solution techniques require that the Jacobian matrix, the matrix of first

Finally, it should be noted that a number of policy instruments can be quite readily modelled in a simple SPE model. Indeed, the trade literature is full of examples. For example, the transportation cost component can be modified to include per unit tariffs, taxes, and subsidies; import quotas can be introduced as upper bounds on shipment variables; and even ad valorem tariffs can be modelled, so long as they are non-discriminatory, by modifying the slope parameters of the demand functions. But there are many policy instruments that the conventional SPE model is unable to accommodate; discriminatory ad valorem tariffs, for instance.

### 3.4.1 The SPE model as an MCP

We now turn to the task of amending (LCP 1) to create a nonlinear complementarity problem. We will refer to the resulting model as an MCP, even though it does not contain a mixture of equalities and inequalities (see Figure 4, the model contains only inequalities). Nevertheless, unlike the linear complementarity problem in Figure 3, our SPE model is now a *non*linear problem due to the functional form we have chosen for the supply and demand functions, i.e. they are constant elasticity functions.

The GAMS code seen in Figure 4 should by now seem quite familiar. The first 30 lines are identical to the LCP model in Figure 3, except that we have introduced two new parameters; eta and sigma, the elasticities of supply and demand, respectively.

We assume for simplicity that supply prices are unitary (i.e. equal to 1). In order to specify the isoelastic functions, we need to compute share parameters based on the base case, or reference data. Hence, lines 33 through 38 declare the parameters to do this. Careful inspection will reveal that the reference demand prices are just the supply prices plus the lowest transport cost to each demand city. Because supply prices are 1, line 42 simply sets the supply function share parameter to be equal to the supply quantity. Similarly, line 44 computes the demand share parameters using the reference demand prices, pbar. Note that two asterisks is the GAMS way of denoting exponentiation (e.g. line 44).

Before we take a look at the variables and the equations, we should reiterate that the reference data we have specified here is purely fictional. Nevertheless, the GAMS code in Figure 4 can be used as a template for specifying a realistic model where supply and demand prices and quantities, and transportation costs are all observed, and the elasticities are econometrically estimated. The set-based structure of GAMS also means that the model is highly scalable to any number of regions (and commodities, for that matter).

partial derivatives, be symmetric. The model is then said to be *integrable*. We should also point out that complementarity techniques are not the only way out of this "requirement for integrability" dilemma. Fixed-point algorithms, for example, received a lot of attention in the1960s and 1970s. But while theoretically elegant, they have turned out to be rather slow and cumbersome in applied modelling situations. Variational inequalities, closely related to the complementarity problem, may also be used (see Nagurney et al., 1996b).

### *Figure 4 GAMS code for a nonlinear complementarity problem*

```
*    Lines 1 through 13 the same as in Figure 3.
14
15    eta(i)   price elasticity of supply
16            /seattle    1.0
17             san-diego  1.0 /
18
19    sigma(j) price elasticity of demand
20            /new-york   1.5
21             chicago    1.2
22             topeka     2.0 /;
23
24  Table dist(i,j)  distance in thousands of miles
25                  new-york       chicago      topeka
26    seattle          2.5           1.7          1.8
27    san-diego        2.5           1.8          1.4  ;
28
29  Scalar f   freight in dollars per case per thousand miles  /90/;
30
31  Parameters
32    c(i,j)   transport cost in thousands of dollars per case
33    alpha(i) supply function share coefficient
34    beta(j)  demand function share coefficient
35    pbar(j)  reference price at demand city j (suppy price = 1)
36            /new-york  1.225
37             chicago   1.153
38             topeka    1.126 /;
39
40  c(i,j) = f * dist(i,j)/1000 ;
41
42  alpha(i) = s(i);
43
44  beta(j) = d(j)*pbar(j)**sigma(j);
45
46  Positive variables
47    x(i,j)       shipment quantities in cases
48    p_supply(i)  shadow price at supply market i
49    p_demand(j)  shadow price at demand market j ;
50
51  Equations
52    supply(i)    supply limit at plant i
53    demand(j)    demand constraint at market j
54    zprofit(i,j) zero profit conditions   ;
55
56  supply(i)..   alpha(i)*p_supply(i)**eta(i) =g= sum(j, x(i,j));
57
58  demand(j)..   sum(i, x(i,j)) =g= beta(j)*p_demand(j)**(-sigma(j));
59
60  zprofit(i,j).. p_supply(i) + c(i,j) =g= p_demand(j);
61
62  Model transport /zprofit.x, supply.p_supply, demand.p_demand/;
63
64  p_demand.l(j) = pbar(j);
65
66  Solve transport using mcp;
```

Source: Rutherford (1995).

The variables and the equations declared in this model, i.e. lines 46 through 54, are the same as we had before in (LCP 1). The difference now is in the specification of those

equations. Consider the left-hand side of the supply equation, i.e. line 56. Whereas before we had $s_i$ on the left, i.e. the fixed supply, we now have a function of own-price, p_supply, that will evaluate to yield the supply quantity. Similarly, the right-hand side of the demand equation is the demand function. Notice the negative sign on the elasticity term to give a downward sloping function. The zero profit condition and the complementarity pairings in the model statement remain unchanged from what they were earlier.

Finally, line 64 assigns an initial (strictly positive) value to the demand price variables. If we didn't do this, then the exponentiation in line 58 would yield a numerical error at the start of the solution process. That is, at this point, prior to the model being solved, the level of the variable p_demand is zero, and zero raised to a negative exponent is undefined.

As we noted earlier, there is really no need to formulate this model as a complementarity problem as it is perfectly able to be solved using conventional NLP techniques. But consider the case of discriminatory ad valorem tariffs or taxes. That is, imagine that each of the three demand cities was to charge each of the two supply cities a *different* ad valorem tariff. Such a problem provides an example of where the formulation of a market equilibrium is not straightforward.[9] No single optimisation problem characterises the equilibrium because *integrability* has been destroyed, i.e. the supply price is a non-unitary multiple of the marginal cost of supply. But there does exist a unique MCP which precisely characterises the equilibrium.

Consider line 60, the zero profit condition. If we had a parameter called $t_{ij}$, denoting our asymmetric ad valorem tariff or tax rate, then it would be a simple matter to incorporate it into the model by modifying the zero profit condition as follows:

```
zprofit(i,j).. (p_supply(i,j) + c(i,j))*(1 + t(i,j)) =g= p_demand(j);
```

There are many other examples in economics where this modelling difficulty arises. Asymmetric demand specifications, regime switching, threshold effects, switching sides of the market (i.e. from being an exporter, say, to being an importer),[10] and tariff-rate quotas are just a few. Exploiting complementarity is a convenient means of resolving the difficulty. Specific taxes (or tariffs or subsidies) do not cause this problem as they can easily be added to the per unit transport cost coefficients.

# 4. TARIFF-RATE QUOTAS

We now return to the topic implied by the title of this paper. Thus far we have developed the intuition underlying the complementarity problem. Moreover, we have shown how the popular spatial price equilibrium model can be formulated as a mixed

[9] An equilibrium to such a problem could be computed by iteratively solving a sequence of NLPs, but this is clumsy and inefficient.

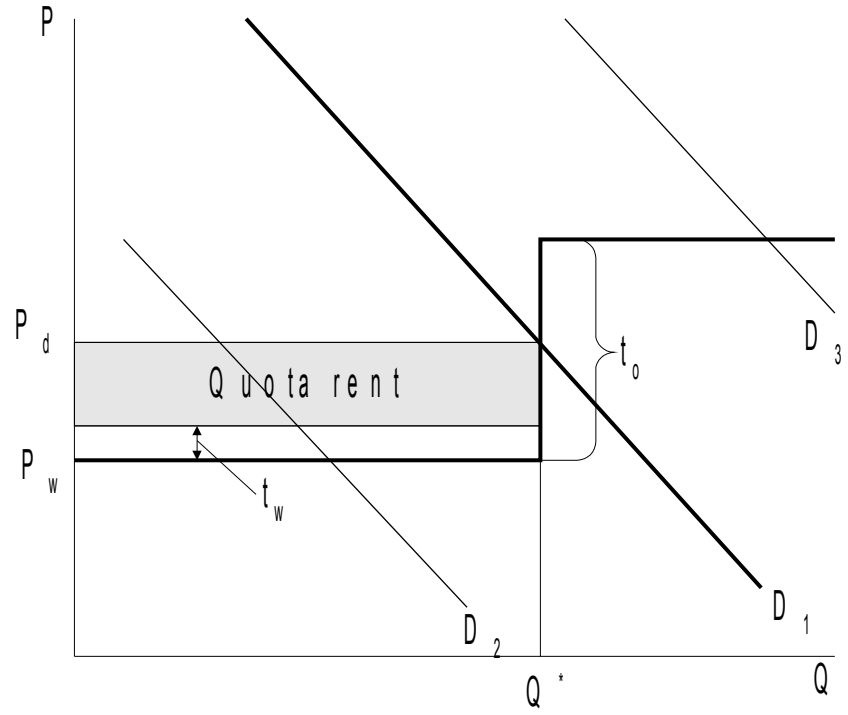
[10] See Anania and McCalla (1991) and Bishop et al. (1994).

complementarity problem, and we have even presented the GAMS code, which can be used as a template for building a realistic model.[1] In this section of the paper we discuss some aspects of tariff-rate quotas (TRQs) that make them difficult to incorporate into models using conventional optimisation techniques. In the following and final section, we bring everything together and present an SPE model complete with TRQs.

## 4.1 The basic tariff-rate quota

As the name suggests, a TRQ embodies both a quantitative restriction in the form of a quota, and a price instrument in the form of a tariff (which may be ad valorem or specific).[2] Figure 5 presents a very simplified depiction of a TRQ.

***Figure 5 A simple TRQ***




Source: NZIER

One can think of this as being a small country supplying a single good to a country that imposes the TRQ. The supply curve is the bold line that begins horizontal at $P_w$, the world price. It then turns vertical at $Q^*$, the import quota quantity, and beomes horizontal again at $P_w + t_o$, where $t_o$ denotes the over-quota tariff rate. For the moment, imagine the importing country's demand curve is that given by $D_1$. Whilst the diagram is clearly not to scale, imagine, for the sake of simplicity, that the supply curve belongs

[1] As we noted at the outset of this paper, we are unable to claim credit for much of what has been presented here. The GAMS templates are no exception as they are taken, almost as originally presented, from other sources.

[2] Throughout this paper we have focused on ad valorem tariffs. We will continue that focus but would point out that the MCP formulation also handles specific tariffs, as well as TRQs that might contain both specific and ad valorem tariffs.

to a single supplier while the demand curve represents the sum of all demand from imported and domestically produced sources.

In this simple case, there are two tariff rates associated with this TRQ. The within-quota tariff, $t_w$, applies to all imports up to the quota quantity, $Q^*$. Above the quota, the over-quota (sometimes called the out-of-quota) tariff applies. But significantly, the TRQ places no quantitative restriction on the import volume that might occur above $Q^*$. However, in practice, the over-quota tariff rates in agriculture are typically so high, they prohibit any trade from taking place above the quota.

But it is not just the tariff rates and quota levels that determine the quantity of trade. The nature of demand in the importing country also plays a role. Consider, the demand curve depicted by $D_1$. In this case, imports are exactly equal to $Q^*$ and a quota rent has been created (the shaded rectangle). The domestic price, $P_d$, is determined in the normal fashion by observing where demand intersects with supply. The government in the importing country collects revenue equal to $t_w$ times $Q^*$.

But what if $D_2$ was the relevant demand curve? Clearly in this case the quota would not be binding and no rent is created. If this situation prevailed, then there is nothing to be gained from trade liberalisation that saw either the quota quantity increased and/or the over-quota tariff reduced. A small increase in trade would be observed if, when $D_2$ was relevant, the within-quota tariff was reduced.

The more interesting situation is when the demand schedule lies somewhere near $D_1$ or $D_3$. When $D_1$ is the relevant demand curve, there is an interesting trade-off to be made, by the supplying country, between the benefits of greater market access, i.e. an increased quota quantity, versus a diminished quota rent. Consider the situation where $Q^*$ is increased up until the point where the world price line intersects with $D_1$.[3] Under this scenario, the quota rent would diminish entirely but this would be offset to some extent by increased exports. Which scenario should the exporting country prefer? The answer is it depends. One needs a model with endogenous regime switching and an endogenously determined quota rental variable in order to be able to ascertain which situation is preferred. It may turn out that the exporting country's favoured position is a *decline* in $Q^*$, which would see the quota rental value increase.

Now consider the case when $D_3$ is relevant. In this situation a judicious *increase* in $Q^*$ is likely to be beneficial to the exporting country, i.e. both the quota rent and the volume of trade increase. On the other hand, the merits of a decrease in $t_o$ may be indeterminate, i.e. the trade volume would increase, but the tariff rental value *may* decline.

To muddy the waters even further, it is usually the case that in reality, TRQs are more complex than that depicted in Figure 5. For example, there may exist several tiers to the schedule. The tariffs and the quotas may be assigned on a bilateral basis as well as on a multilateral basis. For reasons that have to do with historical trade patterns, New Zealand has a disproportionate share of preferential access arrangements in some markets. These may take the form of favourable tariff rates and/or exclusive quota rights. Tariffs may be defined either on a specific basis or on an ad valorem basis.

All of these factors give rise to an interesting problem when deciding upon a negotiating stance to adopt. The answer as to which regime New Zealand favours is likely to be different from one case to the next. But New Zealand may not be in a position to pick and choose. It is therefore imperative that New Zealand negotiators understand what the outcomes of policy proposals are, as they are proposed and

[3] Actually, the relevant line to consider is the line denoted by $P_w + t_w$.

before they are negotiated. In the next section we discuss a modelling framework that is able to shed some light on these questions.

# 5. BRINGING IT ALL TOGETHER

We now present a 3-region SPE model formulated as an MCP, and which incorporates tariff-rate quotas (see Figure 6). The core model specification is changed slightly from the SPE model we saw in Figure 4, with all unnecessary detail stripped away. In fact, we don't even assign data values to the parameter symbols. Rather, we conduct the entire discussion in terms of the symbol names. To keep things uncomplicated we show only bilateral tariff-rate quotas. It is a straightforward modification, however, to integrate multilateral TRQs into the same model. For an example of a trade model formulated as a complementarity problem that embodies a wide range of price- and quantity-based policy instruments, as well as multiple products, see Bishop and Nicholson (2002).

***Figure 6 GAMS code for an MCP with TRQs***

```
1    SETS
2      i   regions                      /r1,  r2,  r3 /
3      ql  quota levels (break points)  /ql1, ql2, ql3 /;
4
5    ALIAS (i,j);
6
7    PARAMETERS
8      s0(i)          reference supply quantity in region i
9      d0(j)          reference demand quantity in region j
10     x0(i,j)        reference biltaeral trade quantity
11     eta(i)         price elasticity of supply
12     sigma(j)       price elasticity of demand
13     qlvl(i,j,ql)   bilateral quota levels
14     t(i,j,ql)      bilateral tariff rates ;
15
16   ... Read in data here ...
17
18   POSITIVE VARIABLES
19     X(i,j,ql)       shipment of product from region i to region j
20     P(i)            regional price
21     QR(i,j,ql)      quota rent (price per unit)  ;
22
23   EQUATIONS
24     ZPROFIT(i,j,ql)  zero profit conditions
25     MARKET(i)        domestic market clearing constraint
26     QUOTA(i,j,ql)    limits on shipments by tariff class ;
27
28   market(i)..
29     s0(i)*P(i)**eta(i) + sum((j,ql), X(j,i,ql)) =e=
30     d0(i)*P(i)**(-sigma(i)) + sum((j,ql), X(i,j,ql));
31
32   zprofit(i,j,ql)$(NOT sameas(i,j))..
33     P(i)*(1 + QR(i,j,ql) + t(i,j,ql)) =g= P(j);
34
35   quota(i,j,ql)$(NOT sameas(i,j)).. qlvl(i,j,ql) =g= X(i,j,ql);
36
37   MODEL trq /market.p, zprofit.x, quota.qr /;
38
39   p.l(i) = 1;
40
41   SOLVE trq using mcp;
```

Source: NZIER

Immediately noticeable is that the declaration of the sets, lines 1 through 3, is slightly different than before. We have dropped the U.S. cities and gone with three generic regions, denoted r1, r2, and r3. As we'll see in moment, each region is both a supplier and a demander of the single good. Hence, this model has three domestic markets that are linked through their ability to trade with one another. Notice the "alias" statement in line 5. It assigns the elements of set i to a set called j (i.e., i and j each have the same elements).

The second set we define, ql, specifies that there are three steps or break points in the quota schedule. This in turn implies that there are three levels in the tariff schedule. To avoid confusion, we will be quite explicit about how this specification is to be interpreted. We could imagine that ql1, the first element of set ql has a quota quantity of, say, 1000 tonnes associated with it. All imports up to that point might attract a tariff of 10%. The second element, ql2, might relate to a quota quantity of 3000 tonnes, which attracts a tariff of 50%. Finally, the third element, ql3, might be infinity, and shipments occurring in this band might attract a tariff rate of 250%. (We would point out that the TRQ instrument normally has an infinite quantity associated with the upper tier, although from a modelling point of view, this is not required.) To recap, our imaginary tariff-rate quota schedule with three tiers operates as follows. Imports up to 1000 tonnes attract a 10% tariff. Imports between 1000 and 4000 tonnes, i.e. a 3000 tonne quota at the second tier, attract a 50% tariff. And all imports over 4000 tonnes get charged a tariff of 250%.

In lines 7 through 14 we declare some parameters. All but the last two should be familiar from the models seen earlier. The reference or benchmark quantity of bilateral trade, x0, is not necessary to specify the model, per se, although such data may be used to validate the model's ability to replicate the benchmark set of data. In any event, it is at line 16 that one would ordinarily assign values to all of these parameters. Alternatively, GAMS could be instructed at this point to read the necessary data from an external file, such as a spreadsheet.

The parameters qlvl and t (lines 13 and 14) are new to this model, and their purpose should be self-evident. These two parameters define the quota levels and tariff rates in the manner we have just explained above. Notice that each of these parameters are defined on sets i and j (i.e. on an origin-destination basis) as well as on ql. This should reiterate the point that these parameters are defined bilaterally.

If one were to also include a multilateral TRQ, it would be accomplished by creating an additional parameter for the multilateral quota levels, say, qlvlm(j,ql), i.e. it would not be defined *on* i as it would apply *to* all i. A multilateral tariff schedule can be accommodated by assigning the appropriate values to the bilateral tariff parameter. Alternatively, one could create a specific multilateral tariff parameter, even though it is unnecessary. Obviously, care needs to be exercised in defining a model with both multilateral and bilateral TRQs. It would make no sense, for example, for a multilateral quota quantity in region j to exceed the sum of all bilateral TRQs emanating from j and applying to all other regions. Finally, the addition of a multilateral TRQ would require an additional variable and complementary constraint; i.e. a multilateral quota rent variable and a multilateral quota constraint.

There are only three variables in this model. The shipment variable, x, we have seen before. But notice that it is now defined on ql as well as i and j. One can think of each origin-destination route as being a road divided into 3 lanes, one for each level of the TRQ schedule. However, the price variable, p, is a significant change from earlier. We now have just one price per region, whereas previously we had a supply price and a

demand price. This comes about because we have removed all of the *intra*-regional price wedges, i.e. there are no transportation costs in this model and the quotas and tariffs don't apply on intra-regional shipments. Hence, the equilibrium supply price is identical to the demand price in each region.

The final variable (line 21) is the quota rent variable, qr. It can be interpreted as the price of a unit of quota rent. Notice that it is defined on both i and j because it applies to quota rents on a bilateral basis. Note too that it is defined on set ql. It should already be apparent that there are potentially three "quota rent" rectangles of the kind seen in Figure 5. The value of each quota rental is just the relevant qr variable multiplied by the relevant quota quantity, qlvl. The key point to note about the quota rent variable is that it is endogenous – a solution to the model will yield the level of qr. It is not the case that the modeller must assign a value to the quota rent before using the model to undertake experiments. Clearly, as was discussed in the previous section, the level of the quota rent variable will be determined by the interplay of a number of factors. Moreover, it will be consistent with the equilibrium outcome.

There are three sets of equations in this simple model. The first, called market, specifies the condition that ensures each market clears and that trade flows balance.[1] It says that the quantity supplied plus the sum of all shipments *from* a region (including intra-regional flows) must be equal to the sum of all shipments *into* a region (including intra-regional flows) plus the quantity demanded. Careful inspection of the summation of the x variable will reveal that the order of the indexes, i and j, is reversed on the right-hand side of the equation from what it is on the left. The supply and demand functions are specified such that prices are equal to 1 in the base case. Hence, there is no need to compute and use the share coefficient terms, alpha and beta, as was the case in Figure 4. But we would stress that this is just for convenience; the functions could be calibrated to any consistent price and quantity levels. As before, the supply and demand relationships are isoelastic functions of own price.

The zero profit condition for this model is quite straightforward. In essence, it says the price in region i multiplied by one plus the quota rent variable plus the tariff rate, is greater than or equal to the price in region j. There are two points to note about this condition. First, it is not defined when i equals j (see the statement that says "not samesas (i,j)". Second, a zero profit condition is defined for each level of the quota schedule, i.e. set ql, as well as each $(i\text{-}j)^{th}$ route, so long as i is not the same as j. Because the benchmark prices are normalised to one in this model, the quota rent variable appears to enter the zero profit condition as a rate, just like the tariff rate. Once again, this would not be the case if prices were modelled at their observed levels.
The final equation is the quota constraint. It simply says that for each i-j-ql arc, where i is not equal to j, the quota level must be greater than or equal to the shipment quantity.

As before, the model statement specifies the complementarity pairing of variables with equations. Also, we set the initial price level to be one in order to avoid undefined exponentiation. The final statement tells GAMS to solve the model called trq, while making sure to treat it as an MCP formulation.

# 6. CONCLUDING REMARKS

Although it took a while to get there, we have now presented a comprehensive

[1] The term "trade flows" is used rather loosely here. It includes intra- as well as interregional flows.

template of an SPE trade model, formulated as an MCP, that can easily be scaled up to realistic dimensions. We finish this paper with a few comments on how to use the model to conduct experiments.

A typical experiment would entail decreasing tariff rates and observing what happens to production, demand, trade volumes, prices, and quota rents. For example, an across the board tariff cut of 30% could be modelled by adding the statement `t(i,j,ql) = 0.3*t(i,j,ql);` immediately after the solve statement in line 41, followed by a second solve statement. The results of the simulation (the second solve) can then be compared with the benchmark case (the first solve). One could get more specific, however, and conduct experiments where only the tariffs on certain routes and/or certain levels of the tariff schedule are modified.

Similarly, one could simulate market access scenarios by increasing quota levels.

Finally, a likely policy scenario would involve increased quota levels in conjunction with decreasing tariff rates.

# 7. REFERENCES

Anania, G. and A.F. McCalla (1991). "Does arbitraging matter? Spatial trade models and discriminatory trade policies." *American Journal of Agricultural Economics* 73:103-17.

Bishop, P.M. and C.F. Nicholson (2002). *A mixed complementarity model of world dairy trade.* Forthcoming staff paper, Department of Applied Economics and Management, Cornell University, Ithaca NY.

Bishop, P.M., J.E. Pratt, and A.M. Novakovic (1994). *Using a joint-input, multi-product formulation to improve spatial price equilibrium models.* Staff paper 94-06, Department of Agricultural, Resource, and Managerial Economics, Cornell University. (Presented at the "New Dimensions in North American-European Agricultural Trade Relations" conference, Calabria, Italy, June 20-23, 1993.)

Brooke, A., D. Kendrick, A. Meeraus, and R. Raman (1998). *GAMS: a user's guide*. GAMS Development Corporation, Washington, D.C.

Chiang, A. (1984). *Fundamental methods of mathematical economics* (3rd ed.), McGraw-Hill.

Cottle, R., J.S. Pang, and R.E. Stone (1992). *The linear complementarity problem*. Academic Press, San Diego.

Dantzig, G.B. (1963). *Linear programming and extensions*. Princeton University Press, Princeton, NJ.

Dirske, S. and M.C. Ferris (1993). *The PATH solver: a non-monotone stabilisation scheme for mixed complementarity problems.* Technical report 1179, Computer Science Department, University of Wisconsin, Madison WI.

Enke, S. (1951). "Equilibrium among spatially separated markets: solution by electric analogue." *Econometrica* 19:40-47.

Ferris, M.C., and T.S. Munson (2000). *GAMS/PATH user guide, version 4.3*. GAMS Development Corporation, Washington, D.C.

Harker, P. and J.S. Pang (1990). "Finite-dimensional variational inequality and

nonlinear complementarity problems; a survey of theory, algorithms and applications" *Mathematical Programming B* 38:161-190.

Hartman, P. and G. Stampacchia (1966). "On some nonlinear elliptic differential functional equations." *Acta Math*. 115:153-188.

Hitchcock, F.L. (1941). "The distribution of a product from several sources to numerous locations" *Journal of Mathematical Physics* 20:225-230.

Kantorovich, L. (1939). "Mathematical methods in the organisation and planning of production," Publication house of the Leningrad State University. Translated in *Management Science* 6:366-422.

Kuhn, H.W. and A.W. Tucker (1951). "Nonlinear programming," in J. Neyman (ed.), *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, University of California Press, Berkeley CA, pp. 481-92.

Lemke, C.E. and J.T. Howson (1964). "Equilibrium points of bimatrix games." *SIAM Review* 12:413-23.

Lofgren, H. and S. Robinson (1999). *Spatial networks in multi-region computable general equilibrium models.* TMD discussion paper no. 35, Trade and Macroeconomics Division, IFPRI, Washington, D.C.

Martin, L.J. (1981). "Quadratic single and multi-commodity models of spatial equilibrium: a simplified exposition." *Canadian Journal of Agricultural Economics* 29(1):21-48.

Nagurney, A., C.F. Nicholson, and P.M. Bishop (1996a). "Massively parallel computation of large-scale spatial price equilibrium models with discriminatory ad valorem tariffs." *The Annals of Operations Research* 68:281-300, special issue on computational economics.

_____ (1996b). "Spatial price equilibrium models with discriminatory ad valorem tariffs: formulation and comparative computation using variational inequalities" in van den Bergh, J.C.J.M., P. Nijkamp, and P. Rietveld (eds.), *Recent advances in spatial equilibrium modelling: methodology and application*s. New York: Springer.

Nicholson, C.F., P.M. Bishop, and A. Nagurney (1994). "Using variational inequalities to solve spatial price equilibrium models with ad valorem tariffs and activity analysis." Working paper 94-14, Department of Agricultural, Resource, and Managerial Economics, Cornell University, Ithaca NY.

Rutherford, T.F. (1993). "MILES: a Mixed Inequality and nonLinear Equation Solver." Working paper, Department of Economics, University of Colorado, Boulder CO.

_____ (1995). "Extensions of GAMS for complementarity problems arising in applied economic analysis." *Journal of Economic Dynamics and Control* 19:1299-1324.

Samuelson, P. (1952). "Spatial price equilibrium and linear programming." *The American Economic Review* 42:283-303.

Takayama, T. and G.G. Judge (1964). "Equilibrium among spatially separated markets: a reformulation." *Econometrica* 32:510-524.

_____ (1971). *Spatial and temporal price and allocation models.* Amsterdam: North-Holland.