18

AIRPORT INVESTMENT AND PRICING POLICIES

A. Cubukgil,* S. Borins** and M. Hoen*
November 1991

1. INTRODUCTION

Planning and managing Canada's airports are two of Transport Canada's most important responsibilities. Airports owned by the federal government have a replacement value, excluding land, of over $10 billion. In the 1988-89 fiscal year, the Airports Authority Group of Transport Canada spent $247 million on capital investments for the expansion, restoration and upgrading of airport facilities. Pearson International Airport's Terminal 3, funded by the private sector, cost $550 million. The three proposed additional runways at Pearson would cost another $469 million. Investments of this magnitude require wise and informed decisions.

This study examines airport planning from an economic perspective, dealing primarily with policies regarding efficient allocation of resources at Canadian airports. One aspect of resource allocation is the use of existing capacity, examined in this study as the "pricing problem." The second aspect is the provision of additional capacity or building of new facilities, which is addressed as the "investment problem." In theory, these aspects are two sides of the same coin, but the distinction between them is important from a more practical policy perspective.

* Transmode Consultants Inc.
** University of Toronto.

The report starts with a background section on the evolution of Canadian airport planning, providing a brief institutional history, a review of the prevalent physical planning paradigm, and a discussion of the three main pillars of the federal government's current airport policy framework: devolution, cost recovery and environmental review. The third section of the report deals with the investment problem, first from a theoretical perspective, followed by two recent case studies (cost-benefit analysis of airside capacity expansion proposals for Toronto and Vancouver international airports). The fourth section deals with the pricing problem, providing a theoretical overview, followed by an assessment of alternative pricing policies and current practices. The conclusion of the report, recommends the incorporation of a continuous cost-benefit analysis framework into the airport planning process, to deal not only with pricing and investment policies but also to mitigate external impacts.

2. Evolution of Canadian Airport Planning

To set the institutional and policy context within which Canadian airports are operated, the following section provides a brief institutional history. Then a more detailed examination of the airport planning process in Canada deals with the physical planning heritage of a system which has not helped to promote economic efficiency. Finally, a discussion of current policy includes the three principal aspects of federal airport policy: devolution, cost recovery and environmental review.

2.1 INSTITUTIONAL BACKGROUND

From the 1960s to the mid-1980s, airports were the responsibility of the Canadian Air Transportation Administration (CATA), which was part of Transport Canada. Airport investments were funded out of CATA's capital budget, which was determined in negotiations between it and the Treasury Board. Airports raised some revenues, primarily through landing fees and rentals of terminal space, but in only the largest airports did revenues cover operating and maintenance costs.

There was no requirement for airports to be self-financing, and they were not. Revenues went to the Consolidated Revenue Fund, and expenditures were paid out of the Fund, but the two were not linked. The costs of the

Air Navigation System, kept separate from those of the airports, were also funded out of the Consolidated Revenue Fund.

The dominant professional groups in CATA were people with hands-on aviation industry experience (for example, former pilots and controllers) and, on the planning and construction side, engineers. Their approach to planning was to set appropriate physical standards for the various facilities (runways, roads, terminals) that make up an airport system and, when these standards were exceeded, to add capacity. These groups did not think in terms of economic concepts, such as use of the price system to ration capacity, or the time value of money.

The demand for air travel grew very rapidly in the 1960s, for a number of reasons. The introduction of jet aircraft dramatically lowered the cost of air travel, and equally dramatically shortened travel times. Real income was increasing. Charter airlines came into being and, by achieving high load factors, were able to lower prices even more. The airport planners and managers in CATA were unprepared for this air travel boom and, by the mid-1960s, major Canadian airports were operating beyond their capacity, with passengers suffering through long queues, particularly in the terminals.

An airport system managed by a government department, accountable to Parliament through the Minister of Transport, must be responsive to the pressures of its stakeholders. In addition, the Department is influenced by the main priorities of government as enunciated by the central agencies. The congestion at major airports meant that airport users — airlines, passengers and general aviation — were complaining. This was an embarrassment to both the Minister and the public servants. On the other hand, with a strong economy, the government's tax revenues were growing and there was no deficit problem. Transport Canada made a strong case to the Treasury Board that it needed additional funding to add capacity to its congested airports. The funding was readily made available.

Because of the long lead time required to plan and build facilities and the even longer life of the facilities themselves, airport planning decisions are made under great uncertainty. It is difficult to predict demand in advance, and planners will either overestimate or underestimate. In this context, the bureaucratic and political costs of providing additional capacity too late are the continued criticism by the stakeholders (airlines, general aviation, the

travelling public). The cost of providing additional capacity too soon is excess capacity for awhile, which does not bother stakeholders. As airports were financed by grants from the Consolidated Revenue Fund, rather than loans, the bureaucratic and political costs were negligible. Indeed, if it was thought that the Treasury Board's fiscal posture was more receptive in the present than it would be in the future, there was a bureaucratic incentive to hasten construction.

Major airports in Toronto and Montreal had a second issue to contend with — should extra capacity be added at the existing airports (Malton, as Pearson was then called, and Dorval) or should new airports be built. The virtue of new airports was that, with substantial space, they could be planned as real showcases, thereby avoiding the physical constraints of the existing sites. Although new airports were more costly than expanding existing sites, cost was not a major problem. In the case of Montreal, the second airport at Mirabel was seen as an economic development project for the Montreal region.

In Toronto, another factor came to the fore, the response of those living near Malton to the prospect of additional noise due to new runways. In 1968 Etobicoke residents put a great deal of pressure on Transport Canada not to expand Malton. For Transport Canada, the path of least resistance became the construction of a new airport. Thus ensued a four-year search for a site, ending when Pickering was chosen in 1972.[1]

In the mid-1970s and into the 1980s the government ministry approach to airport planning, as exemplified by the decisions to build second airports at Mirabel and Pickering, was crumbling. Though the federal government succeeded at building an airport at Mirabel, the proposed Pickering airport met with strong opposition from local residents, environmental groups and, ultimately, the Government of Ontario. Local residents did not want to lose their homes to a new airport. Environmental groups argued that a Pickering airport would destroy the last semi-rural area close to Toronto. They also criticized CATA's planning assumption that air travel demand would continue growing rapidly for the rest of the century. This assumption was consistent with CATA's view of the comparative risks of under-building and over-building. In addition, the Mirabel Airport was tremendously under-used, implying that Pickering might also be a white elephant. Finally, the two-airport system in Montreal inconvenienced connecting passengers, with

the result that Montreal lost some connecting traffic to Toronto. When the Ontario government bowed to public pressure and refused to provide improved ground access to Pickering, the federal government decided, in September 1975, to delay the proposed Pickering airport indefinitely.

By the mid-1970s, the public sector context had changed. Rapid growth in government programs in the late 1960s and early 1970s, followed by slower economic growth in the mid-1970s meant that readily-available funding for CATA was drying up. The federal government was beginning to run deficits. Airports were now being looked upon as a source of revenue, and the first steps were being taken to move towards financial self-sufficiency. Airport fees were raised as part of the government's 1975 budget reduction package, and the air transportation tax was introduced in the late 1970s.

Even though CATA's capital funding was reduced, the demand for facilities was still growing because of the dramatic spurt in air travel caused by the deregulation of the airlines in the late 1970s and early 1980s. In this environment, CATA had to become a more efficient manager of existing facilities. Small capital improvements were undertaken, such as the construction of additional taxiways and improved navigational aids and air traffic control. Airport managers established scheduling committees in which they and the airlines determined how runway capacity would be allocated during peak hours. General aviation was discouraged from using major airports during peak hours. All of these measures were attempts to improve the efficiency of existing airports, while minimizing the dissatisfaction of important stakeholders.

The recession of the early 1980s, which led to sharp drops in traffic volumes for several years, bought some time for CATA and its airport managers. By the mid-1980s several factors were pushing Transport Canada to increase its emphasis on financial self-sufficiency for airports. The new Conservative government committed itself to the privatization of certain Crown corporations (for example, Air Canada, Petro-Canada) and the devolution of some functions previously housed in departments to special operating agencies (for example, the Passport Office) or self-financing corporations. In 1986 CATA was replaced by the Airports Authority Group (AAG) within Transport Canada.[2] The new structure has a mandate to run airports on a basis more closely approximating the self-financing nature of the private sector. Inspired by the privatization of the British Airports Authority, the federal government

has also been moving in the direction of privatization of the airports, and negotiating to turn over airports to local authorities in Vancouver, Edmonton, Calgary and Montreal on long-term (60-year) leases.[3]

While devolution/privatization may be part of the government's political agenda, it is also a result of the difficult state of federal government finances. With an accumulated debt of $420 billion, and debt-servicing obligations of over $40 billion per year as its largest expenditure item, the federal government cannot afford to fund major capital projects at airports. Thus, the third terminal at Pearson International Airport was entirely paid for and is now owned by the private sector.

Airport development is also affected by the federal environmental assessment and review process which is now a requirement in the planning of expansion of any major airport. In this context airport opponents have an opportunity to argue their case. The government, as proponent of additional airport investment, must demonstrate that adverse environmental impacts can be minimized, and that any remaining environmental damage is outweighed by the user benefits. As a result, the AAG has begun to undertake benefit-cost analyses of major airport investments. Since the environmental assessment process is a slow-moving judicial process, lead times for airport planning are much longer than they were 20 years ago.

Both the unavailability of capital and the environmental review process have made airport expansion much more difficult. Meanwhile, demand has been growing since the late 1980s. Consequently, major Canadian airports have once again become congested. All possible minor capital investments (taxiways, navigational aids) have already been undertaken. Airport managers are continuing to use scheduling committees to mollify the airlines and general aviation. But the travelling public is becoming more inconvenienced and angered by delays, particularly at Pearson, the nation's major hub and most congested airport.

Today, the early 1990s, airport planners have come full circle. As in the mid-1960s, they are facing congested airports but unlike then, there is limited federal government money to solve the problem. We have moved from an airport management regime which was the "government ministry model" to one more closely resembling a highly-leveraged "private sector management regime" which is unable to assume more debt. While the

concerns of the travelling public, the airlines, general aviation, and area residents continue to echo through the corridors of the AAG, its response is hampered by the federal government's woeful financial situation.

The following sections of this paper outline a model for airport management that puts the major emphasis on economic efficiency. As we will show, decisions made on that basis are different from those aimed at satisfying stakeholders or at achieving financial self-sufficiency. Furthermore, we will also illustrate how such an approach can mitigate the boom-or-bust airport expansion cycle we have witnessed in the last 30 years.

2.2 THE PLANNING FRAMEWORK

As already noted, the planning process for Canadian airports has been guided largely by physical standards rather than economic efficiency considerations. Such differences are not always recognized, particularly by physical planners who tend to draw on "economics" to justify investments they perceive necessary from their own physical or operational perspective. There is, of course, a fundamental difference between resorting to economic analysis (cost-benefit or economic impact) to justify desired projects, and the application of economic principles towards efficient use of existing or provision of additional capacity. This difference is evident from the recent airport planning experience in Canada, particularly with regard to the expansion of the two largest airports, Toronto and Vancouver. Background to the recent cost-benefit studies of both these airports, which are reviewed in more detail later in this report, reveals a distinct physical planning bias. In both cases, this bias has caused major efficiency losses before capacity could be expanded.

Capacity Considerations

To compare traffic demand to runway capacity requires that "capacity" be defined in physical terms as some rate of throughput (movements per unit time). Following Transport Canada and the U.S. Federal Aviation Administration (FAA) standards, runway capacity in Vancouver was defined to be the rate of throughput at which average departure delay reaches four minutes.[4] This standard definition of capacity reflects the general finding that once average delay time exceeds four minutes it rises rapidly with further increases in throughput; as throughput approaches maximum physical capacity, average

delay time approaches infinity. Given this definition of capacity, and assumptions about airport operating conditions, it is possible to calculate runway capacity.

The 1989 Vancouver study, by the Airside Capacity Enhancement (ACE) Project Team, used Transport Canada's Hourly Runway Capacity Computer Program to determine the throughput rate at which average departure delay at the Vancouver International Airport would be expected to reach four minutes, given a set of assumptions about weather conditions, aircraft mix and air traffic control procedures. This calculated value for existing runway capacity was found to be less than the existing throughput rate and, hence, capacity expansion was deemed to be required. Runway capacity was then recalculated to include short-term capacity improvements and compared to forecast traffic demand. From the finding that traffic demand would exceed runway capacity again within three years, it was concluded that there was a need for a new runway, the economic feasibility of which was to be assessed in a subsequent benefit-cost study.

If capacity expansion is defined to be the level of throughput that produces a four-minute average delay, then a second means of testing the need for capacity expansion is to compare the existing level of average delay to the four-minute standard. The ACE study, based on records of departure delays from May 1988, found average departure delays at Vancouver to be in the range of 6 to 11 minutes, indicating, under the standard definition of capacity, that capacity expansion was overdue. That average delay exceeded four minutes by a wide margin before the need for capacity expansion was identified is a reflection of the lack of any program for monitoring congestion delays at Vancouver prior to the ACE project.

Even if the delay monitoring program had been implemented early enough to identify average delays at the four-minute level, there is no guarantee that the methodology used in the ACE study would have lead to optimal timing of the benefit-cost study. This follows from the fact that capacity expansion may be justified in economic terms before (or after) average delay reaches four minutes. Capacity expansion is justified in economic terms when the (present-valued) benefits of capacity expansion exceed the (present-valued) costs of expansion. Since savings of congestion delay costs are the main benefits of capacity expansion, there is some critical level of average delay at which capacity expansion becomes justified. However, there is

no guarantee that this critical level of average delay is four minutes; the critical level depends on a number of factors, particularly the capital cost of expansion, and may be less than or greater than four minutes.

Hence, definition of "capacity" without reference to the cost of capacity expansion can lead to identification of a need for new capacity either before or after it is justified in economic terms. Economic criteria require that the need for capacity expansion be assessed not by comparing delay time to a physical standard, the economic merit of which is not examined, but rather by explicitly comparing delay cost to capacity expansion costs. In planning for a new runway at Vancouver, such an economic criterion was not employed until the benefit-cost study stage of the planning process. The initial need for capacity expansion, and hence the timing of the benefit-cost study, were determined by application of physical standards rather than explicit economic criteria.

A similar physical planning bias was evident in the background to the proposed capacity expansion in Toronto. Recognizing that air traffic control could not cope with the scheduled movements, and that substantial delays were occurring, a decision was made to cap movements at 70 per hour, which was later increased to 76 per hour. In the meantime, a number of system improvements were planned to increase the capacity of existing runways to 96 movements per hour by the mid-1990s. Demand projections suggested that even this expanded capacity would be exceeded, and that additional runway options should be considered. However, the focus of the debate was always the comparison of hourly capacity estimates (that is, some measure of maximum throughput within certain technical standards) and peak demand. As the analysis later showed, even if peak hour demand did not exceed the estimated "maximum throughput," additional capacity could be economically justified. In other words, since operations at or near the estimated "maximum throughput" cause significant delays to build up in the system, substantial new investment could be justified through savings that result from the reduction of delays.

Aversion to Pricing

The physical standards approach dictates not only the provision of additional capacity, but also the use of existing capacity. When congestion builds up, airport planners recognize the need to ration available capacity, but tend to

revert to administrative rather than economic means. In Toronto, for example, the hourly cap determines the maximum number of flights that can be scheduled, and then available slots are allocated to different users. First, the cap typically represents a level of operation closer to "maximum" rather than "optimum" throughput. There is little recognition of delay costs imposed on different users within the cap, nor is there any mechanism to discourage demand through congestion pricing. Second, available slots are allocated through a reservations committee, which has certain priorities but no mandate to allocate slots to the users who value them most.

The current scheduling practices at Toronto's Pearson International Airport certainly go a long way to avoid congestion, but still fall short of optimal pricing. The physical planning tradition has always displayed a tendency to resist efficient pricing policies. As in most aspects of transportation infrastructure management, this tendency is also evident in Canadian airports. Generally, economists view the engineers' (or planners') aversion to pricing as an inherent disregard for efficiency.

In practice, economists have to take some of the blame for overselling the virtues of pricing, at least in the absence of "rational" investment practices. As discussed in greater detail later in this report, efficient resource allocation in airport planning has two aspects: pricing and investment. Efficient pricing would ensure the best use of existing facilities but, in the long run, economic efficiency criteria can be met only through appropriate levels of investment. In the Canadian airport policy debate, economists have argued that the need for new capacity would diminish greatly if efficient pricing policies were in place. Recent studies, however, have shown that airside capacity investments were overdue both in Toronto and Vancouver, with or without congestion pricing. Thus, the country's major airports suffered from under-investment, which could not be cured through efficient pricing alone.

However, there is a legitimate aspect to the general resistance to congestion pricing at airports. This largely stems from the users' mistrust that, even if justified by demand, funds are not properly invested in additional capacity. Any attempt to increase landing fees is generally perceived as a way of taxing users, rather than creating funds for new (or paying for old) investments in airport facilities. It is, therefore, not surprising that user groups favour a slot reservation system over peak-period pricing. In the absence of a transparent mechanism to channel funds into investment, their fears are indeed

legitimate. Their efforts to ration available capacity among themselves is an attempt to squeeze out certain flights without placing a pricing burden on the rest.

Investment Justification

Physical planners may not always present the most compelling economic rationale for new investments, but they rarely lose their zeal for facility expansion. The two key forces that have held them back in the last decade are no doubt lack of funds and community opposition. Planners tend to counter these forces with economic impact studies in which airports are presented as generators of economic activity. They provide jobs and contribute to the local or regional economy through purchases of goods and services. Through these "direct" wage payments and other expenditures, a series of secondary effects are induced as employees engage in consumption and suppliers interact with other local firms. These ripple effects through the local economy are generally captured by "multipliers" on the primary impacts, taking into account "leakages" along the consumption (or expenditure) chain.

The purpose of economic impact studies is to measure the direct, indirect and induced effects of an airport. In other words, the studies try to capture the contribution of airport-related activities to the local economy, or local economic activity that may disappear with the removal of the airport. The methods and estimation techniques vary greatly, but all these studies try to impute an economic value or worth to an airport. The fundamental principles of impact studies are derived from basic regional economic theories, which strive to understand the spatial dynamics of economic activity, or spatial linkages between firms or industries. These concepts have been embraced by airport planners with somewhat different motives, mainly to establish the significance of their airport's role in the local economy.

Opposition from local residents to any expansion plans, and difficulties in securing investment capital, motivate airport planners to justify their role and enhance their profiles in the community. It is only natural to focus on such issues as local job and revenue generation to mobilize local support for new facilities, particularly in the U.S. where airport authorities tend to be locally governed. Especially when local financing is required, economic impact studies serve not only as a powerful public relations tool but also as

an effective mechanism to solicit public investment funds. It is understandable how economic impact studies became a fad throughout North America, virtually a compulsory undertaking for all local airport authorities through the 1970s and 1980s.

In Canada, economic impact arguments played a major role in justifying the construction of Mirabel Airport, especially since the project was promoted largely as an economic development initiative. The new airport was going to attract development opportunities that would otherwise locate elsewhere, and generate substantial local activity that would otherwise be foregone. Similar arguments were advanced for Pickering where, of course, they were never given a chance to be disproven as in the case of Mirabel.

Although the era of new airports closed with the death of Pickering, economic impact studies continued to play a visible role in the airport planning scene in Canada. In the last decade or so, every major airport in Canada has commissioned at least one economic impact study. The first such study for the Malton (now Pearson) airport in Toronto was in fact funded out of the "left-over budget" from Pickering, within a year of that project's cancellation. Similar studies followed in Edmonton, Calgary, Vancouver and some smaller airports. The earlier study of the Malton airport pre-dated some of the methodological advances in the economic impact culture that swept the U.S. through the late 1970s and early 1980s. In the late 1980s, the authorities felt the urge to commission a new, state-of-the-art economic impact study.

Economic impact studies play an integral role in establishing the importance of airports in the local or regional economy. They no doubt counter local opposition and mobilize political support. Similarly, they are useful in promoting the devolution of airports — which, as discussed in the next section, constitutes a cornerstone of the federal government's airport policy. As local governments understand the economic role of airports, they would naturally be drawn closer to the idea of owning and managing them, as a means of exercising more control or influence over local economic development initiatives. Despite these useful functions, however, a more critical — if not cynical — view of economic impact studies is difficult to avoid when they are being used to justify the building of new, or expansion of existing, airports — in other words, as an investment appraisal tool.

As evident from the Mirabel experience, airports cannot generate economic development; they can only facilitate development if the potential is there in the first instance. The same arguments prevailed throughout the eventually aborted Pickering project. Ironically, airport officials did not appear any wiser a decade later when they were preparing their case for the expansion of Pearson International Airport. Before any serious effort was made to examine the costs and benefits of airside capacity expansion, planners turned their attention to economic impact studies in the hope of proving that Toronto could not afford not to expand its airport. It should have been abundantly clear that the regional economy would be adversely affected by a congested airport, but it was somewhat ridiculous to try to justify a new investment based on potential job and revenue generation in an already overheated economy. As it turned out, benefits from reduced congestion would easily justify significant investment in new runways. The obvious lesson to be learned from this experience is that economic justification for any project lies with demand for that project, not with its consequences or impacts. Thus, expenditures on airport expansion should not be considered a net benefit or a measure of the airport's economic impact, because if the airport were not expanded, those resources would be used on some other construction project.

In conclusion, economic impact studies or statements may have considerable promotional value, continuing to play an important role in support of the federal government's devolution efforts. However, they are inadequate as an investment appraisal or project evaluation tool. As argued throughout this report, justification for airport investments can only be found through sound benefit-cost studies — in other words, economic efficiency must be established, as in all resource allocation decisions. In recent airport planning in Canada, preoccupation with economic impact studies has detracted from more serious and rigorous cost-benefit studies. Even from the perspective of community relations, economic impact studies prove to be of limited value. Local residents affected by negative externalities (for example, noise) take little comfort in positive economic impacts on their community (for example, jobs supposedly created for others).

2.3 THE POLICY CONTEXT

The current policy is moving in a new direction, offering greater comfort to the efficiency-minded economist in the Canadian airport planning scene. The three pillars of current federal airport policy — devolution/privatization,

cost recovery and environmental assessment — provide encouraging signs that the efficiency of the Canadian airport system is likely to be enhanced.

Devolution and Privatization

When the Canadian Air Transportation Administration under Transport Canada was replaced by the Airports Authority Group in 1985, it took charge of some 200 airports across the country. At the time they had an estimated replacement value (in 1985 dollars) of almost $8 billion. With a capital budget of more than $200 million, an operating-maintenance budget of almost $400 million and approximately 4,500 employees, the AAG became a very sizeable operating entity. The total revenues of some $330 million in the first year of operations were generated from terminal fees, landing fees, rentals and concessions. The portion of air transportation taxes allocated to the AAG (approximately $280 million in 1985-86) brought the organization's total revenue to more than $600 million.

In the transformation of CATA into the AAG (and the remaining components — safety, regulation and air navigation — into the Aviation Group), the federal government was motivated primarily by the need to create a commercially minded, businesslike organization. This was evident from the government's policy statement at the time, "Future Framework for the Management of Airports in Canada." The new policy package had two principal thrusts: transfer of Transport Canada-owned airports to local groups, and implementation of the Transport Canada Airports Authority Model in the remaining airports. Devolution is slow in coming, since Bill C-85 (*Airport Transfer Act*) is still before the House of Commons. Nevertheless, during the first few years, the AAG completed the transfer of 50 Arctic B & C airports to the governments of Yukon and Northwest Territories. This year, financial and employee benefit packages appear to have been concluded for the Edmonton, Vancouver, Calgary and Montreal airports, with the actual transfer to be completed shortly. In addition, the following initiatives are under way:

- An agreement has been reached to secure local authority financing (for example, through municipal bonds) for expanded airside and terminal capacity at the Vancouver International Airport.
- Base cases for Quebec City, Moncton, Windsor, Thunder Bay, Winnipeg and Kamloops are completed, and transfer negotiations will soon follow.

- Another round of transfers is expected to be completed in the next fiscal year (1992-93) and, by the following year, as many as 25 more airports will be transferred.

In addition to devolution, the AAG has also undertaken initiatives to secure direct private-sector involvement in the design, construction, financing and operation of facilities. In this regard, the most significant project was the development of the privately owned Terminal 3 at Toronto's Pearson International Airport, which opened in February 1991 with 24 gates and a capacity of 10 million passengers per annum. This $550-million infrastructural investment required less than $10 million in government expenditures, with foregone revenues estimated to be well below the costs of operating and carrying (that is, interest on capital costs) a project of this magnitude. The AAG is currently working on a tender package for the private-sector redevelopment of Terminals 1 and 2 at the Toronto airport. In addition, efforts are being made to involve private-sector interests to buy or lease smaller airport facilities as part of the overall devolution thrust.

As noted above, the operating costs of the AAG during its first year were about $400 million; by 1991-92, costs were down to $371 million and, by 1993-94, they are projected to be below $300 million. The downsizing is largely due to the transfer of airports to local authorities. Although detailed productivity studies are not available, the AAG has also been trying to improve the efficiency of its remaining operations. Together with improved cost effectiveness, there is also a thrust to expand the revenue base through both landing/terminal fees and concessions/rentals.

From our perspective in this research report, organizational aspects are not that critical in the pursuit of efficient pricing and investment policies, at least not in theory. The type of pricing and investment policies recommended throughout this report could have been implemented within the old departmental structure under CATA. However, the institutional record over the last two decades has proven that the bureaucratic environment was not conducive to the pursuit of economic efficiency in the running of existing, or the building of new, facilities.

The more business-minded approach brought about through the creation of the AAG is likely to promote more efficient pricing and investment practices. It should be obvious that a commercially driven organization, as opposed

to the old bureaucratic structure, will foster greater economic efficiency. At the same time, the devolution of both management and ownership helps bring about greater local accountability. This would tend to reduce the dangers of over-investment (for example, building of white elephants like Mirabel), while reducing the tolerance for under-investment through increased responsiveness to airport congestion. In general, the devolution and privatization thrust of the new airport policy framework is a positive development from an economic efficiency standpoint, though not necessarily a theoretical prerequisite to an efficiently run airport system.

Cost Recovery

Another important dimension of federal policy on airport finance and management is cost recovery. The first discussion paper outlining Transport Canada's cost-recovery policies was released in May 1987. This document formed the basis of subsequent consultations with various user and interest groups, leading to the publication of the second discussion paper in April 1990. Following another year of consultations, there now appears some speculation that the implementation of the policy package may be delayed, or even abandoned. In any event, the policies in question are of considerable interest from the perspective of this report.

Transport Canada's cost-recovery policies are no doubt part of the government's overall efforts to reduce the deficit. The underlying principles are stated as follows:[5]

- Ensure that users bear a fair share of the costs of facilities and services from which they derive benefits;
- Relieve the general taxpayer of financial burdens properly borne by users of the transportation system;
- Impose greater discipline on user demands for additional or better facilities and services; and
- Improve the efficiency of the transportation system, an objective that can be met through increased cost recovery because it will enable user demand, investment decisions, and modal choices to be based on a truer perception of the cost of service.

The relevance of these principles to airport pricing and investment, is underlined by the cost-recovery paper's assessment of airport revenues:

- Airports are presently classified into groups for cost-recovery purposes. Major fees (for example, landing fees) are the same for all airports in a particular group but are lower for smaller airports. Landing fees are established on a "residual" basis, meaning that they are justified by the shortfall, for any given group of airports, between total airport costs and all other airport revenues. Historically, landing fees have never fully recovered these shortfalls. This approach has resulted in fees that are not closely related to the costs of the specific facilities and services at particular airports.
- The largest single source of revenue from the air mode is the air transportation tax (ATT), an excise tax collected from domestic and international passengers. Unlike the proceeds of all other excise taxes which are treated as a general source of government revenue, the revenues from the ATT are credited to Transport Canada to help pay for air transportation facilities and services in general.

In determining capital costs, cost-recovery policy focusses on "net book value." While the AAG estimated the replacement value of airports at more than $8 billion, for cost-recovery purposes, the net book value was estimated at about $1.5 billion as of March 31, 1988. Using an average pre-tax return on total net assets of regulated industries in Canada, and including a provision for risk, annual capital costs of the AAG were estimated at approximately $218 million for the year 1987-88. With the appropriate adjustments and the inclusion of non-attributable components, total expenditures (that is, together with operating and maintenance costs) for cost-recovery purposes were estimated at $546 million.

Once the attributable costs are determined, cost-recovery policy deals with their distribution among user groups. The main users of airports are domestic commercial transport services, international commercial transport services, state and military aircraft and general aviation. Proposed cost- recovery policies view most of the airfield facilities at major federal airports as primarily intended to serve commercial transport operators, thus allocating all of the associated capital costs to these users. The operating and maintenance costs of airfield facilities are proposed to be distributed among all users,

per tonne of maximum take-off weight. The costs of special general aviation airports would be attributed to that user group. Terminal building costs would be divided into those associated with the passenger-processing part of the building, the space used by concessionaries (that is, commercial space) and the terminal building's car parking facilities. Based on these considerations, the principal cost-recovery proposals regarding airports are the following:

- The costs of airfields and terminal buildings should be recovered through site-specific charges.
- Airport user-charges should be established on a compensatory basis, in relation to the attributable costs of the airfield and the passenger-processing part of the terminal building.
- Airport airfield costs should primarily be recovered through landing fees, based on aircraft maximum take-off weight, applicable to turbo-prop and jet aircraft. The concession fees on turbo and jet fuel should be eliminated, and landing fees should be increased in such a way as to leave revenues unchanged. By exception, piston-engined aircraft should pay a concession fee on aviation gasoline and, where appropriate, an additional landing fee at large airports.
- Air terminal building passenger-processing costs should be recovered through general terminal charges based on standard aircraft seating capacities. A higher charge should be levied on aircraft in international service to reflect the additional space needs of their passengers (for example, for inspection services, well-wishers/greeters and longer dwell-times).

In addition, cost-recovery policy makes provision for peak-period charging at major airports where traffic exceeds capacity for considerable periods of time. The rationale is as follows:

- Facilities are sized to accommodate a substantial portion, but not all, of peak-period demand;
- Larger facilities, made necessary by peaks in demand, result in additional capital and operating costs; and
- These incremental costs should be borne, to the extent practicable, by the users who occasion them.

Although the cost-recovery statement provides the underlying principles, a more detailed methodology remains to be worked out with regard to both passenger-processing and airfield facilities. In the meantime, minimum landing fees, and two related provisions for Toronto and Vancouver (concerning fees payable by large piston-engined aircraft, and concerning the AVGAS concession fee) have been proposed.

The implications of the proposed cost-recovery policies from a pricing perspective are dealt with in more detail later in this report. There are some conceptual differences between the cost-recovery approach, on the one hand, and theoretical principles of social marginal cost pricing on the other. However, the proposed cost-recovery framework provides the essential elements of an efficient pricing system. Although this is clearly a positive development, it is now doubtful that the cost-recovery policy package will proceed as planned.

Environmental Review

As in all aspects of socio-economic activity, airport-related policies are also taking on an increasingly important environmental focus.[6] In the current fiscal year, the AAG notes the following key initiatives with respect to the environment:

- Guidelines for restrictions on night flights have been established and a major program for Noise Management has been instituted at Toronto Pearson.
- A 5-year program for the destruction of Polychlorinated Biphenyls (PCBs) throughout Transport Canada has been developed and program documentation completed.
- All international airports and each region now have a senior environment officer on staff.

For the next fiscal year, the following are noted:

- The Airports National Environmental Action Plan developed in response to the Federal Green Plan will be implemented. The key elements will be programs to identify and test sites for contamination, testing of underground storage tanks as well as the PCB program.

- While much planning and preliminary survey work will be accomplished, physical progress will undoubtedly be retarded due to the budgetary situation.
- . . . to shift our environmental focus from reactive to proactive, a comprehensive program of environmental audits, air and water quality monitoring, and development of environmental guidelines will be implemented.

Generally, airports and the environment are perceived to be in a perpetual state of conflict. As environmental concerns grow and public policy becomes more sensitive towards environmental quality, airport operators feel more pressure and encounter more constraints on the planning process. However, economic efficiency is not always in conflict with environmental objectives. For example, airplanes generate noise which has an impact on neighbourhoods in the immediate vicinity of airports. These, in fact, are costs borne by local residents, which are in principle not very different from operating costs incurred by airlines themselves. They are all "economic costs" associated with air travel; the difference is that some are borne "internally" by providers or users of commercial services, while others are imposed "externally" on other parties.

All costs, internal and external, have to be incorporated into efficient pricing practices. For example, air travellers should be paying for environmental costs they impose on society at large, in the same manner as they pay for airport facilities they use. In general, as environmental concerns come to the forefront of public policy debates, increasing attention will no doubt focus on external costs (noise, air or other environmental costs) generated by airport activity. However, this development, in and of itself, should not compromise economic efficiency, but place greater pressure for "prices" to reflect both internal and external costs.

Apart from pricing considerations related to existing airport activities, heightened environmental concerns also affect investment policies and practices. For example, as new airports are built, or existing ones expanded, more stringent environmental review and assessment are required. Recently, the proposed airside capacity expansions at both Toronto and Vancouver international airports came under the Environmental Assessment and Review Process (EARP), administered by the Federal Environmental

Assessment Review Office (FEARO). In both Toronto and Vancouver, the Minister of Environment appointed a panel to review Transport Canada's proposals. In each case, Transport Canada prepared detailed environmental impact statements (EIS) which were scrutinized by the panel through consultations and public hearings.

The EARP naturally slows down major projects. While placing an administrative burden, however, the EARP also imposes a greater degree of financial or economic discipline on public investment projects than that which might exist in the absence of such a rigorous review. Most environmental impact statements, particularly with regard to major airport projects, are expected to include a rigorous benefit-cost study. While other Treasury Board guidelines may also require similar financial scrutiny, the rigour with which the cost-benefit studies were conducted in the case of both Toronto and Vancouver airside expansion projects could in large part be attributed to the EARP requirements. It is doubtful that some of the white elephants of the past would pass the scrutiny of today's review standards. Thus, rather than hindering economic efficiency, the environmental review process has, somewhat ironically, enhanced it.

3. The Airport Investment Problem

Airports, like other large, public infrastructure facilities, are characterized by indivisible capacity — capacity that cannot be expanded continuously, but rather only in large lumps. Improvements to navigation and control facilities, taxiways, passenger handling facilities and access roads can enhance the capacity of existing runways and terminals to a limited extent, but, as usage increases, eventually new runways and terminals are required.[7] At that point it is not technically and/or economically feasible to construct half of a runway or half of a terminal: expansion of runway and terminal facilities requires a large, fixed investment. The lumpy nature of capacity expansion means that new capacity may initially be under-used but, as traffic volume increases, congestion builds and congestion delays mount. This section of the report deals with economic criteria for airport investments, first, from a theoretical perspective, then followed by two case studies dealing with airside capacity expansion at major airports.

3.1 THEORETICAL CONSIDERATIONS

Decision Criterion

The relevant economic criterion for evaluating public policies that affect diverse groups is the maximization of net social benefits. This criterion, by definition, accounts for the benefits and costs that accrue to all individuals who are affected by a policy, to arrive at a measure of the net benefit of the policy to society. The concept of net social benefit (NSB) encompasses not only private costs and benefits that accrue directly to providers and users of a service but also external costs and benefits that accrue to third parties. Hence, an evaluation of the proposed construction of a new runway that uses the NSB criterion would weigh the costs of the runway, both to airport operators and to the surrounding community who suffer noise and environmental disamenities, against the benefits of the runway, both to those who use it and, potentially, to those who receive economic spin-off benefits. The NSB of the policy of constructing the runway is then the sum of all social benefits minus the sum of all social costs.

The establishment of net social benefit as an economic criterion for decision making requires that costs and benefits be valued in some common unit, such as dollars. To ensure that the NSB reflects society's strength of preference for a policy, costs and benefits that have no existing market value are assigned the values placed on them by the affected individuals themselves, as revealed by their willingness to pay to receive a benefit or to prevent a cost.[8] When benefits and costs are defined in this manner, the NSB represents the increase in "social welfare" or "economic surplus" attributable to the policy, that is, society's valuation of the policy minus the social cost of providing the policy.

A policy with a positive, or even a maximal, net social benefit is not necessarily a policy that will make everyone better off. Positive NSB requires only that the benefits to those who gain from a policy exceed the costs to those who lose from a policy. The rationale for adopting a policy with positive NSB is, therefore, that it is possible to redistribute the impacts of the policy in such a way that no individual is made worse off, but some individuals are made better off, by the policy. Specifically, the gainers could hypothetically compensate the losers and still have some gain left over. The losers would then be no worse off than before the implementation of

the policy-compensation package, since their social cost, as measured in the NSB, equals their willingness to pay for the removal of the policy and therefore, the amount that they are willing to accept in compensation for the retention of the policy. This is known as the "compensation principle."[9] A policy that has positive NSB is deemed to be socially worthwhile because *if* those who benefit from the policy were to compensate those who lose, everyone would be at least as well off as before the implementation of the policy.

While maximization of the net social benefit may be a sound economic criterion upon which to choose among policies, it may raise political problems. For example, a NSB-maximizing policy, while conferring positive net benefits on society, may impose large losses on some individuals because a positive-NSB policy does not necessarily require that losers be compensated; the formula is that *if* they were compensated and there was still residual benefit to gainers, then the policy would be deemed positive. If the costs of a positive-NSB policy are concentrated among a group of individuals (such as those who inhabit the neighbourhood of a facility), while the larger benefits are spread thinly across the travelling public, each losing individual has a greater incentive to lobby against the policy than does each member of the general public to lobby for the policy. The likely result in such a case is the formation of interest groups opposed to a policy which, in terms of economic efficiency, benefits society as a whole. The implication of such interest group dynamics is that the use of the NSB criterion in policy selection may be constrained by the intrinsic nature of representative democracy.

In addition, the manner in which gains and losses are distributed among members of society may be of social concern. It may be a political objective to select policies that not only maximize the NSB — the size of the economic pie — but also that distribute costs and benefits — shares of the pie — according to some equity criterion,[10] for example, access.

Another type of equity criterion of potential political concern is the impact of a policy on low- versus high-income groups. The political objective of choosing policies that equalize the distribution of income across society is not reflected in the NSB criterion as formulated here. Summing costs and benefits across all individuals in an unweighted manner to obtain the net social benefit implicitly assumes constant marginal utility of income, that is, that low-income individuals value a dollar benefit or cost the same as

high-income individuals. A criterion which sought to equalize society's income distribution would assign more weight to a dollar benefit received by a low-income person than to a dollar benefit received by a high-income person. Therefore, the assumption of equal marginal utility of income is tantamount to overlooking the income-distributional effects of a policy, or assuming that they are negligible.[11]

Investment Timing

The airport planner's long-term decision problem is to determine the optimal quantity and timing of capacity expansion. If the planner's sole concern is economic efficiency, then the optimal capacity expansion path is that which maximizes the net social benefit of expansion. A solution to the problem is therefore a decision rule that identifies the NSB-maximizing quantity and timing of capacity expansion.

In practice, this decision rule is constrained by technical (physical and engineering) and economic (economy of scale) considerations that limit the quantity of capacity expansion to a few feasible options at any given time. The incremental benefit and cost streams (in comparison to some common base-case option) associated with initiating each of these expansion options at that given time can be simulated and the net present value (NPV) of each calculated. According to conventional benefit-cost practices,[12] the optimal policy would simply select the expansion option with the largest positive NPV, or the base case if the NPV of all other options were negative. The selected option would then be subjected to sensitivity analysis to determine whether delaying its implementation would raise its NPV.[13] However, such a policy, which considers the optimal timing of the selected option (the option with the highest NPV for current construction) but not the optimal timing of the other options, would not necessarily expand capacity in a manner that maximizes the NPV of net social benefits.

An optimal capacity expansion policy must optimize over both the quantity and the timing of capacity expansion. In particular, the policy must account for the possibility that start-dates other than the present for *all* options may change the preferred option. That is, the option with the highest NPV for current construction may not be the option with the highest NPV for future construction. In order to account for this possibility, the optimal decision rule for capacity expansion must in general determine

the NPV-maximizing start-date (optimal timing) for each expansion option, and then select the *optimally timed* expansion option with the greatest non-negative NPV.

The optimal timing for a given capacity expansion option is determined through a comparison of the incremental benefit of delaying expansion by one period to the incremental cost of such a postponement.[14] The benefit of delaying expansion by one year is the saving of the opportunity cost of capital in that year.[15] The cost is the lost net benefit (congestion savings less maintenance and externality costs, all over the base case) that capacity expansion would have produced in its first year of operation. The delay is justified as long as the benefit of delaying expansion by one year exceeds the cost of same. If congestion in the absence of expansion increases monotonically over time, and hence the annual benefit of expanding capacity increases monotonically over time, then the cost of delaying capacity expansion by one year increases over time.[16] By contrast, the benefit of delaying expansion by one year (the interest savings on expansion capital) is constant over time.

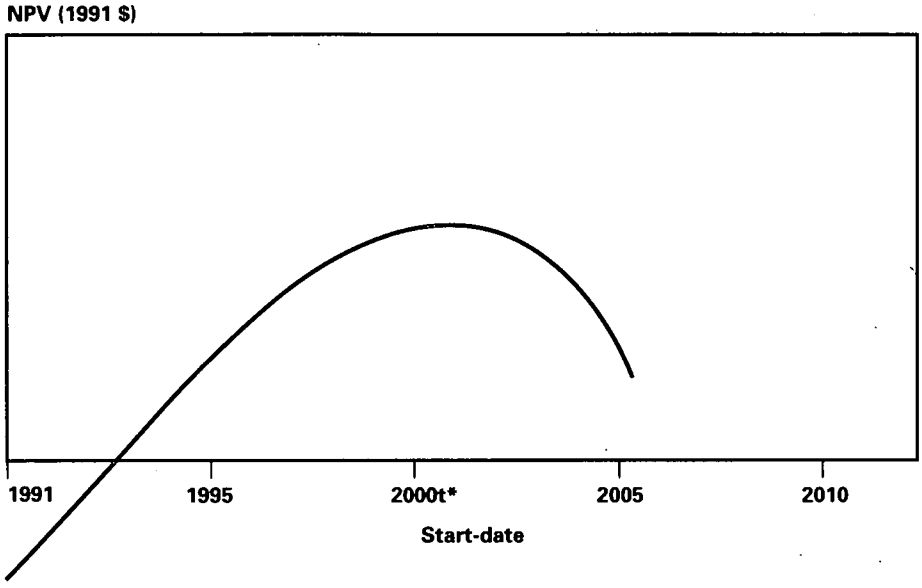
Given these assumptions, it follows that there will be a unique point in time at which the benefit exactly equals the cost of delaying capacity expansion by one year. This point is the start-date that maximizes the NPV of the expansion option, t* (see Exhibit 1). Before this optimal start-date, the cost — foregone benefit — of delaying expansion is less than the benefit — interest savings — of delaying expansion; hence, delaying the implementation of the expansion option increases its NPV. After the optimal start-date, the cost of delaying expansion exceeds the benefit and so further delay leads to a decrease in the NPV of the expansion option. Therefore, assuming monotonically increasing benefits to expansion over time, the optimal year in which to initiate an expansion option is the first year in which the net benefit of expansion exceeds the opportunity cost of expansion capital; that is, the first year in which

$$(B_t - M_t - E_t) \geq rK$$

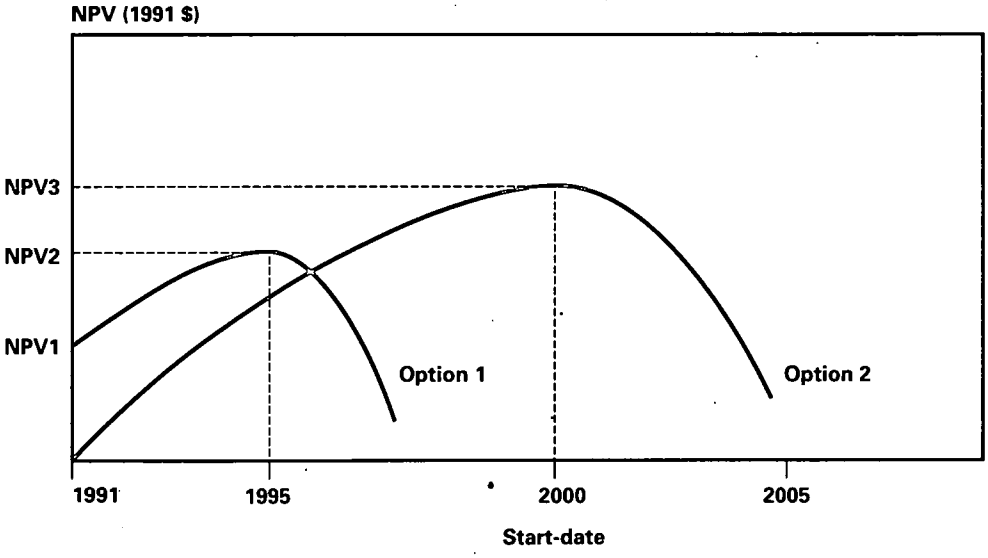
where B_t is annual congestion cost savings, M_t is annual incremental operations and maintenance cost, E_t is annual incremental external cost, r is the social discount rate, and K is the capital cost.[17,18]

Exhibit 1

VARIATION OF NPV WITH START-DATE



VARIATION OF RECOMMENDED OPTION WITH START-DATE



The existence of a unique optimal start-date for any expansion option stems from the assumption that the net benefits to expansion (exclusive of capital cost) increase continuously over time. If the assumption of monotonically increasing net benefits is not satisfied, then the decision rule for optimal timing presented above does not apply, and the optimal timing for initiating an expansion option must be determined by simulating each possible start-date and selecting the start-date that maximizes NPV.[19] However, the assumption of monotonically increasing net benefits is likely to be valid for airport investments, because indivisible capacity means that, in the absence of capacity expansion, congestion rises with demand and, hence, that the benefits of capacity expansion increase over time.

Selecting the recommended expansion option from among the set of optimally timed expansion options ensures that the NPV of capacity expansion is maximized over both the quantity and timing of capacity expansion. Use of a decision rule that chooses between expansion options timed to start at the present date rather than at optimal start-dates can alter the choice of expansion option, and hence expand capacity in a manner that does not maximize NPV. Exhibit 1 also provides an example of non-optimal decision making under a decision rule that maximizes present-year NPV for present-year start-dates rather than over all possible start-dates. In Exhibit 1 (bottom), Option 1 and Option 2 represent two technically feasible expansion options. For each option, potential start-dates are plotted against the 1991 NPV of implementing the option at each start-date.

Option 2 represents a larger capacity expansion than Option 1, with larger capital costs and larger congestion relief benefits, particularly in later years. Because of its larger capital cost, a greater level of congestion cost is required to offset the interest savings benefit of delaying Option 2; thus the NPV of Option 2 is maximized at a later start-date (2000) than is the NPV of Option 1 (1995). Although the greater capital cost of Option 2 causes it to have a later optimal start-date, and to have a lower NPV than Option 1 for early start-dates, the greater benefits of Option 2 in later years gives Option 2 a greater 1991 NPV than Option 1, subject to being able to delay initiation of Option 2 to the year 2000. A policy that chose between the two options based only on 1991 start-dates would select Option 1 and accrue NPV1 in 1991. A policy that further evaluated the optimal timing of its recommended option, Option 1, would find justification for delaying Option 1 until 1995 to raise the 1991 NPV of the project to NPV2. Yet this policy would not maximize

NPV in 1991. An optimal policy would evaluate the optimal timing of both options and recommend implementation of Option 2 in the year 2000, accruing the maximal NPV of NPV3 in 1991 dollars.

Hence, in general, it matters whether optimal timing analysis is performed before, rather than after, the selection of the recommended expansion option. In certain specific cases, the same option would be chosen regardless of whether optimal timing were considered before or after option selection. One such case of particular interest occurs when all options are "overdue" (past their optimal start-dates); the optimal timing for all options is then the present, and so a simple choice of the option with the greatest NPV based on a present start-date will maximize present-date NPV. This may have been the case identified in the recent runway expansion studies for Pearson and Vancouver international airports, both of which found the recommended option to be overdue and recommended immediate implementation.[20] While these recommendations may have been justified in a second-best sense, for a long-run planning policy the first-best solution would implement capacity expansion at, rather than after, the optimum time.

To devise a long-run planning policy that achieves both the optimal quantity and timing of capacity expansion would require looking beyond decision-rules for benefit-cost studies to the timing and frequency of the studies themselves. There would be a trade-off between the administrative costs of more frequent study and the benefits of improved timing. The conventional benefit-cost approach to long-run investment planning offers a robust solution to only half of the planner's long-run investment problem. The conventional approach tackles the question of *what* investment should be made at the current time, and then asks *when* this investment should be undertaken. An optimal solution would take a longer-term view of the long-run investment planning problem by focussing first on *when* each of a number of feasible investments would best be undertaken, and then asking which of these optimally timed options should be undertaken over time. Hence, the optimal long-run investment policy first determines the optimal timing of each feasible expansion option, which occurs in the first year that the net benefit of expansion exceeds the opportunity cost of expansion capital.[21] The optimal policy then selects the option which, constructed at its optimal time, produces the largest NPV in the present.

3.2 VANCOUVER INTERNATIONAL AIRPORT[22]

Scope

Following the economic recovery that began in 1985, aircraft movements at Vancouver International Airport (YVR) grew rapidly, increasing at an average annual rate of 11 percent during the period 1984-1988. The increased aircraft activity stemmed from the growing role of YVR as a hub for regional air services in the wake of airline deregulation and from a surge in Pacific Rim traffic. In response to mounting traffic and observable aircraft delays, Transport Canada initiated the Airside Capacity Enhancement (ACE) Project in 1988. The ACE project documented the existence of airside congestion at YVR during peak periods and identified short-term measures to enhance airside capacity.[23]

These measures included modifications to the existing runway system, air navigation technology improvements, modifications to air traffic control procedures, and a $25 peak-period minimum landing fee intended to shift some piston aircraft to other airports. Together, these measures were expected to boost YVR airside capacity by eight percent; this would constitute the base case for the study. Since even with full implementation of these measures traffic was expected to grow sufficiently to match capacity and delays to re-emerge by 1991, the need for a new runway was identified.

Several options for expanding capacity were formulated for comparison with the base case, each of which included the base improvements to existing infrastructure. These capacity expansion options were imposition of a $25 or $100 peak-period minimum landing fee; construction of a new runway of either 5,000, 8,000 or 9,400 feet in length parallel to the existing main runway; a $25 or $100 peak-period fee in combination with a parallel runway; and $100 peak-period fee in combination with construction of enhanced airside and terminal capacity at alternative Lower Mainland airports. The last option included enhancement of Abbotsford International Airport to allow it to function as a second airport for mainline carrier traffic, enhancement of surface transportation between the two alternate airports, enhancement of Boundary Bay Airport to attract non-commercial traffic, and a $100 peak-period minimum landing fee at Vancouver International to encourage use of the alternate facilities. Peak fees under all options would be levied during the 12-hour period of sustained demand at YVR on weekdays from 7 a.m. to 7 p.m.

The wide range of investment options considered in the study ensured that runway expansion at YVR would be recommended only if it was found to have greater economic merit, that is, net present value (NPV), than all other methods of alleviating congestion. In addition to the base-case efficiency improvements and alternate airport development, consideration of pricing as an alternative to investment was an important aspect of the study. The inclusion of peak-period pricing, both alone and in combination with a parallel runway, ensured that a parallel runway would be recommended only if justified as an alternative to or in addition to peak-period pricing.

If pricing alone were found to have greater economic merit (NPV) than runway construction alone, then comparison of the merit of the runway plus pricing combination with the merit of the pricing alone option would indicate whether runway construction was justified in addition to pricing. On the other hand, if runway construction were found to have a higher NPV than pricing, then comparison of the NPV of the runway plus pricing combination with the NPV of runway construction alone would indicate whether pricing was justified in addition to runway construction. In either case, the runway plus pricing option would not necessarily be superior to either the runway alone or pricing alone options because investment and pricing are alternative measures for dealing with congestion. Since both rely on the existence of delays for their justification, implementation of both runway construction and pricing would be justified only if sufficient congestion remained after implementation of either a runway or pricing alone.

In accordance with the requirements of benefit-cost analysis, the study considered not only a wide range of options but also the social benefits and costs of each option. The benefits of each option in comparison with the base case stemmed from reductions in congestion delay costs, whether by diverting some aircraft from peak periods to off-peak periods or alternative airports, or by expanding runway capacity at YVR. The latter options allow traffic to increase above base-case levels simultaneously with reduction in delay costs. This "generated" traffic, which would not have used YVR without runway expansion, accrues consumer surplus benefits under the runway expansion options that are additional to delay savings that runway expansion accrues to base-case traffic. Options that use peak pricing to alleviate congestion delays at YVR lead to net decreases of peak traffic at YVR and hence do not generate traffic in excess of base-case traffic.

Rather, under pricing options, a portion of base-case traffic is diverted from YVR, and this diverted traffic incurs a consumer surplus loss.

The second major type of benefit of capacity expansion is the incremental macroeconomic benefit of increased airport-related activity that results from generated traffic. Macroeconomic benefits can be attributed to an expansion option only if they would not have accrued to the economy if the resources used to expand airport capacity had been put to an alternate use. In the study of capacity expansion options at YVR, macroeconomic benefits were calculated but were displayed alongside rather than incorporated into option NPVs. Prudent investment planning dictates that decisions be made on the basis of user benefits alone because "the stimulative macro-economic effects of infrastructure projects are very small in relation to the overall volume of macro-economic activity and thus a great deal less certain than estimates of user benefits. As well, uncertainty of the stimulative impact of alternative uses of capital funds creates a risk of double counting benefits."[24]

Approach

Estimation of the capital and O&M costs of airport investment options can be based on the well-defined procedures for engineering costing and economic impact assessment. In the benefit-cost study of capacity expansion options at YVR, one measurement issue relating to capital costs concerned the allocation of the capital and O&M costs of surface transportation improvement to the alternate airport development option. A sensitivity analysis approach was taken to gauge the effect of allocating either 50 percent or 100 percent of surface transportation costs to the option. A second measurement issue concerned the cost of the land on which a parallel runway would be built. The study valued land costs at zero, arguing that there would be no alternative use for the Sea Island lands if a new runway were not built. The Federal Environmental Assessment Panel that reviewed the benefit-cost report argued that the land should have been valued on the basis of airport-related commercial development.[25]

The measurement of the user benefits and costs of airport capacity expansion require special attention involving forecasts of peak-period traffic volumes and average delay times. The reduction in average delay time over the base case engendered by an option can then be applied to the base-case traffic volume to obtain a measure of total delay minutes saved. If an option

alleviates congestion by diverting aircraft movements to off-peak times or alternate airports, delay time savings are calculated by applying average delay time savings to the remaining traffic volume. In both cases, delay minutes saved can be converted to dollar benefits by using readily available information on per-minute aircraft operating costs, aircraft load factors, and the value of business and non-business passenger time.

Consumer surplus benefits and costs to generated and diverted traffic can also be calculated directly from forecast traffic volumes and delay cost savings, provided that assumptions are made about the price elasticity of demand for aircraft movements. The benefit to each generated traffic movement (in excess of base-case traffic) is a fraction of the delay cost savings to base-case traffic movements, where the fraction is determined by the demand elasticity. In the YVR study, fractions in the range of one half to one third were used, reflecting the assumption of demand curves between the linear and log-linear form. Consumer surplus losses to diverted traffic movement were similarly calculated as a fraction of the increase in peak-period fees.

Hence, the key requirements for estimation of the user benefits and costs of airport capacity expansion are the ability to forecast peak-period traffic and the ability to translate peak traffic into average delay times under the capacity specified by each option. In the YVR study, forecasting of peak traffic for the base-case and runway options was relatively straightforward. The number of annual originating and destination passengers was first forecast on the basis of provincial population and disposable income growth. This traffic was then grossed up by a hubbing ratio to account for connecting enplanements and deplanements. Annual enplanements and deplanements were then translated into annual runway movements by making assumptions about aircraft sizes and load factors. Finally, annual runway movements were translated into representative peak-period runway movements on the basis of existing peak patterns.

The traffic forecasting process is complicated by the need to recognize the impact of congestion delay on traffic demand. As traffic increases, congestion delays increase as well, increasing the cost of using the airport, and decreasing the demand for aircraft movements. The effect of delay costs curtails traffic and delays growth in the base case, leading to decreased delay savings benefits (but increased traffic generation) from construction of new runways.

Ideally, traffic and congestion delay would be estimated simultaneously by a structural system in which traffic depends on delay, and delay in turn depends on traffic. However, in the YVR study, traffic and delay were estimated by two separate models, with traffic estimated first without explicit reference to delay, and delay then computed on the basis of traffic. Two strategies were adopted to compensate for the absence of explicit consideration of the effects of delay in the traffic forecasting model. First, the effects of delay costs on traffic were modelled implicitly by adjusting aircraft size and load factor assumptions upward and hubbing ratios downward in the base case to simulate the response of airlines to capacity constrained conditions. Second, after forecast traffic had been used to calculate delays, the sensitivity of study results to capping traffic growth at a level that produced a "maximum tolerable delay" of 20 minutes per aircraft was investigated.

Forecasting of peak-period traffic under pricing and alternative airport development options was achieved by adjusting base-case traffic forecasts. The percent of base-case peak movements by aircraft type that would divert to off-peak periods under a peak fee was estimated based on the response to peak pricing at other airports. Under the alternative airport development option, the diversion attributed to a $100 peak fee was adjusted upward to allow for the increased attractiveness of alternate airports and the enhancement of surface transportation links. The increased attractiveness of alternate airports was analyzed in terms of the types of aircraft that they would be upgraded to handle and the importance to aircraft operators of hubbing on YVR.

To translate peak-period traffic into congestion delay requires simulating use of the capacity provided under each option by the traffic forecast for the option. This was achieved in the YVR study by use of ADSIM, a discreet-event (aircraft-by-aircraft) airfield simulation model developed by the U.S. Federal Aviation Administration and applied to YVR by Hickling Consultants in their study. The simulation model uses queuing theory to predict hourly flow rates and average arrival and departure delays, given data on traffic demand, runway configuration and air traffic control procedures. The model was tested by simulating arrivals and departures over three days in 1989 to determine the extent to which predicted hourly flow rates mirrored actual operations. In each case, simulated flow rates were within one percent of actual, providing confidence in the model as a planning tool.

Simulated delays for 1989, however, significantly exceeded delays recorded by YVR's delay monitoring program. This was attributed to deficiency in the delay monitoring system rather than to deficiency in the simulation model. In fact it seems that the data collected by YVR's delay monitoring program are compromised somewhat by the method of its collection. Control tower personnel record aircraft movements as they occur, but some movements (25.5 percent in 1988) are not recorded because of workload priorities in the control tower during peak periods.[26] Because peak periods are times of greater than average delay, YVR's delay monitoring program systematically underestimates the actual average level of delay.

Hence, although ultimately an integration of traffic demand and delay forecasting models would be desirable, the separate estimation of traffic demand and congestion delay conducted in the YVR study provides a credible basis for measuring the user benefits and costs of airport capacity expansion. Given credible demand forecasts, the existence of simulation models capable of translating demand into average delay, along with the existence of market prices at which to value delay time savings to aircraft operators and passengers, makes estimation of user benefits a fairly mechanical process.

Potentially more difficult to measure are the external costs imposed by airport development on non-users. The YVR study identified three areas of external cost associated with airport development — noise costs; effects on birds, other wildlife and their habitat; and air quality. The methodology used to measure noise costs in the YVR study follows established theory that defines the various components of the social welfare loss produced in a residential neighbourhood by an increase in noise. The first step in the methodology was to determine the number and value of dwellings that move into higher noise contours due to operation of a new runway. The second step was a survey of real estate agents and the literature to estimate by dwelling type and noise contour the percent of householders who would move due to increased noise (6% on average), the depreciation in property values due to increased noise (2% to 6%), and to estimate householder surplus (the value that householders place on a dwelling in excess of its market value; 130% on average). The natural migration rate of those who moved for reasons other than increased noise was also estimated.

An important distinction was drawn between those who would move because of noise and those who would move for other reasons. Both groups of movers would suffer depreciation losses; those who would move because

of noise would also suffer a loss of householder surplus. Those who would stay would not suffer depreciation costs but rather noise annoyance costs. These noise annoyance costs must technically be less than the depreciation and householder surplus costs of moving; otherwise those who stayed would have moved. To be conservative, annual noise annoyance cost was estimated such that its present value equalled the present value of the sum of depreciation and householder surplus costs, under the assumption that the average householder would stay for six years on average after opening of the new runway. In addition, noise insulation costs were calculated for schools and hospitals, as well as moving costs of those who would move because of increased noise.

These various components of incremental noise cost were estimated for both runway expansion at YVR and capacity enhancement at Abbotsford International Airport for representative future years. After a given amount of time, all original residents were assumed to move away and noise costs to drop to zero since those who move in after the new runway was in place receive benefits from depreciated housing prices that offset noise nuisance costs.

Although the YVR study identified the existence of external costs other than noise, only noise costs were quantified. The rationale for not quantifying wildlife and air quality costs was that the net benefits of parallel runway construction were so large that they were unlikely to be offset by environmental costs. Although the Environmental Assessment Panel agreed that environmental costs would not outweigh the estimated net benefits, it did not accept the rationale for excluding them from the analysis:

> By so doing, the analysis implicitly undervalues environmental costs. The federal government's stated objective in the Green Plan is to incorporate environmental criteria into policy and decision-making processes. In this case that has not been done. . . . It is no longer acceptable to exclude these costs from economic analyses."[27]

As the Panel suggested, a reasonable shadow price for valuing wildlife losses is the cost of replacing lost habitat, either through purchase of compensatory habitat or implementation of conservation programs in remaining habitat.

Findings

Exhibit 2 summarizes the findings of the benefit-cost study of the airside capacity enhancement options at Vancouver International Airport. As it indicates, $25 and $100 peak-period minimum landing fees were found to produce net present values of $0.9 billion and $2.1 billion respectively. This reflects underlying estimates that the fees would divert 3.8% and 17.3% of peak-period traffic, respectively.

With an NPV of $3.8 or $3.9 billion, a parallel runway of 8,000 or 9,940 feet was found to produce greater net benefits than a peak-period fee. The amounts by which the NPVs of runway options exceed the NPVs of $25 and $100 pricing options are indicated by the figures in the columns labelled (b) and (c). Both pricing options produce greater net benefits than construction of a shorter runway of 5,000 feet capable of handling limited aircraft types. The pricing options were found to be superior to alternate airport development as well, whether 50 percent or 100 percent of surface transportation infrastructure costs were allocated to the latter option. The finding that a $100 peak fee alone produces a greater NPV than a $100 peak fee plus alternate airport development reflects the high capital cost of surface transportation improvements. It also reflects the underlying assumption that alternate airport enhancement — and hence the peak fee imposed to encourage alternate airport use — would not be fully implemented until 2001.

Whereas peak pricing alone was found to be superior to alternative airport development, and construction of the longer runways alone was found to be superior to peak pricing alone, construction of a longer runway in combination with peak pricing was found to be superior to construction of a runway alone. This reflects the finding that construction of a runway alone would not totally eliminate congestion and delay, either immediately or over the entire study period (to the year 2018). In fact, with construction of a 9,940-foot runway, delay was forecast to reach 1988 levels again by 2005. Hence, with a parallel runway, implementation of peak-period pricing would yield positive incremental net benefits, although these net benefits would not be as great as those attributable to peak pricing alone.

The combination of investment and pricing with the greatest NPV was found to be construction of a 9,940-foot runway with either a $25 or $100 peak-period minimum landing fee. Although the $100 fee in combination with

Exhibit 2

PRESENT VALUES OF BENEFITS AND COSTS OF ALTERNATIVE STRATEGIES (1986 $ MILLIONS)

	Strategy 1 Base case		Strategy 2 — Parallel runway											Strategy 3 Alternative airport	
			2A – 5,000 ft.			2B – 8,000 ft.			2C– 9,940 ft.			2D – 9,940 ft. & $25 fee	2E – 9,940 ft. & $100 fee		
	1B – $25	1C – $100	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)			50%	100%
User benefits:															
Air carriers			367	–62.9	–597	1834	1381	860	1890	1435	908	1890	1890	817	817
Passengers			439	–29.4	–616	2021	1575	960	2136	1605	1023	2136	2136	878	878
Total			**806**	**–92.3**	**–1213**	**3855**	**2956**	**1820**	**4026**	**3040**	**1931**	**4026**	**4026**	**1695**	**1695**
Costs:															
Capital			19	19	19	35	35	35	48	48	48	48	48	1091	2143
Operating			9	9	9	15.5	15.5	15.5	19	19	19	19	19	341	667
Sub-total			28	28	28	50.5	50.5	50.5	67	67	67	67	67	1432	2810
Noise			34.7	34.7	34.7	43.4	43.4	43.4	43.4	43.4	43.4	43.4	43.4	10	10
Total			**62.7**	**62.7**	**62.7**	**93.9**	**93.9**	**93.9**	**110.4**	**110.4**	**110.4**	**110.4**	**110.4**	**1442**	**2820**
Net present value			734	–164	–1285	3761	2862	1726	3915	2929	1820	3915	3915	253	–1124
Landing fee:															
Benefits	921	2140										326	633		
Costs	1.3	17										1.3	17		
Net benefits	919.7	2123										324.7	616		
Net present value	919.7	2123										4240.6	4531.6		

Source: **Hickling Corporation, *Economic Analysis of Airfield Capacity Enhancement Strategies for Vancouver International Airport* (March 1990), p. iv.**

Notes: **The sub-strategies (a), (b) and (c) for Strategy 2 refer to a minimum peak-period landing fee of $0, $25 and $100, respectively, in the base case. The range for Strategy 3 reflects an allocation of either 50 or 100% of the surface transportation costs. The macroeconomic gains are $2,576 million with Strategy 2 and $211 million with Strategy 3.**

a runway produced a greater NPV ($4.5 billion) than the $25 fee ($4.2 billion), the difference between these two amounts was found to be statistically insignificant. Therefore, the $25 fee in combination with the 9,940-foot runway was recommended, reflecting the low level of congestion that would prevail in early years. This was accompanied by the recommendation that the peak fee be reviewed with the intention of revising it upward in future.

The benefits of all options had much more influence on their NPVs, and hence their rankings, than did the costs. For every option except alternative airport development, total benefits exceeded total costs by an order of magnitude. This is not to say that the costs are not large in absolute magnitude. The total cost of a parallel runway would be approximately $110 million, $43.4 million (39%) of which was attributed to increased noise. Of the noise costs, property depreciation comprised 37%, noise annoyance 30%, lost householder surplus 16%, moving costs 8%, and insulation costs 8%. Despite the large share of noise costs in total costs, and the large absolute magnitude of total costs, total costs were small relative to total benefits. For the recommended option, total benefits ($4.3 billion) exceeded total costs ($ 0.11 billion) by a factor of 39.[28] Only under the alternative airport development option did costs approach or exceed benefits, and in that case only as a result of the very high cost of constructing and operating a surface transportation link, estimated at $2.7 billion (1988 present value dollars).

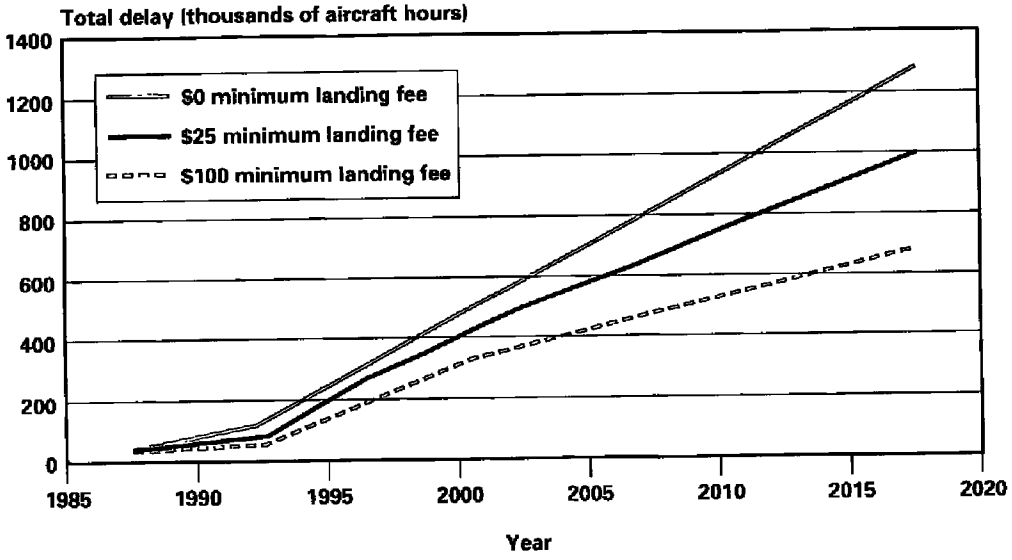
Total annual delay forecasts are presented in Exhibit 3. Even with the combination of a 9,940-foot runway and a $25 peak-period minimum landing fee, delays are forecast to return to their 1988 levels by the year 2010. The finding that, even with a new runway and pricing measures, delays can be expected to re-emerge in the next century prompted the recommendation that steps be taken to preserve the option of future development of alternative airports in the Lower Mainland region.

The findings of the study are subject to uncertainties in the parameters that underlie all forecasting. As noted above, one of the principal uncertainties is that of predicting users' response to mounting delay in the absence of a parallel runway. The benefits of parallel runway construction were based on forecast delays that rise to 127 minutes per aircraft by the year 2018. If growth in traffic, and hence delay, were dampened before delays reached this point, either because passengers found them intolerable or because the airport imposed an administrative cap, then the benefits of parallel runway

Exhibit 3

STRATEGY 1 — BASE CASE

TOTAL ANNUAL DELAY, 1988–2018



STRATEGY 2 — PARALLEL RUNWAY DEVELOPMENT

TOTAL ANNUAL DELAY

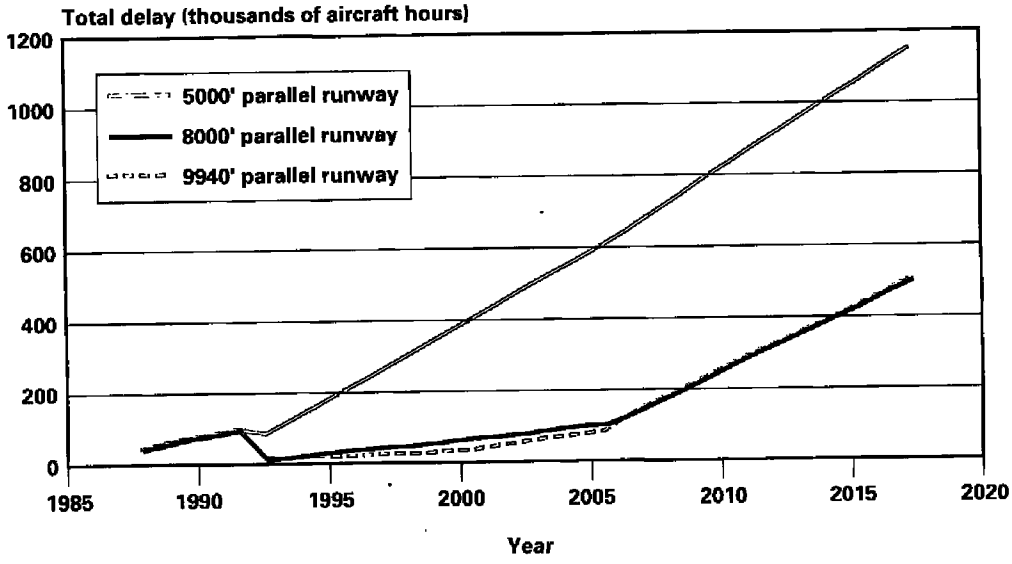
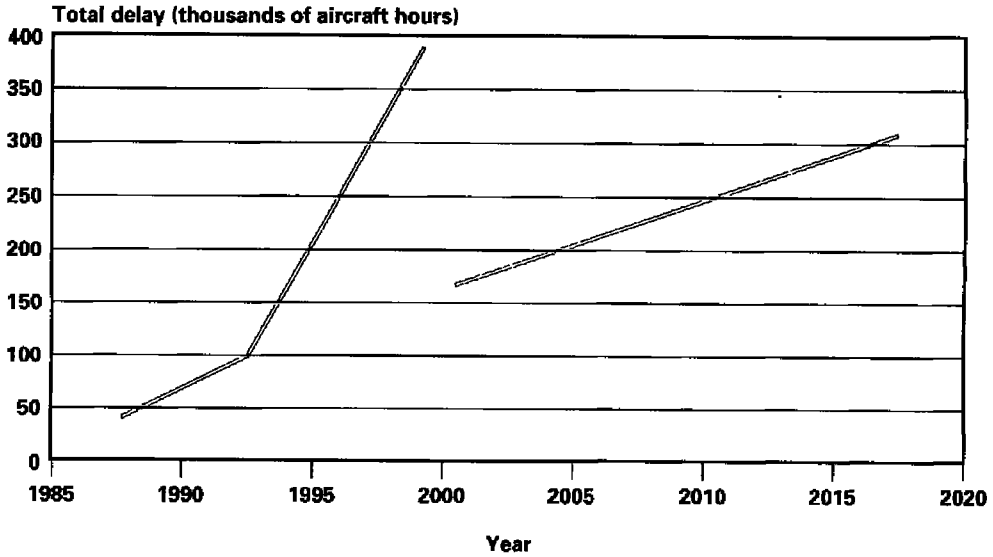


Exhibit 3 (cont'd)
STRATEGY 3 — ALTERNATIVE AIRPORT DEVELOPMENT
TOTAL ANNUAL DELAY



Source: Hickling Corporation, *Economic Analysis of Airfield Capacity Enhancement Strategies for Vancouver International Airport* (March 1990), pp. 116, 118, 119.

construction would be smaller than those upon which the findings of the study were based. However, sensitivity analysis revealed that ranking of options would be unaltered, with a 9,940-foot runway continuing to yield a positive NPV (of $238 million), even under the extreme assumption that traffic would cease to grow when average delay exceeded 20 minutes.

The study also accounted for uncertainty in all underlying assumptions by assigning a subjective probability distribution to each assumption and then using Monte Carlo simulation techniques to derive a probability distribution for the NPV of each option. Monte Carlo simulation allows all underlying assumptions to vary from their expected values randomly and simultaneously and hence simulates the effect of real-world uncertainty on NPV. In this manner, an 80 percent confidence interval was constructed for the NPV of each option on the basis of 80 percent confidence intervals assigned to each underlying assumption by expert panels. This risk analysis lead to the conclusion that for the recommended option "while there is a risk of the net

present value falling beneath the expected value of $3.9 billion, there is virtually no risk that it will be lower than $2.6 billion; a zero or negative net present value is associated with a zero probability."[29]

The YVR study found the first-year benefit ratio (FYBR) for a 9,940-foot runway to be 195%, well above the 10% discount rate, indicating a new runway to be far overdue in economic terms. The FYBR greater than 100% indicates that the runway produces net benefits in its first year of operation greater than its entire capital cost; this is confirmed by the payback period for the runway, reported in the study, of 0.44 years. Although not reported, the FYBR of the recommended option — a 9,940-foot runway plus a $25 peak-period minimum landing fee — would be slightly greater than 195% because the addition of pricing was found to produce small incremental benefits from further reductions in congestion in the first year of operation of the new runway.[30]

From its FYBR it is possible to calculate the annual cost of overdue runway expansion at YVR. By definition, the product of the FYBR (1.95) and the present value of the capital cost of a parallel runway ($48 million) gives the present value of the net benefit generated by the runway in its first year of operation ($93.6 million). The cost associated with overdue runway construction, as measured by the cost of postponing runway implementation by one more year, is the present-valued delay savings foregone in that year ($93.6 million) less the opportunity cost of capital saved by postponing implementation (0.1 × $48 million). The cost of overdue runway construction at YVR is therefore approximately $88 million per year.

The optimal timing of runway construction has to be examined in relation to peak-period pricing as well. This can be determined by calculating a FYBR using the first-year benefit of a runway given that pricing is already in place. This "incremental" first-year benefit of runway construction is the first-year benefit of the (runway plus pricing) combination less the first-year benefit of pricing alone. The YVR study reported a FYBR of 82% for construction of a 9,940-foot runway "with a $100 minimum landing fee in the base case."[31] Since the FYBR of 82% far exceeds the discount rate of 10%, runway construction at YVR would be overdue even if a $100 peak-period minimum landing fee were in place.[32] Hence, the findings of the study indicate that even with peak-period pricing, sufficient congestion delay costs would exist in 1993 (the assumed year of commissioning of the runway) to outweigh

the interest savings on capital that could be achieved by postponing construction by one year. The need for runway construction at YVR is overdue not only because of a lack of peak-period pricing but also because of a lack of physical capacity.

3.3 LESTER B. PEARSON INTERNATIONAL AIRPORT[33]

Scope

Similar to the case for runway expansion at Vancouver International Airport, substantial congestion delays at Pearson suggested the potential need for runway expansion there as well. Between 1984 and 1988, there was a 33 percent increase in passenger volumes and a 40 percent increase in number of aircraft movements per day at Pearson. This rapid growth in traffic stemmed both from buoyant economic conditions in Southern Ontario and from the emergence of Pearson as Canada's primary hub for domestic, transborder and international flights.

The growth in traffic strained existing airport capacity and resulted in increasing delays beginning in 1987. In response, the Minister of Transport introduced an aircraft reservation system and put in place a cap on aircraft traffic of 70 movements per hour. Increased air traffic control staffing led to an increase in the cap to 76 movements per hour in 1990.

At the same time, measures for increasing the efficiency of use of the existing airside infrastructure at Pearson were investigated. These measures included improvements to both infrastructure and operations to maximize the capacity of the existing airfield. These changes were expected to increase the hourly capacity of the existing runways to 96 movements per hour. Using peak-period pricing to shift movements to off-peak hours was found to be largely ineffective due to the fairly inelastic demand of most users of the airport. Despite their limited potential impact on delay, however, minimum landing fees are being introduced at Pearson.

With these improvements to existing airside capacity, traffic demand could be expected to reach 96 movements per hour within five years, leading to the re-emergence of severe congestion problems or the need to impose further caps on use. These findings led to the conclusion that Pearson needed runway expansion.

The base case for the benefit-cost study of runway expansion options at Pearson included all measures required to optimize the capacity of existing runways. These improvements included new taxiways, runway entries/exits, air navigation technologies and procedures, and full staffing of the air traffic control system. Peak-period pricing was not included in the base case, but it was assumed that the cap on runway movements would remain in place along with an administrative allocation system for shifting traffic from peak times to shoulder times and off-peak times.

The existing three-runway configuration at Pearson consists of two east-west (06-24 direction) parallel runways, and one north-south (15-33 direction) runway. The parallel 06-24 runways handle most traffic, but five percent of the time wind conditions prevent use of the 06-24 runways, limiting airport capacity to the single 15-33 runway. The benefit-cost study of runway expansion considered nine options for additional runways. Three options specified a single additional runway in the 06-24 direction, two options specified two additional 06-24 runways, three options specified a single additional 15-33 runway, and one option specified two additional 06-24 runways plus one additional 15-33 runway. The development of alternate airports was not considered because the five other airports in the vicinity of Toronto each face physical (ground or airspace) or institutional constraints that make their expansion infeasible.

The types of user benefits to 06-24 runway construction examined in the Pearson study mirrored those examined in the Vancouver study, and included delay cost savings to base-case traffic and consumer surplus gains to traffic generated by the new runways in excess of base-case traffic. The types of user benefits that were considered for construction of a 15-33 runway were more extensive. Such construction would not only alleviate delays caused by traffic growth but also the flight diversions and cancellations that are currently required during times when wind conditions prevent use of the 06-24 runways. Reduction of these disruption costs was the primary rationale for considering construction of a new runway in the 15-33 direction. Macro-economic benefits were not included in the benefit-cost study, but were documented in a separate study.[34]

Approach

To measure the benefits of runway expansion — both delay savings to base-case users and consumer surplus benefits to generated users —

requires estimates of the reduction in average delay time and the increase in traffic volume induced by new runways. These estimates in turn require forecasting of traffic under base-case and runway expansion conditions, and conversion of these traffic forecasts into average delay times under base- case and runway expansion capacities.

As noted in our discussion of the Vancouver benefit-cost study, one challenge in forecasting traffic is to model the effect of congestion delay on traffic growth. In the Vancouver study two approaches to this problem were attempted. Base-case traffic forecasts of aircraft movements were adjusted downward to reflect the use of larger aircraft with higher load factors by airlines in response to rising delay. Yet even with this adjustment, average delay was forecast to rise to high levels in the absence of a new runway, reaching 128 minutes by the year 2018. To account for the possibility that traffic growth would be severely inhibited by delay before delay reached such high levels, a sensitivity analysis assessed the effects of capping traffic growth when average delay reached 20 minutes. The effect of both attempts to model the effect of delay on traffic growth was to decrease the delay savings benefit of runway construction, but to increase generated user benefits to runway construction, by creating a gap between base-case and runway expansion traffic forecasts.

In the Pearson benefit-cost study, a more stringent approach was taken to modelling the effects of delay on traffic growth than in the Vancouver study. The Pearson study assumed that, in the absence of runway construction, the airport authority would intervene to cap the hourly flow of aircraft before delay reached high levels. The study assumed not only that base-case traffic would be capped at 96 movements per hour, but also that an administrative allocation system would be in place to shift traffic demand from peak to off-peak or shoulder times. These assumptions were reflected in two traffic forecasts: a role-related forecast applied to the runway expansion options and a constrained forecast applied to the base case. The role-related traffic forecast was based on air travel demand forecasts unconstrained by delay. The constrained forecast allocated role-related traffic across peak and off-peak (or shoulder) times under the constraint that hourly traffic not exceed 96 movements per hour. In this manner all role-related ATB (air terminal building) and cargo movements were accommodated in the constrained forecast. Some GA (general aviation) movements which represented the traffic generated by runway expansion, were not accommodated in the constrained forecast.

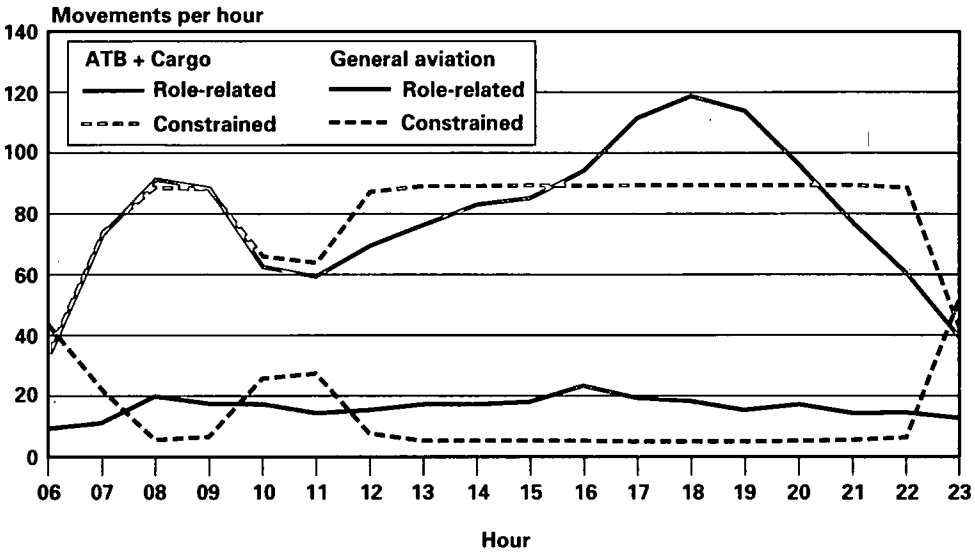
The two traffic forecasts were presented by planning day schedules for the years 1996, 2001 and 2011 that represent the averaged hourly aircraft movements of the seven busiest days in each of the three busiest months of a year. Exhibit 4 illustrates the planning day schedules forecast for the year 2011, and demonstrates how constrained traffic schedules were derived by spreading peak-period traffic over off-peak and shoulder times.

The Pearson study arrived at more conservative estimates of the benefits of runway expansion than did the Vancouver study. In the Pearson study, average delay in the base case was limited to that produced when the existing runways are operating at their maximum hourly capacity (96 movements per hour). This contrasts with the Vancouver study, which allowed hourly traffic demand to increase above maximum hourly capacity until average delay mounted to a maximum tolerable level. The result is that the administrative allocation approach assumed in the Pearson study led to less delay in the base case, and hence fewer delay savings to runway expansion than did the "maximum tolerable delay" approach assumed in the Vancouver study. The administrative allocation assumption also produced less generated user benefits than the "maximum tolerable delay" assumption since, by allocating some peak-period traffic to off-peak periods, it accommodated almost all role-related traffic movements.[35]

Hence the Pearson study, by assuming greater intervention on the part of the airport authority to limit delay in the absence of runway expansion, took a more conservative approach to estimating the benefits of runway expansion than did the Vancouver study. It could be argued that the Vancouver study also allowed for shifting of some users from peak to off-peak times by considering the implementation of a peak-period minimum landing fee. However, this peak fee resulted in a less extensive shifting of traffic away from peak times than did the administrative allocation response to congestion assumed in the Pearson study.[36]

Although the assumptions underlying their base-case traffic forecasts differ, the Vancouver and Pearson studies used the same method to convert base-case and expansion-option traffic forecasts into average delay times. As in the Vancouver study, the Pearson study used a discrete simulation model of airfield operations to predict the average delay times that would occur under the traffic forecasts and capacity conditions specified in the base-case and runway expansion options. The constrained planning day schedules were

Exhibit 4
2011 PLANNING DAY SCHEDULES



Source: Transport Canada, *Toronto Lester B. Pearson International Airport Airside Development Project, Final Report No. 24, Benefit/Cost Analysis*, TP10854E, April 1991, p. 43.

simulated in conjunction with base-case capacity, and the role-related planning day schedules were simulated in conjunction with each runway option capacity. Average aircraft movement delay was converted into average passenger delay by making assumptions about the distribution of aircraft types and load factors. Aircraft delays were then valued using aircraft operating costs, and passenger delays using a value for passenger time. The value of passenger time was constructed as a weighted average of the value of business and leisure travel time. As in the Vancouver study, the average wage rate of business travellers was used as an approximation of the market's valuation of an hour of work time, and leisure time was valued at 40 percent of the value of work time. This resulted in a weighted average value of passenger time of $26.33 per passenger-hour, expressed in 1990 dollars.

The average delay costs produced as such represented planning day delay costs for 1996, 2001 and 2011. To produce an estimate of annual delay costs, the planning day delay costs were used to estimate an average delay cost function using a queuing theory specification that describes the exponential relationship between runway movements (in this case movements per day)

and average delay cost. This average delay cost function was then applied to frequency distributions of base-year and forecast-year daily movements to obtain estimates of annual delay costs for the base case and each runway expansion option.

The types of benefits estimated for 15-33 (north-south) runway expansion were more extensive than the delay savings and resulting consumer surplus benefits estimated for 06-24 (east-west) runway expansion options. This reflected the role of an additional 15-33 runway in reducing the cost of disruptions that occur when wind conditions prevent use of the 06-24 runways. Without the 06-24 runways, capacity is currently limited to the single 15-33 runway; the result is sudden and severe congestion. Depending on the time and duration of such weather-induced disruptions, a large number of flights can be delayed on the ground, in the air on approach to Pearson, or on the ground at other airports. Some flights may be diverted or cancelled.

Estimation of the benefits of a second 15-33 runway was conducted by simulating the effects of a representative "weather incident" on forecast planning day schedules. The disruption costs of this representative incident were simulated with and without a second 15-33 runway, both in the presence of and absence of an additional 06-24 runway. The presence of additional 06-24 runway capacity affects disruption costs by affecting both the magnitude of forecast traffic and the size of the queue that is allowed to accumulate during the disruption. Disruption costs under each runway scenario were then converted to annual disruption costs based on historical data indicating the number of hours of weather-mandated 15-33 runway use over one year.

Simulation of a disruption incorporated the capacity rationing rules currently used during such incidents that give priority to larger aircraft and longer flights; general aviation movements are cancelled or diverted to other airports. The costs of cancellation, diversion, overflight and delay were calculated on the basis of a model developed by Transport Canada for the evaluation of approach aids.[37] Delay costs were estimated on the basis of average queue length, with all departure delays assumed to be taken on the ground, one third of arrival delays assumed to be taken in the air and the balance taken on the ground at another airport. Passenger cancellation costs include delay time, additional handling costs and the foregone benefits of travelling for passengers who do not reschedule, the latter estimated conservatively to be

the amount of their fares. For aircraft, the cost of cancellation is that associated with repositioning. Diversion costs are the extra flight time and ground transport costs associated with diverting general aviation and piston aircraft to nearby airports. Overflight costs are cancellation costs to Pearson-bound passengers who do not board aircraft that plan to overfly Pearson to go to their next destination.

As in the Vancouver study, depreciation in property values provided a basis for valuing increased noise costs. An empirical relationship between noise and property values was established through hedonic regression techniques which regressed housing sale prices on a range of housing characteristics plus a measure of noise exposure, NEF, for a sample of more than 3,000 observations within an eight-mile radius of Pearson. Dwelling market price data by enumeration area were obtained from Census and MLS data, along with the natural rate of emigration. The relationship between increased noise and increased moves out of the area was determined by estimating a dose-response function between those who reported being "highly annoyed" by noise and NEF levels.

Given these relationships between noise levels, property values and natural and noise-induced migration, the study calculated property depreciation, moving, householder surplus and increased noise nuisance costs. Property depreciation is a factor for all those who move, either naturally or induced by increased noise. Moving costs were attributed only to those who moved because of noise. Noise-induced movers also suffer consumer surplus losses stemming from their attachment to the community or their home. These losses were estimated by the difference between the subjective value of a dwelling and its market value, obtained by comparing valuations reported in Census data and MLS data. Increased noise nuisance costs apply to residents who remain in the area, and were estimated to be equal to imputed property depreciation. New residents moving into the area were assumed not to be affected by noise since the associated costs would already be capitalized in the depreciated price they paid for the property. Thus noise nuisance costs were assumed to diminish over time.

Environmental costs other than noise, such as loss of terrestrial and aquatic habitat, were not quantified, but were considered in choosing between 15-33 runway options with marginally differing NPVs, as described below.

Findings

Exhibit 5 presents forecast traffic and simulated average delay under the base case and the 06-24 runway expansion options. The base case includes nearly all traffic accommodated under runway expansion; generated user benefits from 06-24 runway expansion are therefore minimal. However, the delay savings induced by 06-24 runway expansion are substantial; runway expansion would reduce average delay cost compared to the base case, even in the first year of operation. One additional runway would reduce average delay by more than half; two additional runways would reduce average delay to near zero for the entire study period.

Exhibit 6 compares the present value benefits of 06-24 runway expansion options with their present value capital and operating costs. Noise costs are assessed at a later stage of the analysis and shown not to affect the results (see below). The figure demonstrates that while the costs of 06-24 runway expansion are large (in the range of $200 million per runway), the delay savings benefits are larger still, with all runway expansion options producing large positive NPVs. The recommended option specifies construction of two additional 06-24 runways at a present value cost of $354 million, and yields benefits of $1.3 billion for a NPV of $990 million.

This finding is consistent with those of the Vancouver study, the recommended option of which produced user benefits of $4.0 billion. That an approximate doubling of main runway capacity at Vancouver was estimated to produce user benefits three times those estimated for an approximate doubling of runway capacity at Pearson may reflect the more conservative benefit estimation technique used in the Pearson study.

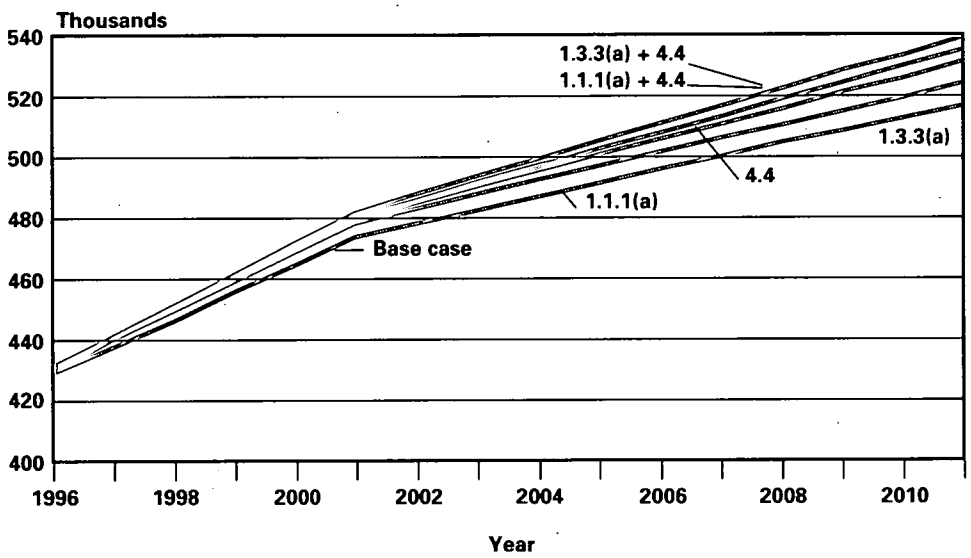
Not only were the estimated benefits of runway expansion higher at Vancouver, but the capital costs were lower, leading to much higher benefit-cost ratios for the recommended option at Vancouver (17.4) than at Pearson (3.8) and, also, much higher internal rates of return (76% versus 30%).

Exhibit 6 also compares the disruption reduction benefits of 15-33 runway options to their capital and operating costs. The benefits of an additional 15-33 runway are apparently not affected by the presence or absence of additional 06-24 runways. For both existing and expanded 06-24 runways, the two longer 15-33 runway options produce large positive NPVs.

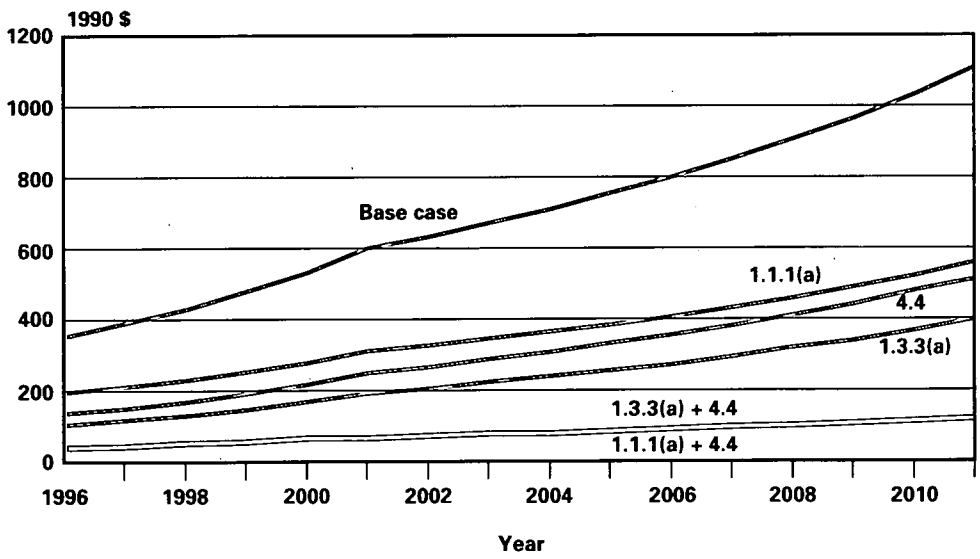
Exhibit 5

DELAY AND TRAFFIC IMPACTS OF RUNWAY OPTIONS

A) ANNUAL MOVEMENTS



B) AVERAGE DELAY COST PER MOVEMENT



Source: Transport Canada, *Toronto Lester B. Pearson International Airport Airside Development Project, Final Report No. 24, Benefit/Cost Analysis*, TP10854E, April 1991, p. 65.

Exhibit 6

BENEFIT-COST ANALYSIS: RESULTS SUMMARY

(PRESENT VALUE — MILLIONS OF 1990 $)

a) 06-24 Runway Options

	Option				
	4.4	1.1.1(a)	1.3.3(a)	1.1.1(a) +4.4	1.3.3(a) +4.4
Benefits					
to existing users	854.8	712.1	937.6	1,329.2	1,244.3
to new users	6.8	3.6	8.8	15.1	15.0
Total	**861.6**	**715.7**	**946.4**	**1,344.3**	**1,259.3**
Costs					
Capital	181.0	146.4	206.5	326.7	386.3
O&M	14.2	12.6	11.2	26.8	23.8
Total	**195.2**	**159.0**	**217.7**	**353.5**	**410.1**
Net benefits (benefits less costs)	666.2	556.7	728.7	990.6	849.1
Benefit-cost ratio	4.41	4.50	4.35	3.80	3.07
Internal rate of return %	33.6	32.5	30.7	29.8	24.6

b) 15-33 Runway Options

	Option					
	2.1.4		3.1.2		3.2.1	
	Existing 06-24	Expanded 06-24	Existing 06-24	Expanded 06-24	Existing 06-24	Expanded 06-24
Benefits						
Reduced disruption	159.5	163.7	279.4	274.1	395.4	410.3
Costs						
Capital	149.0	149.0	157.2	156.8	257.6	257.6
O&M	13.1	13.1	11.7	11.7	19.4	19.4
Total	**162.1**	**162.1**	**168.9**	**168.5**	**277.0**	**277.0**
Net benefits (benefits less costs)	(2.6)	1.6	110.6	105.6	118.4	133.3
Benefit-cost ratio	0.98	1.01	1.66	1.63	1.43	1.48
Internal rate of return %	9.8	10.1	16.4	16.1	14.6	15.3

Source: Transport Canada, ***Toronto Lester B. Pearson International Airport Airside Development Project, Final Report No. 24 Benefit/Cost Analysis,*** **TP10854E, April 1991, pp. 69, 103.**

Although these NPVs are not as large as those obtained for the 06-24 runway options, they nonetheless provide justification for construction of a 15-33 runway. The 15-33 option with the greatest NPV, option 3.2.1, was not recommended, however, due to environmental considerations that had not been quantified. Option 3.2.1 would require extensive fill within the Etobicoke/Spring Creek ravine, resulting in a much higher loss of terrestrial and aquatic habitat than would Option 3.1.2. Option 3.2.1 would also expose new areas to noise while Option 3.1.2. would not. For these reasons, the option with the second highest NPV, Option 3.1.2, was recommended. The choice of Option 3.1.2 over Option 3.2.1 leads to a loss in NPV, and an implicit valuation of environmental costs of approximately $30 million.

Noise costs were modelled for the recommended 06-24 and 15-33 runway options. Noise nuisance cost to remaining householders was found to be the largest noise cost component, accounting for approximately 65% of total noise cost. Total incremental noise cost amounted to only $5.1 million for the addition of an 06-24 runway and was negligible for the addition of a 15-33 runway. The inclusion of noise costs in the benefit-cost analysis had an insignificant effect on benefit-cost results, reducing the NPV of the recommended 06-24 option by only 0.5%. Variation in the cutoff level of noise exposure indicated noise costs to be two orders of magnitude less than the net benefits of runway expansion, regardless of the noise cutoff used.

Sensitivity analysis was performed to test the impacts of changes in many key assumptions underlying forecasts of 06-24 and 15-33 runway benefits and noise costs. Variation of model parameters within reasonable limits was found not to affect study results. Reasonable reductions in aircraft operating costs or the value of passenger time were shown to have no significant effect on the economic attractiveness of the preferred options; even if no value were attached to passenger time, the NPVs of the preferred options would be positive. For the 06-24 option, a cap on base-case traffic growth at 1996 levels was also investigated as an extreme reaction to delay; even under this assumption, sufficient delay cost savings would exist to justify two additional 06-24 runways. In the case of 15-33 runway expansion, the key variable was the amount of time during which weather conditions confine traffic to the 15-33 runways. A study of weather data indicated that such conditions occur 4.7% of the time. However, an additional 15-33 runway was shown to break even if only needed 2.9% of the time.

Delaying implementation of the recommended 06-24 runway option by one year was found to decrease its NPV by $40 million. This large cost of delaying runway implementation is consistent with the $45 million cost of a one-year delay calculated above for Vancouver and reflects the high cost incurred in running a congested airport. A first-year benefit ratio (FYBR) for the recommended Pearson runway expansion options can be calculated by dividing the (present-valued) first-year net benefit by the (present-valued) capital cost of expansion. For the recommended 06-24 runway option this yields a FYBR of 23.2% (76/327). This FYBR greater than the discount rate (10%) indicates that 06-24 runway expansion is overdue, although not as overdue as runway expansion at Vancouver, with a FYBR of 82%. Inclusion of administrative allocation in the base case of the Pearson study ensures that 06-24 runway expansion is overdue because of a lack of capacity, not a lack of use of existing capacity. The FYBR for the recommended 15-33 runway option is 11.2% (17.5/156.8), indicating that current timing of 15-33 runway expansion is close to optimal.

4. The Airport Pricing Problem

Having discussed the conditions governing the efficient level of capacity, we now turn to the question of the efficient use of a given level of capacity. The following section presents the theoretical solution to this short-run planning problem and a translation of that solution into requirements for an efficient short-run utilization policy. Several short-run policies are evaluated according to their ability to meet efficiency requirements ranging from administrative allocation to social marginal cost pricing. Finally, we review the current pricing practices at Canadian airports and evaluate the proposed cost-recovery policies of Transport Canada from the standpoint of economic efficiency.

4.1 THEORETICAL PRINCIPLES

Given a fixed level of capacity, airport planners must find the solution to two problems: determining the level of use of the fixed capacity, and allocating this level of utilization among users. Efficient use and allocation are determined by the trade-off between users' valuations of using the facility and the social cost of usage.

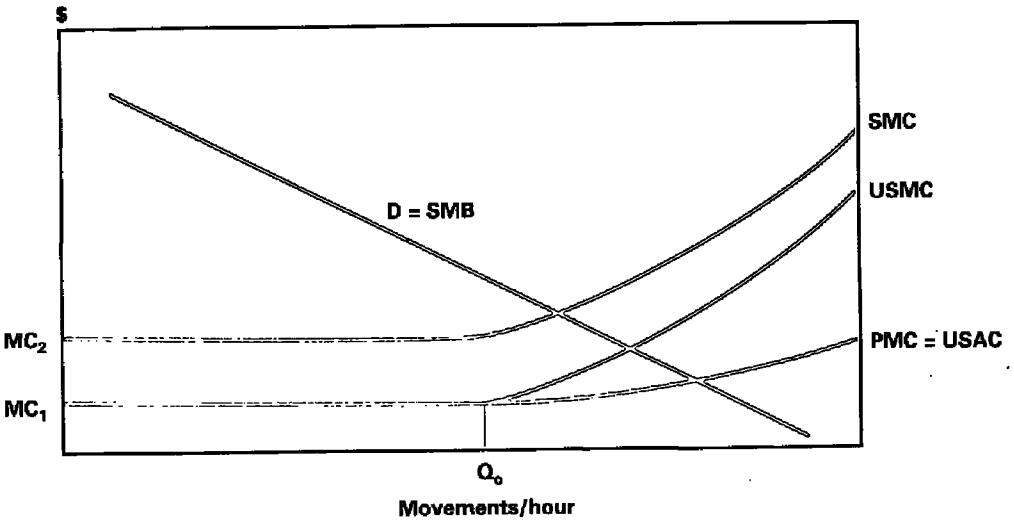
Exhibit 7 illustrates the short-run planning problem. For some fixed facility size, the x-axis indicates the level of utilization in movements per hour. The demand curve (D) represents aircraft operators' demand for usage of the facility and is derived from the demand for air transportation.[38] The demand curve plots users' valuations of using the facility (their willingness to pay for usage) in descending order. The demand curve is therefore the marginal valuation curve of users as a group: for any level of use the demand curve indicates the valuation of the marginal user (the user who values facility usage least) under the assumption that facility usage is allocated to users in the order of their valuations, with the user who values usage of the facility most (the "high-valued" user) allocated usage first. As use increases, the valuation of the marginal user falls and, hence, the demand curve slopes downward. Assuming that users are the only beneficiaries of airport use,[39] the demand curve is not only the users' marginal valuation curve but also the social marginal benefit (SMB) curve: at each level of facility use, the demand curve plots the benefit to society of increasing use by a small increment.

We now turn from the social marginal benefits of expanding facility use to the social marginal costs. The costs relevant to determining the efficient use of fixed airport resources are all social costs that vary with airport usage, including aircraft operating costs, passenger-time costs, airport operation costs and externality costs (such as noise costs). When considering marginal user costs, it is important to distinguish between costs that are borne by the marginal user and costs that the marginal user imposes on other users. The former are termed private marginal costs (PMC) and the latter, which arise from the increased congestion that an additional user imposes on all other users, are termed marginal congestion costs. The sum of private marginal cost and marginal congestion cost is termed users' social marginal cost (USMC), where "social" in this case denotes costs borne by all users.

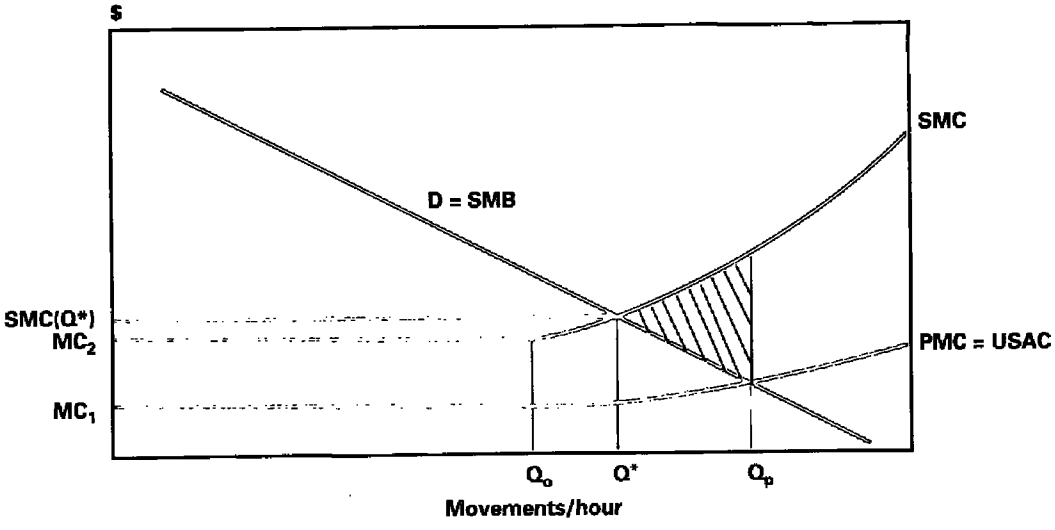
As use is expanded from zero to Q_o there is no congestion and the private marginal cost (PMC) is constant at MC_1 (see Exhibit 7). All users are able to pass through the facility at the maximum speed technologically possible, making the private cost of operating an additional aircraft the same as that of operating all previous aircraft.[40] Since PMC is constant, MC_1 also equals users' social average cost (USAC) in the absence of congestion. When use expands beyond Q_o movements per hour, the facility becomes congested, and the magnitude of congestion increases as the number of movements per hour increases.

Exhibit 7

Social Marginal Benefits and Costs of Facility Use



Social and Private Optimum Utilization Levels



Congestion affects all users equally: for a given level of utilization, all users experience the same level of congestion. The congestion cost borne by each user, when added to users' average operating cost without congestion (MC_1), is termed users' social average cost (USAC). Since the marginal user bears the same congestion cost as each other user, PMC equals USAC. Also, seeing that marginal users bear only the average, not the total,

increase in delay cost resulting from their usage, PMC is less than the cost to all users of marginal usage (USMC). If marginal users were to withdraw from the facility, the cost savings would include not only the operating and congestion costs borne by marginal users (USAC) but also the congestion cost that marginal users impose on all other users. Under congested conditions, the full cost to users of a marginal increase in use (USMC) is therefore greater than the cost to marginal users (PMC) and, hence, the USMC curve lies above the PMC curve over the congested range of use greater than Q_o, and coincides with PMC for use less than Q_o.

The concept of social marginal cost of expanding use can be widened to include costs that vary with use other than those directly borne by users. These non-user marginal costs include marginal airport operation costs (such as air traffic control and runway resurfacing) and marginal externality costs (such as noise costs). These costs to non-users of marginally increasing facility use are assumed to be constant at all levels of utilization and equal to $MC_2 - MC_1$ in Exhibit 7. The social marginal cost (SMC) curve is obtained by shifting the users' social marginal cost (USMC) curve upward by an amount equal to $MC_2 - MC_1$. The SMC curve plots the full social cost of increasing facility use by one movement per hour at each level of utilization; this social marginal cost includes the operating and passenger time costs of marginal users, the marginal congestion cost imposed by marginal users on all other users, the marginal airport operating cost of serving marginal users, and the marginal noise and other externality costs imposed by marginal users.

The efficient level and allocation of facility use are determined by the facility's social marginal cost (SMC) curve and social marginal benefit (demand) curve. The efficient level and allocation of use is that which maximizes net social benefits. Net social benefits are maximized when use is maximized subject to the constraint that no user value usage of the facility less than the facility's social marginal cost. This constraint is satisfied when marginal users (those who value usage least) value usage no less than the social marginal cost of serving them. Maximizing use subject to this constraint means expanding use in a manner that allocates usage to users in order of their valuations until the marginal users' valuation (and hence the SMB) equals the social marginal cost. Graphically, this solution is also represented in Exhibit 7 by the intersection of the demand curve with the SMC curve at utilization level Q^* and social marginal cost level $SMC(Q^*)$.

Note that congestion is not eliminated at the efficient level of facility use, Q*. The objective of maximizing net social benefits does not imply eliminating congestion, but rather expanding facility use until the social marginal cost of expansion (which is driven upward by increased congestion) exceeds the valuation of marginal users. Hence, the efficient level of use is not the no-congestion level, Q_o, but rather Q*, which represents the optimal level of congestion. Note also that the efficient level of use, Q*, is not the level of use that would arise in the absence of intervention by the airport authority.

New users have an incentive to enter the facility as long as their marginal valuations of usage exceed their private marginal costs. Hence, in the absence of regulation, facility use would expand beyond its socially optimal level to the private equilibrium level, Q_p. At Q_p social marginal cost exceeds social marginal benefit by a wide margin; the magnitude of the welfare loss (the loss of NSB) of operating at Q_p rather than at Q_o is depicted by the shaded area in Exhibit 7. This welfare loss under unregulated conditions represents a "market failure" that results from the significant external costs (particularly congestion) that are generated by use of the facility that are not borne by users. Attainment of socially optimum facility use requires intervention by the airport authority, either through utilization restrictions or through utilization pricing that shifts the social marginal cost of utilization to the users.

The optimal short-run policy is one that uses the facility at the flow rate of Q* and restricts use to users who value usage greater than SMC(Q*). In procedural terms, the implementation of such a policy requires the following steps:

- Allocate usage of the facility to users in order of their valuation of usage, giving priority to high-valued users.
- Continue to expand usage until the valuation of the next prospective user is less than the social marginal cost of expanding usage.

The first step requires that use of the facility be rationed efficiently among users, regardless of the overall level of utilization. The second step requires that the level of utilization be the efficient level, where the marginal valuation of usage equals the social marginal cost of usage.

4.2 CAPACITY ALLOCATION OPTIONS

Administrative Allocation

A common method of managing fixed airport capacity is through the use of traffic quotas set by airport authorities. Typically, a scheduling committee made up of individual airlines allocates the pre-set traffic quota among carriers in the form of slots and timetables which jointly satisfy the traffic quota.[41] This quota/slot allocation method has been used at a number of major Canadian and American airports, as well as at many airports in Europe and Asia.

Whether such an administrative method of allocating use can ration a traffic quota efficiently (giving priority to high-valued users) is open to question. A fundamental obstacle to efficient allocation is the inability of the airport authority to know the valuations of potential users of the facility and, hence, the inability unilaterally to allocate the quota in a manner that gives priority to those users who value usage the most. Incentive problems preclude the airport authority from simply asking each user his or her valuation since each user has an incentive to over-report in an attempt to receive a larger allocation. Rules for estimating the valuations of individual users are also fallible. For example,

> under the current scheduling committee system it is quite possible to prescribe a rule under which a charter jet, filled with tourists who are indifferent between landing at 5:00 pm or 8:00 pm, obtains a 5:00 pm peak hour slot, while a CEO of a large local company, travelling by small private aircraft, has his/her flight delayed for three hours, thus missing an opportunity to close a deal which would have brought substantial employment to the community. Although this person may have valued the 5:00 pm landing slot higher than the charter flight, a scheduling committee has no way of telling that.[42]

A more pervasive problem results from low-valued, general aviation flights delaying large, high-valued commercial flights. Inefficient quota allocation occurs because there are usually separate traffic quotas for general aviation and commercial traffic, and the slots for general aviation operations are allocated on a first-come-first-served reservation basis. Under such a system, general aviation flights are allocated slots in a manner that does not take

account of their valuations of those slots in comparison to other general aviation flights, nor in comparison to commercial flights.

Assuming that some mechanism can be found to incorporate general aviation into the slot allocation system in such a manner that slots can be allocated to all flights — general aviation or commercial — on the basis of their valuations, there remains the problem of designing an administrative allocation system that can identify and assign priority to high-valued users. Since users will not reveal their valuations without sufficient incentive, such an allocation must place users in a situation which induces low-valued users to reveal their low valuations through their willingness to be compensated for relinquishing their claims to slots. Conversely, high-valued users can be identified by their willingness to pay for slots. Hence, in general, users can be induced to allocate a traffic quota efficiently among themselves through a competitive mechanism which requires high-valued users to compensate low-valued users for the right to use peak-period slots.

It can be argued that the bargaining process inherent in a scheduling committee allocation system is a competitive mechanism that can achieve an efficient allocation of a traffic quota. The competitive nature of the bargaining process requires each airline to make concessions with respect to the number of flights that it will operate during peak periods. Each airline will concede its low-valued flights in order to retain its high-valued flights. Under certain conditions, such a bargaining process is likely to produce an efficient allocation of the traffic quota. The conditions are that the number of airlines be small, their valuation distributions similar, and each airline know the cost and demand conditions of the others.[43] As the number of airlines increases and airlines with differing valuations are introduced, the outcome of the bargaining process becomes less predictable, and it is uncertain whether a bargaining process alone will reach the efficient allocation. However, assuming that the airlines are able to make side payments (pay monetary compensation to each other), the traffic quota is likely to be allocated efficiently.

In addition to the requirement for efficient allocation, efficient use of fixed capacity requires that the quota itself be set at the efficient level, so that the social marginal benefit and social marginal cost of facility use are equal. As the discussion of bargaining has indicated, it may be possible for the airport authority, without knowing users' valuations of airport usage, to design an

administrative mechanism that rations a given traffic quota efficiently. It is, however, impossible for the airport authority to ensure that the quota is set at the efficient level without knowing the valuation of marginal users under the quota.

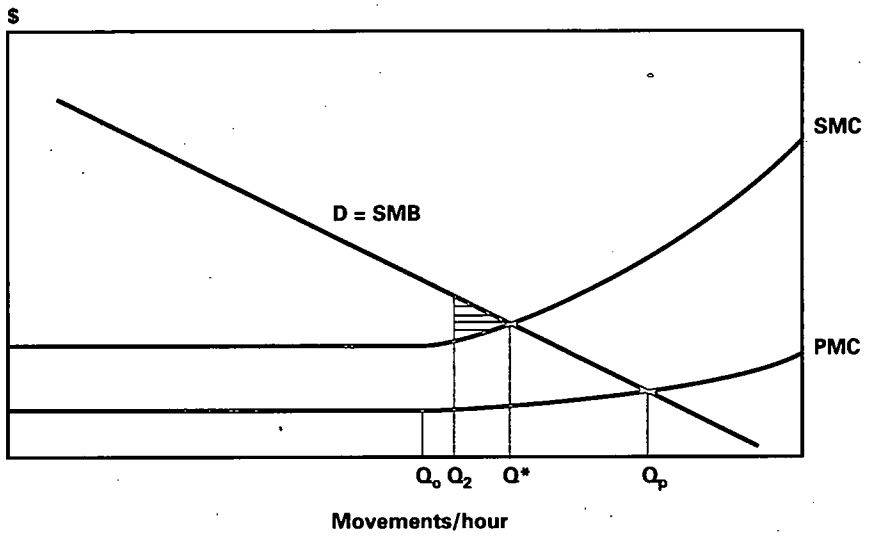
As illustrated in Exhibit 8, if the authority sets the traffic quota at Q_2 movements per hour, and the Q_2 movements are rationed efficiently, there will still be a welfare loss equal to the shaded triangle compared to the efficient quota level Q*. This is because, for a traffic quota of Q_2, the marginal users' valuation of using the facility is greater than the marginal social cost. The result is that there are (Q* - Q_2) users not using the facility who value usage more than the marginal social cost (including the cost of the extra congestion that they impose on other users) of serving them; hence, the traffic quota should be expanded. But if the airport authority does not know the valuation of the marginal user, the authority does not know to expand the quota.

Hence, although slot allocation systems based on bargaining among users may allocate a fixed traffic quota efficiently, the airport authority's inability to observe the marginal user's valuation of usage precludes the adjustment of the quota to its efficient level. The result is that administrative methods may meet the first requirement for efficient use of fixed capacity (that a given level of use be allocated efficiently among users), but are unlikely to meet the second requirement (that utilization be set at the efficient level). Under these circumstances, the magnitude of the welfare loss will depend upon the airport authority's ability to estimate users' aggregate demand curve, and thereby the optimal traffic quota, Q*. If, instead of estimating Q*, the airport authority follows a policy of setting the traffic quota at the level where there is no congestion, Q_o, then a welfare loss is certain to ensue.

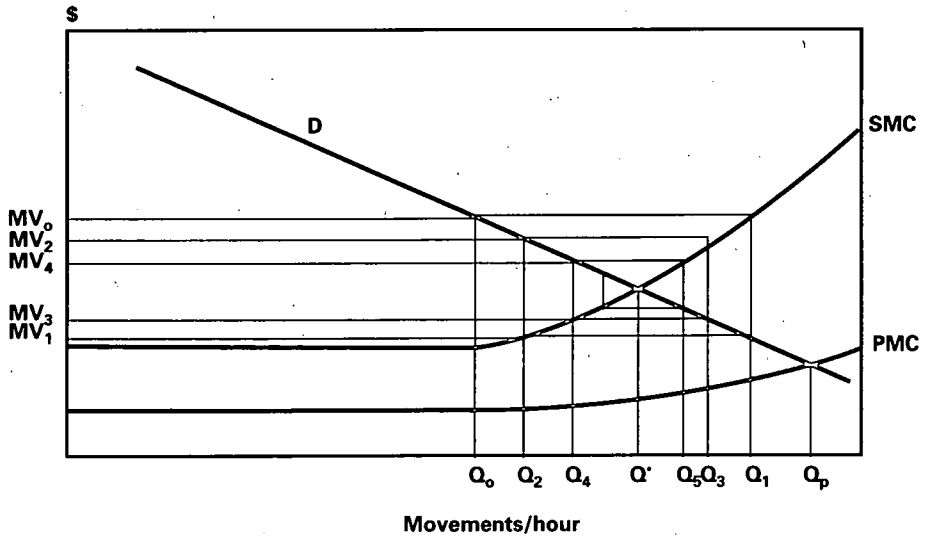
In addition to the uncertainties surrounding the efficiency with which it uses fixed airport capacity, an administrative allocation system may raise anti-competitive and cost-recovery concerns. Incumbent airlines have an incentive to resist the expansion of traffic quota if it will mean the admission of new airlines to the allocation process. If, for example, the airport authority seeks to expand the traffic quota beyond Q_o while simultaneously admitting new airlines to the allocation process, the authority may be criticized by incumbent airlines on the grounds that the expansion of quota would lead to an increase in congestion greater than the benefits of increased usage. The criticism would be well-founded from the incumbents' point of view,

but invalid from the point of view of social benefits. Hence, administrative allocation systems require an explicit mechanism for including potential entrants in the allocation process to prevent the system from becoming a barrier to entry.

Exhibit 8
An Inefficient Quota



Adjustment of Quota to its Efficient Level



Slot Lottery or Auction

The problem of efficient administrative allocation can be solved by replacing the bargaining mechanism with an explicit market in which users buy and sell peak-period slots. Since in order to sell slots, users must be given property rights over slots, this raises the question of how property rights over slots are to be assigned initially.

Incumbent airlines may argue that, in view of their substantial investments in existing facilities, slot rights should be assigned on the basis of existing usage patterns. However, ceding slot rights to incumbent carriers raises the same competitive concerns that arise under administrative allocation. In fact, anti-competitive tendencies may be strengthened under a system that gives slot rights to users: an incumbent airline, for strategic reasons, may not be willing to sell slot rights to an entrant airline, even if the entrant is willing to pay more than the incumbent's valuation of the slot.

A system that assigns slot rights on some basis other than current usage could prevent anti-competitive activities. Rights could be assigned randomly through a periodic slot lottery conducted by the airport authority. Following a lottery, slots could be bought and sold among airlines, but the ownership of rights would not extend past the next lottery date. However, slot lotteries raise new equity concerns: the lottery system could assign windfall gains to low-valued entrants who could sell their lottery-won rights to high-valued incumbents. The possibility of windfall gains requires that safeguards, such as a registration fee, be built into the lottery system to deter illegitimate entrants. While providing windfall gains to some users, slot lotteries in their pure form would not provide revenue to the airport operator for cost recovery.

With a market, rather than a bargaining, mechanism to allocate a traffic quota, there is much greater certainty that efficient allocation will occur. The market mechanism would ensure that the traffic quota, once assigned by the lottery, would be redistributed among users on the basis of willingness to pay, since low-valued users would always be willing to sell slots to high-valued users. Furthermore, the periodic re-assignment of rights through lotteries would act as a deterrent to strategic behaviour by low-valued incumbent users.

The problem of setting the traffic quota at its efficient level would still remain since the airport authority must determine the level of the traffic quota to know how many slots to assign in the lottery. The market

mechanism for rationing would, however, provide the airport authority with a clearer signal for adjusting the quota towards its optimal level than does the bargaining mechanism. The market clearing price for slots should be a good estimator of the valuation of the marginal user.[44] If the airport authority is able to observe this market clearing price — perhaps by acting as a market maker — then the authority could use this information to adjust the traffic quota in the direction of its optimal level at the next lottery.

Such an adjustment process is illustrated in Exhibit 8. The airport authority initially lotteries a quota of Q_0, and from the re-sale market for slots observes a marginal valuation of MV_0. The authority then calculates the level of quota at which social marginal cost equals MV_0, Q_1, and sets the quota at Q_1 at the next lottery.[45] After that lottery, the authority observes a marginal valuation, MV_1, and sets the quota for the following lottery at Q_2. This process continues, the airport authority using revealed marginal valuations to adjust the quota in the direction of its efficient level, Q*. Following the cobweb path shown in Exhibit 8, it is evident that the quota converges to Q*.[46]

Use of this algorithm for adjusting quota to its efficient level does not require the airport authority to know users' entire demand curve, but rather only to measure users' marginal valuation at a given quota by observing the market price of slot rights. Hence the algorithm, although presented here for the case of unshifting demand, should be robust to intertemporal shifts in demand. The algorithm's reliance on the market price as a measure of users' marginal valuation is, however, potentially fallible. Each user has an incentive to drive the quota above its socially optimal level, Q*, to its privately optimal level, Q_p. To this end, users can potentially manipulate the quota adjustment mechanism by inflating the market price above its real level and paying unobservable refunds on the side. Therefore, although the use of a slot lottery system with a re-sale market provides a more efficient allocation of quota and a clearer signal for adjusting quota than does administrative allocation, there may still be incentive problems in measuring users' marginal valuation.

The equity and cost-recovery problems associated with random allocation of traffic quota can be overcome by replacing the slot lottery with a periodic auction of slots. The auction mechanism would allow an initial allocation of slots on the basis of users' willingness to pay, and hence would avoid equity problems. Allowing post-auction trading of slot rights would compensate

for any allocative imperfections in the auction mechanism (such as those stemming from the dynamic nature of the auction process), ensuring that the final allocation is efficient. The proceeds of the auction would go to the airport authority, not to the windfall gainers, and could be used for cost recovery.

Although a system of slot auctions with a re-sale market would allocate a given traffic quota as efficiently as any other allocative mechanism with a re-sale market, the problem of determining the efficient level of quota remains. Under a slot auction system, the process of adjusting quota to its efficient level is complicated by the presence of two signals of users' marginal valuation: the minimum winning bid in the auction and the market price in the re-sale market.

SMC Pricing

Each of the short-run facility use policies examined thus far — administrative allocation, slot lottery and slot auction — allocate a given traffic quota efficiently to differing extents, thus satisfying the first condition for efficient facility use. However, they all have a problem with satisfying the second condition since they all have to measure users' marginal valuations to determine the efficient level at which to set the traffic quota. This stems from the very nature of allocation systems that rely on quantity methods to control use by limiting use to a specified number of movements per hour. The alternative is to control use by setting a minimum marginal valuation, rather than a maximum utilization level. This is achieved simply by setting a price for usage and then letting any user who is willing to pay the price use the facility. Each price induces a unique level of use by users, as specified by the demand curve. Hence, the airport authority can control use at least as effectively by controlling the price that users pay for utilization as by controlling the level of use.

The immediate benefit of using price, rather than quantity control, is that the first condition for efficient use — that of allocating usage to users in the order of their valuations — is satisfied with certainty. This is because the price control acts as a rationing mechanism which induces low-valued users to sort themselves from high-valued users: for any given price only those users who value usage more than the price will be willing to pay and hence use the facility. Price control is more certain to achieve efficient allocation

than are the other mechanisms considered thus far because the other mechanisms can initially produce inefficient allocations and, hence, must rely on re-sale markets to ensure efficient final allocation. Such re-sale markets, since they depend on trade among users to ensure allocation on the basis of willingness to pay, are prone to high transaction costs, imperfect information and strategic behaviour. By contrast, controlling use through pricing, because it puts a limit on user valuation rather than on use, always produces an efficient allocation: each user who values usage more than the price is guaranteed usage.

The problem is then to set the price at its efficient level, that is, the level that satisfies the second condition for efficient use. The efficient price is that which produces a level of use at which the social marginal benefit equals the social marginal cost, that is, the level Q* shown earlier in Exhibit 7. Hence the efficient price is SMC(Q*); a price set at SMC(Q*) is efficient because it imposes on each user the full social cost of marginal usage. This internalization of the user's congestion cost corrects the market failure (the divergence of social and private optimum) that occurs when users pay only their private marginal costs, and produces an equilibrium that is socially, rather than privately, optimal.

In order to make users pay the price, SMC(Q*), for facility usage, it is not necessary to charge a user fee or toll equal to SMC(Q*), since the user already pays the private marginal cost, PMC(Q*), to use the facility. It is only necessary to impose a toll of SMC(Q*) - PMC(Q*) to cause the users to internalize the social marginal cost of usage. Since private marginal cost is the same for all users and therefore equals users' social average cost (USAC), the optimal toll can be restated as SMC(Q*) - USAC(Q*), the social marginal cost less average user cost at the optimal level of utilization.

However, since the airport authority does not know the demand curve, only the cost curves, SMC(Q*) is not known and must be found through an iterative process analogous to that used to adjust the traffic quota. This algorithm for adjusting the user toll to its optimal level is encapsulated by the policy of continually setting the toll equal to SMC(Q) - USAC(Q), regardless of the utilization level (Q) that currently exists. Such a policy is known as social marginal cost pricing (SMC pricing). Convergence of the user toll to its optimal level under social marginal cost pricing can easily be demonstrated. Assume that facility usage is initially set at the maximum no-congestion

level, Q_o. Social marginal cost pricing dictates initially setting a user toll equal to $SMC(Q_o) - USAC(Q_o)$, which effectively shifts users' average cost curve, USAC curve, (and hence their PMC curve) upward by $SMC(Q_o) - USAC(Q_o)$. In response, users will increase use to the private optimal level Q_{p1}, where the new PMC curve intersects the demand curve. If the airport authority then resets the toll to $SMC(Q_{p1}) - USAC(Q_{p1})$, users will respond by decreasing use to the new private optimum level Q_{p2}. The algorithm continues, with the airport authority using the users' quantity response as the basis for setting new SMC prices, until Q_{pn} converges to Q^* and the toll converges to its optimal level, $SMC(Q^*) - USAC(Q^*)$.

The SMC pricing policy that forms the basis of this algorithm requires only knowledge of the facility's social and private marginal cost curves, not the demand curve. The algorithm does not require observation of users' marginal valuation because the trial prices used in the algorithm are essentially trial marginal valuations. The algorithm adjusts marginal valuation to its optimum level by observing quantity responses, rather than adjusting quantity to its optimum level by attempting to observe marginal valuation responses. The robustness of this price adjustment algorithm, in comparison to quota adjustment algorithms, is that it does not require the airport authority to observe an intangible quantity (users' marginal valuation in response to trial quotas), but rather only a tangible one (the level of usage in response to each trial price).

Thus far, the analysis of facility use has assumed a static demand curve. Introduction of differing levels of demand by time of day, season and calendar year is needed to model more closely the fluctuating, peak-load nature of demand that the airport planner faces. SMC pricing implies optimal user tolls that vary with the level of demand. A shift in the demand curve necessarily leads to a change in the socially optimal level of facility use (where the demand curve intersects the SMC curve), and hence a change in the optimal toll. For example, an increase in the level of demand leads to an increase in the socially optimal level of use, but, at that new level of use, marginal congestion cost is greater and consequently a higher toll is required to ensure that usage remains limited to those users who are willing to pay the social cost of their usage. In this manner, a policy of social marginal cost pricing recognizes the social justification for allowing congestion to increase when users' valuations of usage increase, but limits the increase in congestion to that justified by users' increased willingness to pay. Hence,

SMC pricing dictates higher tolls when demand is high and lower tolls when demand is low. The implication for practical application of SMC pricing is that tolls should be higher during peak periods and peak seasons than during off-peak periods and seasons and also, that tolls should rise over the lifetime of a facility as demand increases.

4.3 PRICING POLICIES IN ACTION

As demonstrated, social marginal cost (SMC) pricing is the most efficient mechanism for regulating use of fixed airport facilities, since it ensures that usage is allocated only to those who value usage (in terms of their willingness to pay) at least as much as the social marginal cost of their usage. Charging a user social marginal cost means charging the user those social costs that would be avoided if the user did not use the facility. As applied to landing fees, SMC pricing therefore comprises three elements: the marginal airport operating cost of serving an aircraft, the marginal noise cost that the aircraft generates and the marginal congestion (delay) cost that the aircraft imposes on other aircraft.

The pattern of landing fees across aircraft types and time periods dictated by SMC pricing results from the relative magnitudes and variation by aircraft type and time period of each of the three elements of social marginal cost. Analysis of social marginal costs at Pearson International Airport indicates that marginal airport operating cost is small relative to the other two social marginal cost elements and is relatively constant (in the range of $5–$10) across aircraft types and time periods. Marginal noise cost varies by aircraft type (in the range $25–$200) and may vary to a lesser extent by time of day. Marginal congestion cost varies primarily by time of day (and season), and to a lesser extent by aircraft type, with large congestion costs (in the range $130–$220) during peak times and negligible congestion costs during off-peak times. Marginal congestion costs for light aircraft are almost as high as those for heavy aircraft, because a light aircraft occupies a runway for almost as long as a heavy aircraft, and the opportunity cost of a minute of runway time during a congested period is the same for all aircraft. The opportunity cost of runway time is high during a congested period because it factors in the high cost of delaying a heavy aircraft.

For congested airports, SMC pricing therefore implies a regime of landing fees characterized by a base fee equal to marginal airport operating and noise cost that applies during all times and varies by aircraft type,

supplemented by an additional marginal congestion cost fee during peak times (and seasons) that varies somewhat by aircraft type. From the point of view of alleviating airport congestion, the key feature of SMC pricing is the large fee differential between peak and off-peak use for all aircraft.

Efficient allocation of airport resources is not the sole objective of government policy. As discussed earlier in this report, Transport Canada's proposed cost- recovery policy for airports sets the goal of recovering airport capital and operating costs. Under a cost-recovery constraint, an airport pricing system performs the dual role of allocating airport resources efficiently and generating the revenue required to recover airport capital and operating costs. Social marginal cost pricing automatically recovers airport operating costs because they are a social cost that varies with marginal airport use. Because SMC pricing deals with the problem of allocating airport resources efficiently in the short run — that is, given a fixed level of airport capacity — it does not charge users directly for the capital costs incurred in providing capacity.[47] Rather, users are charged marginal congestion cost, which varies with the level of traffic. Under SMC pricing, the airport authority must look to the revenue from congestion (peak-period) charges to recover capital costs and fund capacity expansion.[48]

The issue of whether the revenue generated from SMC pricing is in general sufficient to cover the cost of capacity investment has been examined from a theoretical perspective. The cost-recovery theorem developed by Mohring and Harwitz (1962, 1970) proves that, under certain conditions, the revenue generated by optimal congestion tolls will exactly equal the capital cost incurred by optimally timed investment.[49] These conditions are that capacity expansion be perfectly divisible, and characterized by constant returns to scale. Since airport terminal and runway capacity investments are typically highly indivisible (lumpy), there is no theoretical guarantee that, in general, the revenue generated by SMC pricing will exactly cover airport operating and capital costs. Depending on the level of congestion of an airport, the revenue from congestion fees may under-recover, exactly recover or over-recover capital cost.

Whether the revenue from marginal congestion cost fees is sufficient to offset capital costs clearly depends on the level of congestion at an airport. However, to determine whether capital costs will be recovered at a particular airport, it is not sufficient to compare annualized capital cost to the level

of congestion fees in a single year, since the level of congestion, and hence the annual revenue from congestion fees, rises as traffic demand grows over time. Once the level of congestion reaches a critical level (when annual congestion fee revenue exceeds the opportunity (interest) cost of expansion capital), capacity expansion is justified; with expanded capacity, the level congestion and congestion fees returns to a low level. As Oum and Zhang (1990) have pointed out, whether the total revenue from congestion fees over an investment cycle (the time between initial and subsequent capacity expansion) is sufficient to recover capital cost depends on the time path of traffic demand growth.[50] For a given average demand growth rate, capital costs are more likely to be recovered if demand (and hence congestion) grows rapidly at the beginning of the investment cycle and then levels off, than if demand grows slowly at the beginning of the investment cycle and rapidly at the end.

Using a simulation model with demand and capacity parameters at Pearson International Airport, and assuming an average annual demand growth rate of 3.5 percent, Oum and Zhang found that the revenue from marginal congestion fees recovered capital costs, regardless of the time path of traffic growth. For an initial capacity of two runways, even the most pessimistic assumption regarding the pattern of traffic growth (slow growth in traffic in early years followed by acceleration in later years of the cycle) leads to (bare) cost recovery, generating a cost-recovery ratio of 1.01. More optimistic assumptions regarding the pattern of traffic growth generated financial surpluses, with cost-recovery ratios in the range 1.1 to 1.3. They also found that the cost-recovery ratio increased with the level of initial capacity, ranging as high as 2.3 for an initial capacity of four runways. The positive relationship between airport capacity and cost recovery of capacity increments is attributed to the fact that, at larger airports, capacity increments represent smaller percentage increments in capacity. With capacity expansion less lumpy in percentage terms, traffic grows sufficiently for congestion to re-emerge sooner after capacity expansion at larger airports than at smaller airports, with the result that larger airports generate congestion fee revenue sufficient to recover capacity expansion costs more rapidly than do smaller airports.

The implication of Oum and Zhang's findings is that for major airports with demand and capacity conditions analogous to those at Pearson, SMC pricing is likely to recover airport capital as well as operating costs. However, this

finding is not applicable to significantly smaller airports where the relative lumpiness of capacity expansion and/or slow demand growth make cost recovery through congestion fees unlikely. Therefore, the conclusion is that marginal congestion cost peak fees, and hence SMC pricing, are likely to recover capital costs at large, congested airports, but unlikely to recover capital costs at small or uncongested airports.

Where SMC pricing does not recover costs, but cost-recovery is required, it is necessary to diverge from SMC prices to recover the revenue shortfall generated by SMC pricing. The most efficient such divergence is Ramsey pricing which, in the absence of externality (for example, congestion, noise) costs, differentially marks up prices above the airport's private marginal (operating) cost in inverse proportion to the demand elasticities of separable user segments. In the more general case where externality costs are present, Ramsey pricing marks up prices differentially over private marginal costs and a fraction of marginal externality costs. In both cases, Ramsey pricing recovers costs efficiently by marking up prices above marginal cost proportionally more for users who value usage more and less for users who value usage less. This ensures that the amount of traffic "choked off" by the mark-ups, and hence the efficiency loss due to divergence from SMC prices, is minimized.

Since Ramsey mark-ups vary in inverse proportion to the price elasticity of user demand, in practical terms Ramsey pricing implies landing fees that vary primarily by aircraft type, with larger aircraft charged higher fees than smaller aircraft. Ramsey prices may also vary by time of day, since an aircraft of a given type may have more inelastic demand for peak-period use than for off-peak use. Demand elasticities, and hence Ramsey prices, can also vary by stage-length (length of flight) and type of use (commercial, general aviation, military or government).

At major congested airports, SMC pricing will probably recover capital costs without the need to resort to the second-best Ramsey pricing policy. At non-major airports SMC pricing will not fully recover capital costs, in which case Ramsey pricing represents the most efficient way to recover the revenue shortfall. Therefore, in practical terms, the essential features of the landing fee policy dictated by efficient allocation of airport resources are:

- landing fees at major airports characterized by a large fee differential between peak and off-peak times of day (and seasons) for all types of aircraft, with relatively small variation in peak fees by aircraft type (reflecting differential runway occupancy times), and larger variation in off-peak fees by aircraft type (reflecting differential noise costs); and
- landing fees at non-major airports characterized by large variation in fees by aircraft type, and smaller variation in fees for each aircraft type between peak and off-peak periods.

Transport Canada's proposed cost-recovery policy for airports provides a basis for airport pricing that differs *conceptually* from the basis provided by efficient allocation of airport resources. Because Transport Canada's proposed policy is concerned primarily with cost recovery rather than with efficiency, it dictates that prices be based on direct allocation of variable and *fixed* costs to users, rather than based on the social *marginal* costs that airport users generate. The policy is characterized by:

- fees determined by allocating site-specific operating and capital costs to users;
- capital costs of runways allocated to commercial, state and military aircraft only (not to general aviation);
- airfield operating and maintenance costs allocated to all users on a per tonne of maximum take-off weight basis;
- terminal building capital and operating costs allocated to airlines, concession operators and passengers on the basis of usage; and
- peak-period fees at *major* airports that recover capital and operating costs in excess of those that would be incurred if demand were evenly distributed throughout normal operating hours of the day and throughout the year.

Although conceptually different from SMC and Ramsey pricing, in practice, the structure of landing fees suggested by Transport Canada's (TC's) proposed pricing policy is similar to that implied by SMC and Ramsey pricing. For non-major airports, TC's cost-recovery policy does not propose peak-period fees, but rather fees that vary on the basis of aircraft type both explicitly, in the provision that runway capital costs not be allocated to general aviation, and implicitly, by allocating airfield operating costs to users on the

basis of maximum aircraft take-off weight. This variation in landing fees by aircraft type is consistent with Ramsey pricing. For major airports, TC's cost-recovery policy dictates variation in landing fees by aircraft type during off-peak periods, supplemented by a peak-period fee that allocates part of capital costs to peak users. Provided that the peak fee is to be applied to all peak users, TC's policy for major airports is structurally similar to SMC pricing.

The structural resemblance between TC's proposed pricing policy and SMC/Ramsey pricing indicates that TC's policy recovers costs in a relatively efficient manner. However, how efficient it will be depends to some extent not only on its qualitative similarity but its quantitative similarity to SMC/Ramsey pricing. In particular, the TC policy dictates a peak/off-peak fee differential based on allocation of capital costs between peak and off-peak use, whereas SMC pricing bases the peak/off-peak fee differential on the different criterion of the marginal congestion cost differential between peak and off-peak times.

A specific comparison between peak fees under TC's proposed policy and peak fees generated by SMC pricing is not possible at present, due to the lack of an explicit formulation of TC's proposed peak fees. However, a more general comparison of magnitudes is possible. TC's peak/off-peak fee differential will likely be less than that generated by SMC pricing because cost-recovering SMC pricing allocates essentially all of capital cost to peak users (who generate the majority of marginal congestion costs), while the TC policy will partition capital costs between peak and off-peak users. In terms of the degree of peak pricing, the proposed TC policy, therefore, represents an intermediate state between the current pricing system, which does not differentiate prices by time-of-day, and the peak period pricing suggested by SMC pricing.

Exactly how close TC's proposed policy is to SMC pricing may not, in any case, be the key consideration in assessing its relative efficiency. Borins (1984) simulated the dynamic effects of various non-SMC pricing policies on the economic surplus generated by Pearson International Airport and found that "the social welfare surface surrounding the optimal policy [SMC pricing] is relatively flat for a substantial range, and that existing policies are also on this range . . . the welfare surface resembles a broad plateau . . . rather than peaking very sharply at the optimum resembling a mountain like the Matterhorn."[51] In concrete terms, Borins found that "the relative deviations

[from the level of economic surplus induced by SMC pricing] of non-optimal pricing policies are quite small, always less than 1 percent for the low elasticity model and less than 5 percent for the high elasticity model."

These findings indicate that the efficiency loss, in terms of short-run use and long-run timing of investment, of pricing policies that lie between the existing policy and the optimal (SMC) policy is small in relative terms. The key consideration in assessing TC's pricing policy is not whether it corresponds exactly to the optimal SMC/Ramsey pricing policies, but rather that it represents an improvement over the existing policy that moves in the direction of optimal pricing. TC's policy approximates the price structure implied by SMC and Ramsey pricing by recovering costs in a manner that encourages efficient allocation of airport resources by discouraging use of congested facilities by low-valuing users during peak times, and allowing for use of uncongested facilities and off-peak times by all users.

5. A Framework for Efficient Planning and Decision Making

So far in this report, we have examined the airport planning process from two perspectives. First, the long-run problem is one of making investment decisions that maximize return on capital. Second, the short-run problem is one of efficient use of existing capacity through appropriate pricing schemes. In theory, this dichotomy between short-term and long-term decisions (or policy traces) should not exist, but, in practice, it is quite prevalent. Airport administrators are given the responsibility of operating existing facilities within certain policy and management guidelines. They function within the realm of operating budgets, planning revenues and expenditures to meet certain financial objectives. In the short run, excess demand is handled through rationing or, if possible, by improving the operating efficiency of existing facilities to increase throughput. The longer-term considerations regarding capacity expansion are dealt with in the realm of capital spending. Such decisions are subject to public investment guidelines, and are rarely seen as an extension of the airport's operating strategy. They present themselves as discreet events, which have tended to be in response to long-term growth projections in the past and, in more recent times, to severe congestions.

A rational model, based on sound economic principles, has to deal with the management of existing and construction of new facilities in an integrated framework. Although a distinction is drawn with respect to time horizons,

investment and pricing both serve the same objective of efficient allocation of resources. The lumpy nature of airport investments may pose problems in the derivation of cost functions, which, in turn, may pose difficulties in determining efficient prices. In practice, however, sound application of cost-benefit principles in investment decisions and social marginal cost principles in pricing decisions, forms a common basis that provides adequate levels of capacity and allows for that capacity to be utilized.

Investment and pricing policies are evaluated against the criterion of economic efficiency. An efficient policy is one that maximizes net benefits to society. In the context of pricing policy, economic efficiency dictates operating an airport at the level of use where the social benefits of expanding use are just exceeded by the incremental social costs, including the cost of incremental congestion. Social marginal cost (SMC) pricing holds use and congestion to this efficient level by rationing use among users on the basis of their willingness to pay the social marginal cost. As users' willingness to pay (demand) increases, the efficient level of use and congestion increases, and the social marginal cost price rises. When congestion reaches a critical level, the congestion cost savings that can be achieved by expanding capacity outweigh the capital, and other social costs, associated with capacity expansion. Capacity expansion then is justified because it provides positive net social benefits. By controlling congestion growth, pricing policy affects the timing of investment. A pricing policy that diverges from social marginal cost can therefore lead not only to inefficient use of fixed capacity but also to inefficient timing of capacity expansion.

Under a cost-recovery constraint, pricing policies not only affect the timing of investment, but also provide the revenue required to fund investment. Social marginal cost pricing, because it recovers congestion — not capital — costs, will not necessarily recover the capital cost over a period of capacity expansion. However, for major Canadian airports, any revenue shortfall from SMC pricing is likely to be relatively small, and in some instances could lead to a revenue surplus.

If cost recovery is to be achieved, the alternative to SMC pricing is pricing based on allocating capital (and operating) costs to users. In Transport Canada's cost-recovery policy, landing fees vary among different aircraft types on the basis of maximum aircraft take-off weight rather than the amount of runway time used and congestion induced. A potential inefficiency associated

with capital cost-based pricing is that prices do not necessarily rise with increased congestion as SMC prices do and, therefore, do not signal the relative scarcity of capacity to users. However, the relative loss of social welfare associated with these inefficiencies would be small for major Canadian airports. The argument for capital cost-based pricing is further strengthened if different prices are charged for peak and off-peak use, based on some estimate of the proportion of capital costs necessitated by peak use. Such a peak/off-peak fee differential performs a similar function to SMC pricing in shifting low-valued users from peak to off-peak times to use capacity more efficiently and delay the need for capacity expansion. In general, the proposed cost-recovery package provides a practical framework that would yield fairly efficient pricing practices at Canadian airports.

In investment policy, economic efficiency dictates the comparison of the social costs and benefits of potential capacity expansion options. The established tool for conducting this exercise is cost-benefit analysis (CBA). There are some technical difficulties in CBA, such as the need to estimate demand curves to obtain an adequate welfare measure of benefits and costs, and the need to choose a social discount rate to compare benefits and costs that accrue at different times. The use of CBA as an evaluation tool raises a number of issues that require further consideration in the airport planning context.

First, the timing of cost-benefit studies, as well as of actual investments, requires special attention. The Vancouver and Toronto studies both revealed that runway capacity expansions were overdue. Severe congestion has been imposing serious costs on airport users, which should have been detected earlier. As long as new investment considerations are delayed until congestion levels reach intolerable (or even noticeable) levels, appropriate remedies will always be implemented too late. This is exacerbated by the long lead times required for planning new capacity, and by the fact that congestion costs rise exponentially as capacity use approaches maximum physical capacity.

The high costs associated with overdue investment are demonstrated by the results of the Pearson study, which found the first-year benefit of two new runways in 1996 to be $140 million, as compared to a total capital cost of $469 million.[52] Assuming a social discount rate of 10 percent, in 1990 present-value dollars, the first-year benefit is $79 million and the capital cost $327 million. The cost of overdue runway expansion at Pearson is therefore approximately $46 million per year. This figure represents delay

cost savings foregone by postponing runway expansion by one year ($79 million) less the opportunity (interest) cost of capital that is saved by postponing expansion by one year ($32.7 million).

A more general measure of the extent to which a project is overdue is the "first-year benefit ratio," the ratio of the first-year benefit to the total capital cost. If a project is overdue, the first year benefit exceeds the opportunity cost of capital (the discount rate times the total capital cost), and hence the first-year benefit ratio exceeds the discount rate. An optimally timed project is indicated by a first-year benefit ratio equal to the discount rate, and a premature project by a first-year benefit ratio less than the discount rate. In the case of Pearson runway expansion, the first year benefit equals 79/327 or 24%. Since the 24% return on capital that runway expansion produces in its first year of operation exceeds the 10% return on capital assumed in an alternate use, postponing the project would lead to a decrease in project net present value, and the project can therefore be deemed overdue. In relative terms, runway expansion at Vancouver was found to be even more overdue than at Pearson, with a reported first-year benefit ratio of 82% for the recommended runway option.[53]

The second problem with the application of cost-benefit principles to airport investment concerns the range of investment options considered in cost-benefit studies. In the case of Toronto, for example, investment decisions at the Lester B. Pearson International Airport have to be examined in the context of the regional airport system. Since there may be opportunities for diverting traffic to other airports (for example, Hamilton, Toronto Island or Buttonville), and the construction of a second stand-alone airport always remains a consideration, investments at Pearson are difficult to evaluate in isolation from other components of the airport system. This difficulty was partially overcome in the cost-benefit study by arguing that proposed airside development options were medium-term solutions, which could not be substituted by diversion or relocation of traffic to other airports. Although this constitutes a reasonable argument, it superficially limits the time horizon over which medium-term airside development options are evaluated.

In the Pearson case, the benefits and costs of options were evaluated over only 15 years from the year of implementation. Although the need for runway expansion was strong enough to justify expansion over the 15-year horizon in this case, in general this is an unreasonably short period of time

over which to evaluate the economic viability of airport investment options. A more systematic approach capable of integrating evaluation of medium- and long-term airport investment options is therefore required. The Vancouver study used a 30-year time horizon and considered the development of alternate airports as an option. However, the percentages of different types of traffic that would divert to alternate airports, while partially based on access time-cost differentials, were largely assumed rather than being predicted by a multi-airport system model.[54]

The third problem that plagues the application of cost-benefit analysis to major airports relates to uncertainty associated with forecasts of benefits and costs. As noted earlier in this report, there have been times when official air traffic forecasts did not materialize, eliminating the need for additional capacity, or exposing examples of excess capacity (for example, Mirabel). Forecasting will always remain an uncertain art. The only practical solution is to evaluate investment options against different growth scenarios, and to look at alternative time horizons when considering each investment option. Such sensitivity analysis is an effective means of overcoming the inherent uncertainty attached to underlying parameters and assigning a degree of confidence to the results, as long as variation of underlying parameters over reasonable ranges does not affect the ranking of options. Sensitivity analysis also identifies important demand and other parameters that can significantly affect the results if they vary beyond critical ranges. Decision makers or experts can then assess the probability that these sensitive parameters will fall outside of critical ranges.

A source of uncertainty which is difficult to deal with is congestion delays under differing traffic levels and capacity options. This uncertainty results not only from the unknown delay properties of expanded infrastructure and forecast traffic volumes but also from inadequate knowledge of actual delays incurred on existing capacity by current levels of demand. The studies of Pearson and Vancouver airports both faced the problem of a lack of historical data on congestion delays. Both studies used discreet (aircraft by aircraft) simulation models of airfield operations to estimate delays, not only for capacity expansion options, but also for base-case (existing capacity) options. As long as delay forecasts are based solely on the predictions of simulation models, significant uncertainty about the delay forecasts will remain. Delay forecasts could be more accurate if more reliable data on actual delays were available to verify the delay predictions of simulation models.

Most of the problems associated with airport benefit-cost analysis can be alleviated by adopting a framework that facilitates continuous monitoring of costs and benefits. A social benefit and cost monitoring system would:

- Provide the data required to time cost-benefit studies optimally and to increase the level of confidence associated with study forecasts, and would provide the basis for an airport planning model that could incorporate multiple airports.
- Deal with incidents of externalities and other distributional issues. User costs and benefits could be compared to external impacts, such as neighbourhood noise costs or spin-off community benefits. The system could be used to disseminate information to the public, and to facilitate dialogue between opposing interests. It could serve as an effective means of mediating between conflicting interests, or at least provide grounds for compromise.
- Serve a useful purpose in pricing decisions by providing the data required for examination of the efficiency and revenue implications of alternative pricing schemes. Data on the social costs of airport operation could also be used to quantify externalities such as noise to provide the basis for taxation and compensation schemes.

In conclusion, from an economist's perspective we see the planning process as a continuous and comprehensive cost-benefit analysis. We see benefit-cost analysis not only as a "technocratic instrument" for investment appraisal, but an effective framework within which external impacts can be monitored and quantified, political compromises among different interest groups can be reached, and appropriate compensation mechanisms can be devised to make all concerned parties better off. The discipline of the new environmental review and assessment procedures has imposed the need for rigorous cost-benefit analysis in evaluating large projects, such as the international airports in Toronto and Vancouver. We propose that the scope of these cost-benefit studies be expanded to become a framework for continuous monitoring of airport costs and benefits, which we believe would constitute the basis for more rational decisions with respect to the economic efficiency of the airport system.

ENDNOTES

1. See S. Borins, "Organization with Environment: A Review of Walter Stewart's *Paper Juggernaut*," *Canadian Public Policy* 6, 1 (Winter 1980), pp. 115–23 and S. Borins, "Self-Regulation and the Canadian Air Transportation Administration: The Case of Pickering Airport," in *Studies on Regulation in Canada*, ed. W. T. Stanbury (Montreal: Institute for Research on Public Policy, 1978), pp. 131–51.

2. See Transport Canada, Airports Authority Group, *Our First Year: 15 October, 1985–15 October, 1986* (Ottawa: Supply and Services Canada).

3. Geoffrey Rowan, "Community airports head for fast lane," *The Globe and Mail*, October 2, 1991, pp. B1, B6.

4. Transport Canada, Airside Capacity Enhancement Project Team, *Vancouver International Airport, Airside Capacity Enhancement Project, Airside Demand/Capacity Analysis* (June 1989).

5. Transport Canada, *Proposed New Cost Recovery Policy: Phase II* Discussion Paper, TP10041 (April 1990).

6. Transport Canada, Airports Authority Group, *Business Plan Framework, 1991/92–1993/94.*

7. For a review of indivisibilities in airport capacity, see T. H. Oum and Y. Zhang, "Airport Pricing: Congestion Tolls, Lumpy Investment, and Cost Recovery," *Journal of Public Economics* 43 (1990), pp. 353–74.

8. The latter applies, for example, to the valuation of externalities such as noise, which are not subject to property rights and hence have no established market price. Where market prices of costs and benefits do exist, they are used because they reflect the maximum that individuals are willing to pay in the presence of the market.

9. This principle was first introduced by N. Kaldor in "Welfare Propositions of Economics and Interpersonal Comparisons of Utility," *Economic Journal* 49 (1939). The compensation principle is a relaxed version of the Pareto principle, which requires that a policy make at least one person better off and no one worse off without compensation.

10. Equity objectives of Canadian transport policy are outlined in Royal Commission on National Passenger Transportation, *Getting There: The Interim Report of the Royal Commission on National Passenger Transportation* (Ottawa: Supply and Services Canada, April 1991), p. 221.

11. For a review of methods for incorporating distributional weightings into the evaluation criterion, see W. G. Waters, II, "Investment Criteria and the Expansion of Major Airports in Canada," *Canadian Public Policy* 3 (1977), pp. 23–35.

12. See, for example, Transport Canada, *Toronto Lester B. Pearson International Airport Airside Development Project, Final Report No. 24, Benefit/Cost Analysis*, TP10854E (April 1991), report prepared by Transmode Consultants Inc.; and Hickling Corporation, Economic Analysis of Airfield Capacity Enhancement Strategies for Vancouver International Airport (1990).

13. By NPV we refer in all cases to net present value evaluated in the current year, regardless of the start-date considered. Delaying the start date of an option will increase its NPV if the interest savings on expansion capital exceed the foregone net benefits of expansion.

14. See Oum and Zhang (1990).

15. If the capacity expansion project is to be evaluated using a fixed, finite economic life of n years, rather than a fixed horizon date or an infinite horizon, then there will be an additional benefit of delaying capacity expansion by one year: the net benefit in year n+1. The n+1 year benefit is not considered here for simplicity of exposition, and consistency with the literature (for example, Oum and Zhang (1990) use an infinite horizon). For applications where consideration of the n+1 year benefit is justified, use of the formula presented below understates the benefits of delaying expansion, and hence leads to later than optimal capacity expansion. Since the n+1 year benefit will normally be heavily discounted, the magnitude of the understatement is likely to be small. Simulation methods can be used to determine the exact optimal timing.

16. This assumes that increases in congestion cost savings over time are greater than increases in externality (noise) costs.

17. For discussion of this criterion see S. A. Marglin, *Approaches to Dynamic Investment Planning* (Amsterdam: North-Holland, 1963), and S. F. Borins, "The Effect of Pricing Policy on the Optimal Timing of Investments in Transportation Facilities," *Journal of Transportation Economics and Policy* 15 (1981), pp. 121–33.

18. If the expansion option requires a gestation period for construction before benefits are realized, then K is the capital cost compounded by r over the gestation period.

19. In the latter case, the optimal postponement period will be the longest period over which the average rate of increase of annual net benefits exceeds the discount rate. See E. J. Mishan, *Cost-Benefit Analysis* (London: George Allen & Unwin, 1982), p. 269.

20. Transport Canada, *Toronto Airside Development Project* (1991) and Hickling Corporation, *Economic Analysis of Airfield Capacity Enhancement Strategies for Vancouver International Airport* (1990).

21. If only a single expansion option is being evaluated, and if benefits are monotonically increasing over calendar time, and a long planning time horizon is in use, then this condition for optimal timing is not only necessary but also *sufficient* to recommend expansion. If the single option is optimally timed, so that its first-year benefit exceeds rK, then the present value of its benefits over all future time exceeds K (since the present value of a perpetuity of rK is 1/r·rK = K), and hence the option automatically has a positive net present value and can be recommended for implementation without explicitly calculating its NPV. Thus, conceptually, it is the optimal timing rule which forms the foundation for investment evaluation, with NPV analysis required to handle the more general case when more than one investment option is to be evaluated.

22. Hickling Corporation, *Economic Analysis of Airfield Capacity Enhancement Strategies for Vancouver International Airport* (1990).

23. Transport Canada, Airside Capacity Enhancement Project Team, *Vancouver International Airport, Airside Capacity Enhancement Project, Airside Demand/Capacity Analysis,* TP 9411E, (June 1989).

24. Hickling Corporation, *Economic Analysis of Airfield Capacity Enhancement Strategies for Vancouver International Airport* (1990), p. 8.

25. Federal Environmental Assessment Review Office, *Vancouver International Airport Parallel Runway Project: Report of the Environmental Assessment Panel* (1991), p. 40. The Panel also noted that inclusion of land costs would not have changed the study results.

26. Transport Canada, *Vancouver International Airport, Airside Capacity Enhancement Project* (1989), pp. 3–1, 3–3.

27. Federal Environmental Assessment Review Office, *Vancouver International Airport* (1991), p. 40.

28. In addition to this high benefit to cost ratio, the internal rate of return of the recommended option was also high at 114 percent.

29. Hickling Corporation, *Economic Analysis of Airfield Capacity Enhancement Strategies for Vancouver International Airport* (1990). It should be noted that the degree of confidence that can be placed in such a statement depends on the accuracy of the subjective probability distributions assigned to underlying assumptions. Risk analysis of the type conducted in the study does not eliminate uncertainty: it addresses one level of uncertainty but creates another.

30. An average delay of two minutes was forecast to remain in the first year after implementation of the runway (1993), which was found to be sufficient to generate incremental benefits to peak-period pricing. Hickling Corporation, *Economic Analysis of Airfield Capacity Enhancement Strategies for Vancouver International Airport* (1990), pp. 112, 116.

31. Ibid.

32. The FYBR of 82% implies that the annual cost attributable to overdue runway expansion in the presence of a $100 peak fee is approximately $44 million.

33. Transport Canada, *Toronto Airside Development Project* (1991).

34. Transport Canada, *Economic Impact of Alternative Traffic Management Options for Lester B. Pearson International Airport,* TP10668E (March 1991), report prepared by Hickling Corporation.

35. Estimation of benefits was made even more conservative by including only the benefits of using versus not using the airport, not the benefits of using peak versus off-peak times, in the calculation of generated user benefits.

36. No assessment of the allocative efficiency of administrative allocation versus congestion pricing is implied.

37. Transport Canada, *Benefit-Cost Model for Airport Approach Systems,* TP6887E (September 1986).

38. The analysis assumes that the demand curve is fixed and hence applies individually to periods of relatively constant demand.

39. For the purposes of this analysis, our assumption ignores the macroeconomic spin-off benefits that accrue from airport operation.

40. MC_1 is the passenger time and operating cost of using the airport of the average aircraft, and reflects the assumption of a constant aircraft mix over utilization levels.

41. The term "slot" generally denotes the right to land or take off (not both) once during a specified hour, daily, for one winter or summer season. S. G. Hamzawi, "Methods to Relieve Airport Congestion," *Proceedings of the Canadian Transportation Research Forum*, (1988).

42. D. W. Gillen, T. H. Oum and M. W. Tretheway, "Airport Pricing and Capacity Expansion: Economic Evaluation of Alternatives," *Transportation in Canada* (Ottawa: Transport Canada, 1990) TP10451E, p. 86.

43. For example, consider the case of an airport served by two airlines with the same operating configuration, costs and passenger demands. If the airport manager sets a peak-period traffic quota and informs the two airlines that they will not be permitted to fly until they work out a schedule that conforms to the quota, the two airlines will probably agree to split the quota in half and to each half assign their highest valued flights, producing an efficient allocation of the quota. In game theoretic terms, such an agreement represents a focal equilibrium.

44. The market clearing price will equal the marginal users' valuation less their PMC. Hence, assuming the airport authority knows the PMC curve, PMC can be added to price to obtain marginal valuation.

45. We assume here that the airport authority knows the SMC curve and that the demand curve is constant.

46. This algorithm is presented as an illustration of the use of market prices to guide quota setting. Although this algorithm does not necessarily converge (consider steepening the demand curve in Exhibit 5), other algorithms based on the same price information could be designed that would converge.

47. The practice of "long-run marginal cost" pricing — charging users for the "marginal" cost of capital as well as the marginal cost of airport operation, noise and (depending on the formulation) congestion — is not efficient. As Gillen, Oum and Tretheway (1990) note, with lumpy capacity investment "there is no smooth long run marginal cost curve . . . the long-run marginal cost curve is the collection of short run marginal cost curves." Therefore, a policy that allocates capital costs to users is not necessarily efficient. The efficient pricing policy charges users the short-run social marginal cost at the prevailing traffic level. This implies "charging higher prices [due to higher congestion] before any capacity investment, and lower fees after capacity is in place." Charges that vary over time with the level of congestion signal the relative abundance or scarcity of capacity to users and, hence, are more efficient than charges that allocate to users a fraction of capital cost that remains constant over time. See Gillen, Oum and Tretheway (1990), p. 91.

48. We ignore the possibility of using marginal noise cost fees to recover capital costs for the sake of simplicity, and because they may be required to fund householder noise compensation schemes.

49. H. Mohring and M. Harwitz, *Highway benefits: An analytical framework* (Northwestern University Press, 1962) pp. 84–86, and H. Mohring, "The Peak Load Problem with Increasing Returns and Pricing Constraints," *American Economic Review* 60 (1970), pp. 693–705.

50. The optimal timing of capacity expansion is itself determined by the level, but not the growth path, of traffic demand. See Oum and Zhang (1990), pp. 353–74.

51. S. F. Borins, "The Economic Effects of Non-optimal Pricing and Investment Policies for Substitutable Transport Facilities," *Canadian Journal of Economics* 17 (1984), pp. 80–97.

52. Transport Canada, *Benefit-Cost Analysis of Airside Development at Lester B. Pearson International Airport* (1991), pp. 36, 61.

53. Hickling Corporation, *Economic Analysis of Airfield Capacity Enhancement Strategies for Vancouver International Airport* (1990), p. 174. Hickling defined the first-year benefit ratio as the ratio of "all benefits in the first year after commissioning a project divided by the total costs incurred to that date, including interest" (p. 172). This formulation is equivalent to the ratio of present valued first-year benefit to present valued capital cost that we present. The first-year benefit ratio, a measure of project timing, should not be confused with the internal rate of return (the discount rate at which NPV equals zero), which is a measure of the rate of return of a project over its entire economic life and, hence, a measure of project merit.

54. Ibid., pp. 54, 63, 91.

19

TRAVEL DEMAND BEHAVIOUR: SURVEY OF INTERCITY MODE-SPLIT MODELS IN CANADA AND ELSEWHERE

Eric J. Miller and Kai-Sheng Fan*
December 1991

1. INTRODUCTION

1.1 STUDY PURPOSE

The purpose of this study is to review the evolution of the intercity travel demand modelling state-of-the-art over the past 20 years, within Canada and, as applicable, elsewhere. In particular, the review summarizes the lessons which have been learned from this modelling work concerning Canadian intercity travel in general and intermodal substitutability in particular.

Given the emphasis on modal substitutability, this review focusses on the choice of mode in the travel demand modelling process. Trip generation/ distribution components are discussed, but these aspects of the overall modelling process are not reviewed in detail. In particular, the issue of the "induction" of new, previously unmade trips through the introduction of service improvements on one or more modes is not comprehensively explored.

* Department of Civil Engineering, University of Toronto.

This review is not intended to assess the relative accuracy of the various Canadian intercity demand models which have been developed and applied over the years. Rather, it is intended to assess what has been learned from these models that is applicable to understanding and forecasting Canadian intercity travel demand. Thus, the study spends virtually no time reviewing the actual forecasts generated by these models. Instead, it focusses on the elasticities, values of time and other fundamental indicators of travel behaviour that can be extracted from these models. In so doing, the review also inevitably deals with methodological issues associated with the specification, estimation and application of these models, since the empirical results (elasticities, etc.) obtained from these models can only be evaluated within the theoretical and methodological context within which they are obtained.

The primary focus of the review is necessarily on Canadian intercity mode-split models. Every Canadian multimodal[1] model of consequence reported in the literature is reviewed in this report. U.S. models of significant relevance to the Canadian context, in particular those which appear to represent the current state of practice, are also reviewed in detail. Non-North American models are generally not reviewed in detail, typically due to a lack of transferability to the North American context and/or a lack of detailed information concerning the models' functional forms, assumptions, etc.

Time and resource constraints and, more particularly, lack of access to original model data sets prevented any new analysis from being undertaken within this study. Given the complex, non-constant nature of the elasticities associated with virtually every model reviewed here, this review contains only empirical elasticities which have been reported in the models' documentation, since sufficient information to compute meaningful elasticities given a model's functional form and estimated coefficients was rarely available.

1.2 REPORT ORGANIZATION

Section 2 of this report provides a brief background discussion of several issues that are of particular importance in assessing the state-of-the-art of intercity travel demand. In particular, the fundamental issue of choice of model aggregation level is discussed in some detail. Primarily motivated by this discussion of the aggregation issue, the review of specific intercity mode-split models is presented in two sections. Section 3 deals with aggregate modelling efforts, while Section 4 deals with (typically more recent)

disaggregate models. Finally, Section 5 summarizes the findings of this detailed review with respect to empirical results, methodological issues and directions for further model development.

2. Discussion of Issues

2.1 INTRODUCTION

This section discusses several issues which have a direct impact on the results which are obtained from intercity passenger travel demand models and, hence, must be considered in any evaluation or discussion of these models. These issues include:

- spatial aggregation level and model transferability;
- travel market definition; and
- model specification.

2.2 LEVEL OF AGGREGATION

Travel demand models are typically developed at one or the other of the following two levels of *spatial aggregation*:

- the *aggregate* level, in which total trips (by mode) between zones are modelled directly; and
- the *disaggregate* level, in which trips by individuals are modelled directly (and the aggregate zone-to-zone flows required for policy analysis are then generated by explicitly or implicitly adding up all the trips made between these zones by these individuals).

By far the majority of intercity passenger travel demand models fall into the aggregate category, although in the last 5 to 10 years, models have been at least partially disaggregate in nature. Aggregate models typically possess several practical advantages. In particular, their input data requirements are generally much more modest than those of disaggregate models and are also generally more consistent with the information which often has been available for model construction. Further, such models are generally easier to apply, since they are developed directly at the level of policy interest, that is, the level of city-to-city flows.

Aggregate models, however, can be criticized with respect to several aspects.[2] The most fundamental of these is that aggregate models inherently run the risk of having incorporated within them unknown amounts of *aggregation bias*. Figure 2-1(a) illustrates the concept of aggregation bias. In this figure, a hypothetical demand curve is assumed. The precise nature of the curve is not of immediate importance, except that it is non-linear in nature (certainly not an unreasonable assumption). The demand curve shown expresses the probability of an individual choosing a particular mode of travel for a trip as a function of the person's income, where all other factors affecting this modal choice (modal levels of service, trip purpose, etc.) are assumed to be held constant. Two individuals (1 and 1'), possessing very different income levels are shown (I_1 and $I_{1'}$), along with their mode choice probabilities (P_1 and $P_{1'}$). The average income for these two individuals (I) and their average choice probability (P) are also shown. Points to note from this figure include the following:

- A disaggregate model would attempt to reproduce the demand curve shown in Figure 2-1(a) by statistically relating the observed response of each individual trip-maker to his/her individual characteristics.[3] Such a model will inevitably contain some error, due to the use of "approximate" functional forms, omission or mismeasurement of explanatory variables, etc. But such a model, if properly constructed, will not be inherently biased in terms of its model parameter values.
- An aggregate model, on the other hand, would typically be developed by statistically relating the observed *average* response for an aggregation of trip-makers as a function of the *average* characteristics of these trip-makers; that is, point {I,P}, combined with comparable points for other groups, that is, average values for other zones or zone-pairs, as the case may be.
- Point {I,P} does *not* lie on the true demand curve and, in general, will not lie on the curve unless the curve is linear (a very unlikely event). Thus, any model based on aggregate data such as {I,P} will not likely be able to reproduce the true relationship between modal choice and income. Figures 2-1(b) and 2-1(c) illustrate two extreme but not inconceivable examples in which the assumed aggregations lead to either no apparent relationship between mode choice and income (Figure 2-1(b)) or a positive relationship (that is, increasing modal use with increasing income, as in Figure 2-1(c)), when in both cases the same true underlying relationship of a negative relationship between mode choice and income (that is,

modal use declines with increased income, as in Figure 2-1(a)) is generating the observed aggregate results. Clearly, models developed from the data shown in either Figures 2-1(b) or 2-1(c) will be seriously in error — that is, biased. And, in general, *any* spatially aggregate model will be subject to some unknown level of bias.

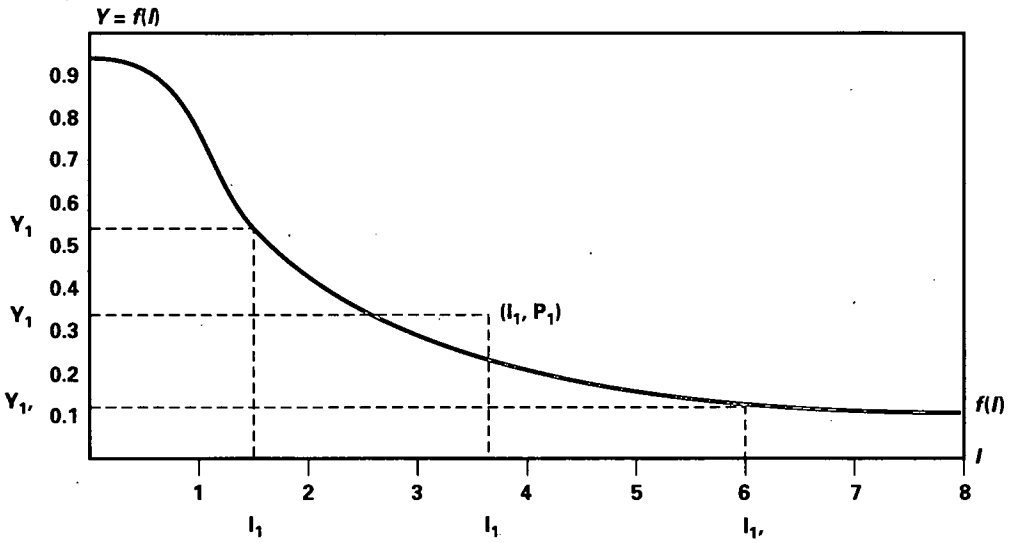
The problem of aggregation bias is lessened in situations in which zones are relatively homogeneous with respect to the variable(s) being aggregated. This is rarely the case, however, in intercity models, in which a "zone" is typically an entire urban area ("Toronto," "Montreal," etc.), within which potentially important explanatory variables, such as income, occupation, family composition, access/egress times and costs, etc., all will vary dramatically and in complex ways.

The impact of aggregation bias can also be minimized if the model is more or less restricted to policy variables involving relatively little bias (for example, city-to-city travel times and costs) and it can be argued that the "net" effect of all other factors (income, etc.) are "captured" within the model's parameters in a way which is unlikely to change significantly over the forecast period. Many aggregate intercity demand models at least approximately fall into this category in that they contain relatively few (if any) socio-economic variables (which are particularly sensitive to aggregation biases). The critical question, of course, is whether the net effect of all other factors is, in fact, constant over time (or otherwise properly controlled for) within such models.

The second major issue with respect to aggregate models, which really represents an extension of the aggregation-bias problem, is that of model transferability. It should be clear that, regardless of whether an aggregate model is "fatally" biased or not, its potential to be transferred from the area for which it has been developed to another area is extremely limited, since the net effects embedded within the aggregate model are likely to be quite different from one area to the next. Thus, a model developed for one travel corridor or one country is very unlikely to be readily transferable to another corridor or country. This has certainly been the case for urban travel demand models and, as is discussed further in subsequent sections of this report, the available evidence indicates that this also seems true for intercity models. Hence, it is likely that, at best, only very generalized results might be transferred from one region of the country to another, or from one country to another.

Figure 2-1
AGGREGATION BIAS EXAMPLE

CASE (a)



CASE (b)

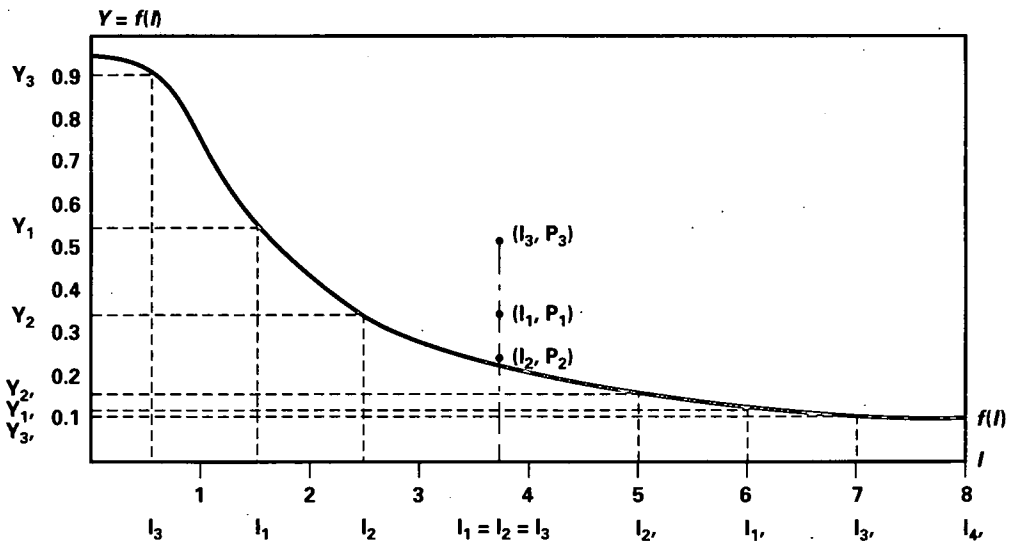
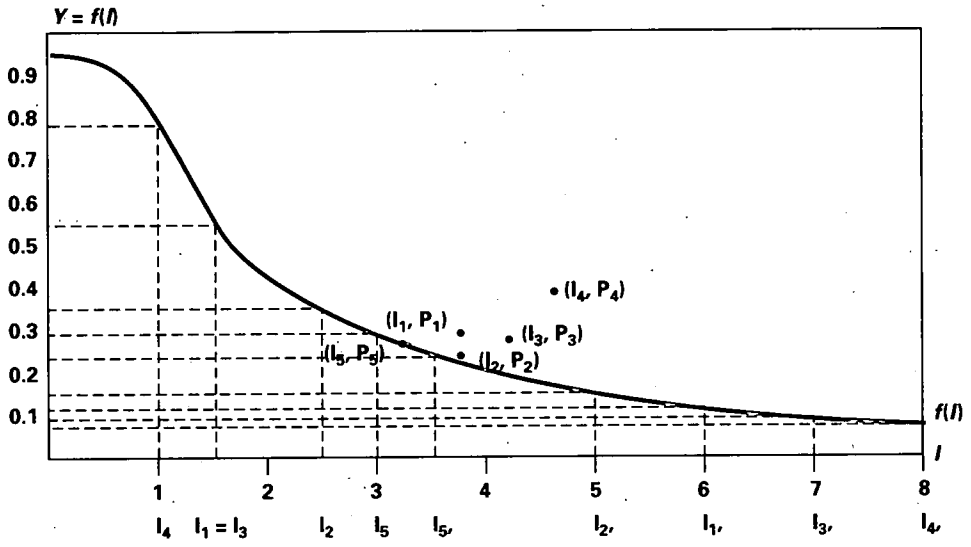


Figure 2-1 (cont'd)
AGGREGATION BIAS EXAMPLE

CASE (c)



2.3 TRAVEL MARKET DEFINITION

It is widely recognized in the demand modelling literature that intercity travel must be disaggregated by trip purpose. At a minimum this means developing separate models for business and non-business purposes in recognition of the very different behavioural processes, choice elasticities, etc., that exist within these two very different travel markets. Further disaggregation of non-business trips (into purposes such as "visit friends and relatives," "personal business," "vacation," etc.) typically depends upon data availability and the importance of distinguishing between these various sub-markets within the given corridor or region under analysis.

Further market categorization, however, is generally possible and usually desirable. In particular, it is likely that significant differences in modal availability and decision processes exist between short-distance and long-distance intercity trips, although where the break point between short- and long-distance trips lies is not necessarily well understood. Similarly, it is not clear that linear travel corridors and more general regional or inter-regional

travel systems behave in similar ways that can be captured equally well by the same model. Thus, some form of spatial categorization, on the basis of distance and/or travel system structure, may well be necessary to understand properly the intercity travel market as a whole.

This discussion of market categorization is, of course, another example of the aggregation issue discussed in the previous section. By categorizing the market "properly," one is attempting to identify travellers who are relatively homogeneous in both their decision-making process (time-cost trade-offs, etc.) and the environment within which these decisions are made (available modes, relevant modal characteristics, etc.). A particular difficulty with this market segmentation process is that it generally must be done prior to formal model estimation, with the result that statistically rigorous selection of the "optimal" categorization scheme is often difficult to achieve. Further, segmentation means that two or more models based on sub-samples within the overall set of observations will be developed, with a possible loss of statistical significance in parameter estimates. This makes market segmentation a tedious, inevitably somewhat ad hoc process which probably has not been explored in most modelling efforts as extensively or as consistently as one might wish.

An alternative to market segmentation is the use of categorizing variables directly within the model functional form. These can include spatial, purpose or socio-economic variables. Thus, for example, it is very common to include income directly within an intercity mode-choice model as an explanatory variable, rather than to estimate different models for different income groups. Inclusion of such variables within the model itself, however, usually involves very strong (and typically simplistic) assumptions concerning the effect which these variables have on the decision process being modelled. Further, they too typically involve a fairly ad hoc, trial-and-error search for the "best" combination of variables, although at least in this case some parametric statistical tests (*t*-tests, etc.) are available to aid the search.

2.4 MODEL SPECIFICATION

A considerable portion of the intercity travel demand modelling literature has focussed on the question of choice of functional form for these models. As discussed in greater detail in Rice et al. (1981), much of the early discussion of this issue was somewhat spurious in that the various model forms

considered at the time were simple algebraic variations of one another (for example, most one-stage "direct" demand models could be algebraically decomposed into a two-stage model and vice versa). As the range of modelling methods has expanded, however, significantly different model functional forms have emerged that will produce significantly different modelling results (descriptively and predictively), even when calibrated using the same data base. Further, the ability to capture these different functional forms as parametric variations of more general functional forms — and hence to test statistically the relative merits of these alternative functional forms — has grown over the years.

The selection of model functional form is fundamental to the modelling process in that it determines the data required to estimate the model, the estimation procedure used to determine model parameters (and the tractability and efficiency of this procedure) and, most importantly, the overall behaviour of the model in terms of its predictions of future system behaviour under the range of policy tests of interest. Functional forms should be selected on the basis of theoretical plausibility, goodness-of-fit to observed data, predictive feasibility and predictive performance. The last of these is, of course, of greatest practical importance but is the most difficult to assess, particularly when one is dealing with hypothetical alternatives such as the introduction of high-speed rail services into North American travel corridors. Thus, in practice, one tends to rely on theoretical reasoning and empirical descriptive results (that is, goodness-of-fit to observed data) in developing a model that one hopes will then predict well into future, unobserved situations.

3. Aggregate Intercity Travel Demand Models

3.1 INTRODUCTION

Before the 1980s, virtually all operational intercity travel demand forecasting models were totally aggregate in nature. As discussed further in Section 4, disaggregate mode choice models began to be developed in the 1970s and early 1980s. This trend has continued (for all the theoretical reasons discussed in Section 2) to the point that disaggregate mode choice models are now the operational norm.[4] Trip generation/distribution models, however, typically remain specified at the aggregate level. In general, these later models have not changed substantively in the last 20 years.[5]

Non-Canadian aggregate models from this early era of the 1960s and 1970s are well reviewed elsewhere,[6] and it would serve little purpose to repeat such a review, given the weak theoretical content of most of these models, the lack of transferability of their results, and the extent to which they have been superseded by more recent, methodologically sounder methods (at least, with respect to mode choice models). Rather, this section focusses on major Canadian efforts in the development and use of aggregate intercity travel demand models.

Two major operational aggregate models were developed during the 1970s in Canada: the Canadian Transport Commission (CTC) model, developed in 1969–1970 for the Windsor–Quebec City corridor; and Transport Canada's PERAM model, developed in the mid-1970s for Canada-wide intercity travel. Subsection 3.2 reviews the CTC model in some detail, both because it is representative of the aggregate modelling state-of-the-art circa 1970 and because it defines the point of departure for most of the Canadian modelling efforts which have followed.

Unfortunately, very little detailed information concerning PERAM is available publicly. Major studies which made use of PERAM (such as the 1984 VIA Rail Review[7] and the Southern Ontario Multimodal Passenger Study (SOMPS)[8]) typically provide qualitative overviews of the model, as well as the end forecast results — neither of which provide the sort of detailed technical information of direct interest to this review. PERAM, however, was derived from the econometric investigations of Gaudry and Wills (1978) which forms part of the material discussed in the next paragraph and, hence, is discussed in this context.

Another type of aggregate modelling work which seems to be more or less uniquely Canadian consists of investigations by several Canadian economists (Gaudry, Wills, Oum, Gillen) of the functional form of intercity travel demand models and the implications which the choice of functional form has on intercity demand elasticities. This work is of direct importance here, both from the methodological point of view of what it implies for model specification and selection, and in terms of the empirical results obtained concerning Canadian travel demand elasticities. This work is summarized in subsection 3.3.

3.2 THE CTC MODEL[9]

Although slightly over 20 years old, the model developed by the Canadian Transport Commission (CTC) from 1969 survey data for the Windsor–Quebec City corridor as part of its Intercity Passenger Transport Study represents an important starting point for reviewing Canadian intercity passenger travel demand models, for several reasons, including:

- It represents the first significant intercity demand model developed in Canada.
- It is one of the best documented models in Canada.
- It is representative of the state-of-the-art in intercity demand modelling as of the late 1960s and early 1970s.

The CTC model is a two-stage model of common carrier demand (air, rail and bus modes) in the Windsor–Quebec City corridor, defined by the following system of equations:

$$V_{AB} = K_T P_A P_B^{1.08} L_{AB}^{1.30} e^{-0.7/r_A} e^{0.23(D-T)} (C - P)^{-0.41} W^{0.205} \quad [3.1]$$

$$MS_i = w_i/W \quad [3.2]$$

$$w_i = K_i T_i^{-3.05} C_i^{-4.85} e^{-3.9/F_i} \quad [3.3]$$

$$W = \sum_i w_i \quad [3.4]$$

where:

V_{AB} = total annual trips generated from city A to city B by common carrier

K_T = constant (= 2.73)

P_A, P_B = populations of city A and city B (thousands)

L_{AB} = index of linguistic pairing between cities A and B ($0 \leq L_{AB} \leq 1$)

r_A = fraction of families with annual incomes greater than $12,000

D = highway driving time (centre to centre) (hours)

T = average total trip time by common carrier, weighted by modal split (hours)

C = average total trip cost by common carrier, weighted by modal split (dollars)

P = perceived cost of car ($0.03/vehicle-mile; 2.15 persons/vehicle)

W = level of service of the common carriers as defined in the modal split model

MS_i = fraction of traffic (modal split) using mode i (for city-pair A-B, the AB subscripts are deleted for simplicity of presentation)

w_i = level of service of mode i

W = system impedance or overall common carrier level of service

T_i = total user trip time (includes access, egress, terminal waiting and block times) (tens of hours)

C_i = total user trip cost (includes access, egress and fare costs) (tens of dollars)

F_i = perceived daily departure frequency

K_i = modal constant (air = 31.8, rail = 10.0, bus = 1.65)

The two-stage structure (total demand, mode split), the use of ad hoc "trip induction" terms (W) in the total demand equation, and the multiplicative nature of both equations [3.1] and [3.3] are all typical of the modelling state-of-the-art at the time of this model's development. Also typical is the estimation procedure (linear regression using "linearized" versions of these equations) which uses relatively few observations (in this case 34 city-pairs) to estimate a relatively large number of model parameters (eight in the total demand equation, six in the mode-split model). Somewhat unusual features of this particular model include the lack of explicit demand forecasts for the

car mode (primarily due to data limitations) and the use of the "linguistic pairing" term to deflate the level of interaction between city-pairs within the corridor on the basis of their relative linguistic compatibility.

The exponents on the population terms in equation [3.1] of 1.0 and 1.08 imply that total common carrier demand between any two city-pairs has constant unit elasticity with respect to either city's population. This, in turn, implies that if the population of the two cities were to double, then the common carrier demand for the city-pair would quadruple. The unit elasticity result is extremely unlikely and is undoubtedly the result of estimating the model on a very small sample of cross-sectional observations. A larger sample of observations taken over time would almost certainly yield improved estimates of the population effects, which would almost certainly involve exponents (and hence elasticities) significantly less than one in value (assuming that the same function form was used).

Similarly, the "trip induction" terms in this model are equally suspect. Note that a decrease in rail travel time, for example, has three effects in this model. It increases rail's common carrier modal split, since w_{rail} will increase relative to W. It also, however, increases the total common carrier demand, both through a decrease in T (average common carrier trip time) and an increase in W (common carrier level of service). This increase in common carrier demand presumably consists partially in shifts in exisiting trips from the car mode to a common carrier mode and partially in the generation of new, previously unmade trips. Problems with this approach include the following:

- Without explicitly modelling the car mode, it is very difficult to ensure that the "shift from car" effect is being captured properly.
- The cross-elasticities of common carrier volumes by mode with respect to car travel times and costs implied by this model are 0.23 and 0.41, respectively. That is, they are constant across the three common carrier modes. This is an extremely unlikely result. The constant car cost cross-elasticity is also implausible. The magnitude of the cost cross-elasticity also appears large, at least with respect to urban modelling results, in which (short-run) car usage is generally found to be quite cost-inelastic. The increase in the car time cross-elasticity with total trip time is a much more plausible result, although the appropriateness of the order of magnitude of this cross-elasticity is not easy to judge.

- The model implies that the total common carrier demand elasticity with respect to aggregate level of service (W) is 0.205. This, again, is most likely an overestimate of the trip induction effect, again due to estimation of the model from a small, cross-sectional data set. It is very unlikely that the true, net increase in trips attributable to level-of-service changes can be statistically identified from cross-sectional data, since the potential for spurious correlations, etc., is simply too great.
- There is no logical constraint in the way that common carrier level-of-service terms (that is, T, C and W) enter the total demand equation to ensure that total demand elasticities have the correct signs with respect to level-of-service variables. This is illustrated in Table 3-1, in which zero or marginally positive total demand elasticities occur for certain modes, service variables and origin-destination pairs. In each case, this result implies that increasing the travel time or cost (as the case may be) for a given mode results in no loss or even a slight increase in total common carrier ridership — a result which can only happen if one or more of the unchanged modes gains some new, "induced" riders, over and above those it gains from the mode experiencing the level-of-service change. This is clearly an illogical result, but one to which models of this general form are particularly prone.

Ignoring "trip induction" effects, the direct and cross-elasticities of a mode i's share of the market with respect to change in level-of-service variable k for mode j are given by:

$$e_{iik} = \begin{cases} \beta_k(1 - P_i) & \text{for } k = \text{time or cost} \quad [3.5.1] \\ (\beta_k/F_i)(1 - P_i) & \text{for } k = \text{frequency} \quad [3.5.2] \end{cases}$$

$$e_{ijk} = \begin{cases} -\beta_k P_j & \text{for } k = \text{time or cost; } j \neq i \quad [3.6.1] \\ -(\beta_k/F_j)P_j & \text{for } k = \text{frequency; } j \neq i \quad [3.6.2] \end{cases}$$

where:

β_k = model coefficient for the kth variable

P_i = modal split for mode i (fraction)

F_i = frequency for mode i

Points to note concerning these modal split elasticities include the following:

- Cross-elasticities are constant across the unchanged modes. For example, the cross-elasticity of the air modal split with respect to a change in rail service is the same as the bus mode's split with respect to this same change (since equations [3.6.1] and [3.6.2] do not depend on the unchanged mode i, only the changed mode j). This is illustrated in Table 3-1, in which the cross-elasticities found in any column are the same (marginal differences in total demand sensitivities aside). This is generally viewed as a rather undesirable property of mode-share models of this nature. As discussed in subsection 4.2, this is, however, a property which is common to many mode-split models.
- In general, the model implies that direct elasticities are greatest in magnitude when modal shares are smallest (that is, equation [3.5.1] states that e_{iik} has a maximum value of β when mode i has zero modal share) and decrease linearly with increasing modal share. Conversely, cross-elasticity magnitudes increase linearly with the modal share of the mode being changed (that is, mode j), reaching a maximum when the changed mode has 100 percent of the market. The general nature of these relationships is intuitively reasonable, although the strict linear nature of the relationship between mode share elasticity and mode share may be overly simplistic.
- Given these elasticity relationships in combination with the very large model coefficients in equation [3.3], very high elasticities, especially direct elasticities, can be anticipated. This is confirmed in Table 3-1, in which the direct-fare elasticities are considerably greater than 1.0 in magnitude, and the bus and rail direct-time elasticities are generally greater than 1.0 in magnitude.

Table 3-1
TIME AND FARE ELASTICITIES, CTC MODEL

Effect on volume	Montreal–Toronto					
	Schedule time			Fare		
	Air	Rail	Bus	Air	Rail	Bus
Air	–0.62	0.84	0.22	–2.75	1.61	0.40
Rail	0.29	–0.35	0.22	1.27	–2.59	0.40
Bus	0.29	0.84	–2.15	1.27	1.61	–3.87
Total	**–0.21**	**0.01**	**0.00**	**–0.90**	**0.01**	**0.07**
Effect on volume	**Ottawa–Montreal**					
	Schedule time			**Fare**		
	Air	**Rail**	**Bus**	**Air**	**Rail**	**Bus**
Air	–0.46	0.72	0.36	–2.97	1.45	0.73
Rail	0.07	–0.82	0.36	0.43	–1.66	0.73
Bus	0.07	0.72	–1.44	0.43	1.44	–2.91
Total	**0.01**	**–0.21**	**–0.16**	**0.06**	**–0.42**	**–0.32**
Effect on volume	**Toronto–Ottawa**					
	Schedule time			**Fare**		
	Air	**Rail**	**Bus**	**Air**	**Rail**	**Bus**
Air	–0.58	0.07	0.36	–2.71	1.22	0.66
Rail	0.27	–1.52	0.36	1.26	–2.64	0.66
Bus	0.27	0.71	–1.88	1.26	1.22	–3.48
Total	**–0.19**	**0.04**	**–0.01**	**–0.87**	**0.07**	**–0.03**

The linearization of the mode-split model involves defining a "base" mode, dividing equation [3.2] for each of the "non-base" modes by equation [3.2] for the "base" mode and then taking logarithms of both sides, yielding:

$$\log (MS_i/MS_b) = \log (K_i/K_b) + a^* \log (T_i/T_b) + b^* \log (C_i/C_b) + c^*(F_b/F_i) \quad [3.7]$$

where the subscript b indicates the base mode, and K_i, K_b, a, b and c are the model parameters to be estimated.

As shown by Wills (1981), application of ordinary least-squares to equation [3.7] will result in biased model coefficient estimates that will vary depending on the base mode chosen. Wills demonstrated that a multi-step generalized

regression procedure eliminates this bias. The key point, however, is that all models which were estimated using this form of linearization and ordinary least-squares regression will contain some level of bias and may generate unreliable forecast results.

3.3 INVESTIGATIONS INTO MODEL FUNCTIONAL FORM

Gaudry and Wills (1978) showed that many of the common intercity model functional forms developed to that point represented special cases of a very general model form constructed through the use of Box–Tukey or Box–Cox transformations of the dependent and independent variables within a general linear regression model. Generalized mode-split and total demand equations were eventually estimated for four modes (air, rail, bus, car) using 1972 data for 92 Canadian city-pairs. Focussing on the mode-split model, the general equation assumed is:

$$MS_m = \frac{\exp\left[\alpha_{m_0} + \sum_k \alpha_k (C_{m_k} + \mu_{1,k})^{(\lambda_{1,k})}\right]}{\sum_m \exp\left[\alpha_{m_0} + \sum_k \alpha_k (C_{m_k} + \mu_{1,k})^{(\lambda_{1,k})}\right]} \qquad [3.8]$$

where MS_m is mode m's mode share, the various α terms are model parameters, the μ and λ terms are the transformation parameters which control the specific functional form of the model, and C_{m_k} is the kth explanatory variable for mode m. Only three explanatory variables are used in the models estimated: the fare, F, the travel time, H, and the frequency, D.

Various specific models were then estimated involving different assumptions concerning the μ and λ parameters. In particular, the multiplicative model (such as used in the CTC model) is recovered if these parameters are all set equal to zero. Similarly, setting the λ terms equal to zero and the μ terms equal to one yields the standard multinomial logit model (see Section 4 for further discussion of this model).

Table 3-2 summarizes the estimation results for the five models tested, where these models have been arranged in order ranging from the most general on the left (labelled "TLCS-2") to most restricted on the right (the "log-linear" or multiplicative model and the logit model, which represent different but equally restrictive assumptions on functional form). As indicated by the log-likelihood values for the various models ($L_1(\lambda, \mu)$), the more general the model, the better the overall fit of the model. Up to a point this is a

straightforward result: the more parameters a model has (that is, the less restricted the model), the better it will generally fit a given data set.

There are, however, at least two points to note with respect to these results. The first is that the logit model performs particularly poorly relative to the other models. This may reflect the very aggregate nature of the model developed, although there is no reason in principle why any of the other models could not also be applied at a more disaggregate level as well as the logit model, perhaps with similar results.

Table 3-2
Estimation Results, Selected Models

Parameter	Model					
	A TLCS-2	B TLCS-1	C CLCS-2	D CLCS-1	E Log-linear	F Logit
(*F*) $\lambda_{1,1}$	–0.2399	–0.2660	–0.2626	–0.1930	0.0	1.0
(*H*) $\lambda_{1,2}$	–1.0982	–0.2660	–0.0513	–0.1930	0.0	1.0
(*D*) $\lambda_{1,3}$	0.0298	–0.2660	0.5712	–0.1930	0.0	1.0
μ_k	35.757	8.6862	0.0	0.0	0.0	0.0
(Air) α_{10}	4.7986 (3.629)	0.7288 (1.044)	21.762 (5.001)	0.9343 (1.462)	1.3910 (1.695)	113.00 (3.486)
(Rail) α_{20}	5.0241 (4.230)	0.5087 (1.124)	21.115 (4.957)	0.0380 (0.097)	1.0136 (1.618)	112.28 (3.459)
(Bus) α_{30}	4.2821 (3.761)	–0.3085 (–0.793)	20.354 (4.837)	–0.8106 (–2.504)	0.2516 (0.452)	111.76 (3.448)
(Car) α_{40}	0.0 (–)	0.0 (–)	0.0 (–)	0.0 (–)	0.0 (–)	0.0 (–)
(*F*) α_1	–1.8164 (–14.13)	–1.7231 (–14.96)	–1.8274 (–17.09)	–2.2254 (–18.70)	–2.9653 (–15.57)	-0.841×10^{-3} (–11.31)
(*H*) α_2	–0.0153 (–5.941)	–0.3932 (–5.105)	–0.8358 (–4.596)	–0.3605 (–4.144)	–1.3148 (–5.576)	–0.0141 (–9.092)
(*D*) α_3	1.3779 (5.735)	0.4414 (5.067)	13.503 (5.368)	0.1331 (5.022)	0.4221 (4.607)	0.0114 (3.521)
$L_1(\lambda, \mu)$	543.87	539.95	538.66	532.97	528.71	456.32
$L_1(\hat{\lambda}, \hat{\mu}) - L_1(\bar{\lambda}, \bar{\mu})$	0	3.94	5.21	10.90	15.16	87.55
r^2	0.7301	0.7223	0.7197	0.7079	0.6987	0.4909
Skewness	3.155	3.331	3.268	3.466	3.731	6.888
Kurtosis	0.566	0.711	0.624	0.765	0.990	2.667
DF	0	2	1	3	4	4

Source: Gaudry and Wills (1978).
Notes: Numbers in parentheses are *t*-statistics; r^2 = correlation coefficient; DF = degrees of freedom.

The second, more general point is that the choice of functional form can have a very dramatic impact on the conclusions drawn from the model. Table 3-3 presents the market share elasticities evaluated at average sample values for fare, time and frequency for the four modes for two of the models developed. As indicated in this table, these elasticities can vary dramatically as a function of λ, which in turn affects the nature of the model functional form. Figure 3-1 further illustrates this point by plotting the CLCS-1 fare and time elasticities as a function of λ.

"Optimal" values of λ for the CLCS-1 and TLCS-1 models are –0.193 and –0.266, respectively.

Linear interpolation of Table 3-3 yields the service elasticity estimates for the two models at optimal λ shown in Table 3-4. These estimates imply that all modes are fare-elastic (with rail and bus modes being very fare-elastic); car and air modes are time-inelastic, while rail and bus modes appear to have time elasticities between approximately –0.8 and –1.1 (depending on model assumed); and all three common carrier modes are frequency-inelastic.

Gaudry and Wills (1979) continued this general form of investigation into model functional form, in this case using a Box–Cox dogit model of the general form:[10]

$$MS_m = \frac{e^{V_m} + \theta_m \sum_{m'} e^{V_{m'}}}{(1 + \sum_{m'} \theta_{m'}) \sum_{m'} e^{V_{m'}}} \qquad [3.9]$$

$$V_m = \beta_{m_0} + \sum_k \beta_m [(X_{m_k} + \mu_k)^{\lambda_k} - 1]/\lambda_k \qquad [3.10]$$

where the β terms are parameters of the modal utility functions, the μ and λ terms are, as before, variable transformation parameters controlling the overall functional form, and the θ terms are the dogit parameters that further influence the shape of the overall function. In particular, they alter the modal cross-elasticities relative to the ordinary logit case. One interpretation of these parameters is that they capture "captivity" effects within the travelling population (that is, a certain proportion of the population may only take one mode, regardless of service levels, or may be prevented from taking a given mode due to accessibility constraints; for example, people without cars generally will not be able to use the car mode for intercity travel).

Table 3-3

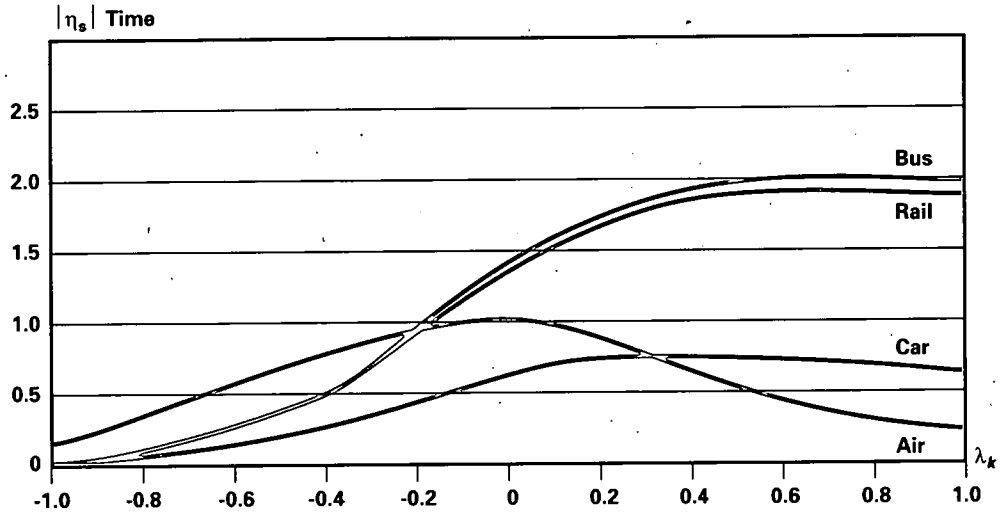
SERVICE ELASTICITIES,[a] SELECTED MODELS

	Mode-split equation; model CLCS-1; market share elasticities						Mode-split equation; model TLCS-1; market share elasticities						
μ_k=0	Car			Air			μ_k=10	Car			Air		
λ_k	Fare	Time	Frequency	Fare	Time	Frequency	λ_k	Fare	Time	Frequency	Fare	Time	Frequency
–1.0	0.48	0.02		0.50	0.16	–0.14	–1.0	0.48	0.18		0.49	0.83	–0.32
–0.5	1.03	0.16		1.31	0.61	–0.24	–0.5	0.88	0.51		1.09	1.08	–0.32
–0.2	1.25	0.37		1.86	0.86	–0.28	–0.2	1.14	0.53		1.65	0.76	–0.34
0	1.26	0.56	N/A	2.07	0.92	–0.29	0	1.22	0.56	N/A	1.99	0.63	–0.35
0.2	1.21	0.65		2.22	0.77	–0.31	0.2	1.21	0.60		2.33	0.53	–0.35
0.5	1.09	0.66		2.37	0.48	–0.32	0.5	1.09	0.63		2.39	0.38	–0.33
1.0	0.82	0.56		2.36	0.17	–0.26	1.0	0.82	0.56		2.36	0.17	–0.26
μ_k=0	Rail			Bus			μ_k=10	Rail			Bus		
λ_k	Fare	Time	Frequency	Fare	Time	Frequency	λ_k	Fare	Time	Frequency	Fare	Time	Frequency
–1.0	1.34	0.03	–0.52	1.32	0.03	–0.24	–1.0	1.31	0.30	–0.58	1.29	0.30	–0.49
–0.5	2.48	0.29	–0.53	2.50	0.29	–0.36	–0.5	2.11	0.94	–0.43	2.12	0.96	–0.46
–0.2	2.84	0.75	–0.46	2.91	0.77	–0.40	–0.2	2.55	1.09	–0.37	2.61	1.12	–0.46
0	2.74	1.21	–0.39	2.83	1.26	–0.40	0	2.65	1.24	–0.32	2.73	1.28	–0.46
0.2	2.53	1.52	–0.33	2.64	1.59	–0.40	0.2	2.53	1.45	–0.28	2.64	1.50	–0.43
0.5	2.14	1.74	–0.25	2.27	1.82	–0.38	0.5	2.15	1.69	–0.21	2.28	1.77	–0.38
1.0	1.46	1.75	–0.12	1.59	1.84	–0.38	1.0	1.46	1.75	–0.18	1.59	1.84	–0.28

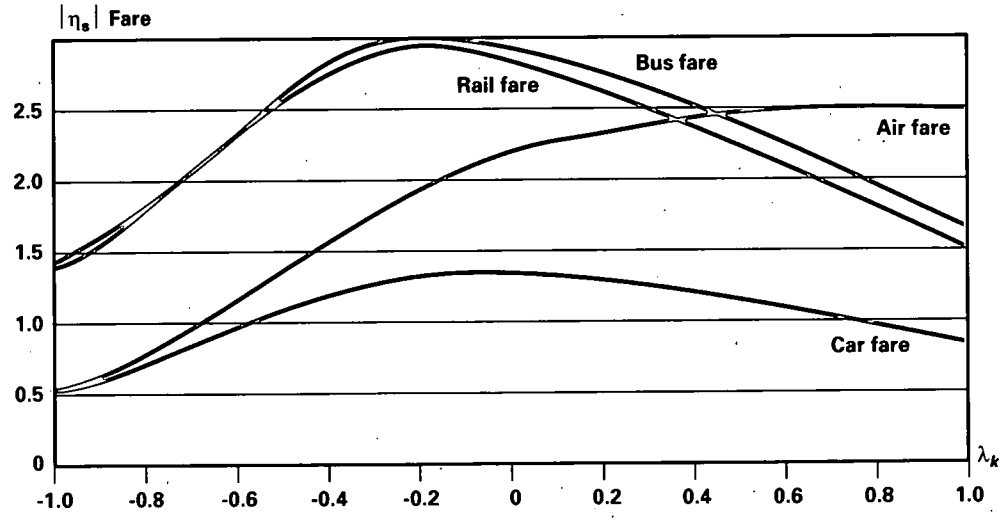
Source: Gaudry and Wills (1978).

a. By definition, elasticities are completed with the opposite of the sign of the estimated coefficient.

Figure 3-1
Variation in Fare and Time Elasticities with Functional Form, Gaudry and Wills (1978) Model CLCS-1



Mode-split equation: model CLCS-1; absolute value of time elasticities.



Mode-split equation: model CLCS-1; absolute value of fare elasticities.

Table 3-4

ESTIMATED SERVICE ELASTICITIES AT OPTIMAL λ VALUES, SELECTED MODELS

	Fare		Time		Frequency	
	CLCS-1	TLCS-1	CLCS-1	TLCS-1	CLCS-1	TLCS-1
Car	1.25	1.08	0.38	0.53	N/A	N/A
Air	1.87	1.53	0.86	0.83	–0.28	–0.34
Rail	2.84	2.45	0.77	1.06	–0.46	–0.38
Bus	2.91	2.50	0.79	1.08	0.40	–0.46

Source: Gaudry and Wills (1978).

As in the previous study, various special case models (including the multiplicative and logit models) were estimated. Table 3-5 presents the estimation results for the six models tested using 1976 observations for 56 Canadian city-pairs (all pairs were within approximately 1,000 miles of one another). In this case general conclusions include the following:

- The dogit model is only very marginally preferable to the logit model in this application.
- Unconstrained use of transformed variables provides little additional explanatory power relative to the more conventional untransformed case. Moreover, transformations do not alter the dogit-logit comparison.
- There is very little difference between the multiplicative and linear exponent (that is, logit) models in terms of model goodness-of-fit.

These results differ significantly from those reported in the same paper for a time-series urban application, in which the dogit model was found to be clearly superior to the logit model. The authors speculated that "a minimum observed market share of 5% for all alternatives may be a rough [lower] bound for the convincing use of logit model the tails of which are often too 'thin' in particular applications."[11] This is perhaps of particular importance in intercity applications where rail and/or bus often constitute "minority" modes within given travel markets and often are found to be poorly modelled by ordinary logit models (see Section 4 for further discussion of this point). Dogit models, on the other hand, are specifically designed to have choice probability distribution-tail thicknesses which vary to fit the observed behaviour and hence might better capture the behaviour of such "minority" modes.

Table 3-5
ESTIMATION RESULTS, LOGIT MODEL

Parameter	Model variants[a]					
	DU	DCM	DCL	DEU/LU	LCM	LCL
θ_1 car	10^{-5}	10^{-5}	10^{-7}	0.000	0.0	0.0
θ_2 public	0.039	0.044	0.041	0.000	0.0	0.0
λ_k fare	0.701	0.0	1.0	0.786	0.0	1.0
λ_k time	0.233	0.0	1.0	0.362	0.0	1.0
λ_k frequency	–5.438	0.0	1.0	–4.674	0.0	1.0
μ_k all variables	52.08	0.0	0.0	41.44	0.0	0.0
Fare coeff.	–2.394	–1.924	–1.530	–1.855	–1.545	–1.130
Fare *t*-stat.	(–2.34)	(–2.90)	(–1.04)	(–2.06)	(–2.90)	(–0.95)
Fare elast.[b]	–0.77/0.59	–0.89/0.89	–0.42/0.28	–0.62/0.45	–0.77/0.77	–0.34/0.23
Time coeff.	–0.2457	–1.093	–10.378	–2.960	–1.02	–9.485
Time *t*-stat.	(–8.43)	(–4.67)	(–8.40)	(–9.79)	(–5.43)	(–9.44)
Time elast.[b]	–0.72/1.04	–0.50/0.50	–1.09/3.71	–0.80/1.32	–0.51/0.51	–1.08/3.67
Freq. coeff.	17×10^{-12}	0.465	26.911	22×10^{-11}	0.371	21.603
Freq. *t*-stat.	(2.91)	(2.55)	(1.07)	(2.97)	(2.53)	(1.05)
Freq. elast.	n.c.	n.c.	n.c.	n.c.	n.c.	n.c.
ct car coeff.	0.0	0.0	0.0	0.0	0.0	0.0
ct car *t*-stat.	(–)	(–)	(–)	(–)	(–)	(–)
ct pub. coeff.	–1.270	1.621	29.648	–1.136	1.273	23.743
ct pub. *t*-stat.	(–3.26)	(127)	(1.00)	(–3.64)	(1.24)	(0.98)
$L(\Omega)$	134.52	131.78	129.95	133.31	129.69	128.56
$L(\Omega) - L(\omega)$	0.0	2.74	4.57	1.21	4.83	5.96
r^2(mean)[c]	0.74	0.70	0.69	0.77	0.74	0.70
r^2(origin)[c]	0.82	0.80	0.79	0.83	0.81	0.80
Skewness[d]	2.45	3.10	2.59	2.09	2.56	2.06
Kurtosis[d]	–0.28	0.15	–0.39	–0.18	0.24	–0.44
DF	0	4	4	1/2	6	6

Source: Gaudry and Wills (1979).

a. In model names, D = dogit, L = logit, C = constrained, U = unconstrained, M = multiplicative, L = linear exponent. Gaudry and Wills (1979) *t*-statistics are computed in each variant for the given values of θ, λ and μ.

b. The share own and cross-elasticities for alternative 2 (public) are given for market shares assumed to be equal.

c. Calculated on variables transformed by λ and μ in model format (13)-(14).

d. Skewness and kurtosis for the distribution of errors.

More recently, Gaudry (1989, 1990) has continued this comparison between logit, dogit and the "inverse power transformation logit," in which transformations are applied to the modal utility functions, rather than to the individual variables within these utility functions. This last approach represents a fairly economical means of developing a very general functional form, since it involves the introduction of only two parameters per mode (one relating to asymmetry effects and one to captivity effects with respect to each mode).

As in the 1978 Gaudry and Wills study, the logit model is found to be inferior to the more generalized forms, using 1976 data for 120 Canadian city-pairs. In particular, significant asymmetry and captivity effects are found which imply that use of an ordinary logit model would under-predict the response to significant rail service level changes.

Finally, a somewhat similar exercise was presented in Oum and Gillen (1983) in which a very generalized utility function (in this case the translog reciprocal indirect utility function) was used to derive a system of "average expenditure share" functions for five expenditure sectors: aggregate goods, aggregate services (excluding intercity travel services), and intercity air, rail and bus services. This system of expenditure share functions was estimated using Canadian time-series data constructed for the period 1961–1976, for a set of assumed parameter restrictions, corresponding to a range of hypotheses concerning travel behaviour.

In assessing these modelling results one should note that this model differs in several important ways from all other models discussed in this paper. First, the dependent variables are shares of expenditures, not trip shares. Second, the model is by far the most aggregate of all models considered in this paper in that it considers total Canadian expenditures, without any level of spatial disaggregation (for example, by origin-destination city-pair, which is the norm in other models considered here). Finally, price is the only modal characteristic which enters the model, meaning that the model cannot be used to assess the impact of level-of-service changes, impute values of time, etc. This also raises the question of the impact of supply-side changes — such as changes in travel times, frequencies, etc. — on model results, given the extended time period over which the model is applied.

In general, the estimation results indicated strong support for the most general model developed. In particular, it implied that the intercity travel sector

should not be modelled independently of other economic sectors and that "the current common use of homothetic functional forms such as log-linear models cannot be justified."[12]

Table 3-6 presents the price and income elasticities computed by Oum and Gillen using the generalized model for selected years in their time-series. In general, these results indicate that all three modes are price-elastic. They also indicate that air price elasticities declined slightly over the study period, while rail price elasticities increased at a more significant rate. The negative bus cross-elasticities are explained on the basis of the complementary nature of this mode relative to the other two (an explanation which leaves this reviewer less than fully convinced).

In summary, the various modelling exercises reviewed in this section all point to the need to use more generalized functional forms than those typically used in "operational" intercity travel demand models. In particular, the "ordinary" multinomial logit model with its large number of fairly rigid assumptions concerning functional form, cross-elasticities, etc., seems to be routinely dominated by more general model forms. This is a particularly important point given, as discussed in the next section, the prominent role the multinomial logit model has played (and continues to play) in intercity demand modelling.

The impact of this conclusion, however, is weakened by the unfortunately high level of aggregation adopted in every one of the studies cited above. The Oum and Gillen study, as already noted, is the most extreme in this regard (the omission of the very important car mode from the Oum–Gillen model is also potentially troublesome). In general, however, all the models reviewed above have not disaggregated the travel market at least into business and non-business travel, typically due to data deficiencies. As is clear from the disaggregate modelling results presented in the next section, however, this market segmentation into business and non-business travel is fundamental to achieving an adequate representation of intercity travel demand, given the clearly different variables, elasticities, etc., associated with each of these markets. Further, it may well be that more extensive market segmentation (by trip distance, party size, etc.) may also be equally important to model development.

Table 3-6

PRICE AND INCOME ELASTICITIES FOR TRAVEL DEMAND, OUM-GILLEN MODEL

Year	Quarter	Air						Bus						Rail					
		E_{11}	E_{12}	E_{13}	E_{14}	E_{15}	E_{1y}	E_{22}	E_{21}	E_{23}	E_{24}	E_{25}	E_{2y}	E_{33}	E_{31}	E_{32}	E_{34}	E_{35}	E_{3y}
1961	1	-1.2815	-0.0185	0.0311	1.0797	-2.1753	2.3645	-1.4168	-0.0757	-0.3176	-0.5064	0.2888	2.0278	-1.1051	0.1062	-0.2274	-0.0099	0.6473	0.5889
	2	-1.2537	-0.0165	0.0280	0.9691	-1.9605	2.2335	-1.3885	-0.0701	-0.2962	-0.4771	0.2736	1.9584	-1.0927	0.0946	-0.2002	-0.0173	0.5779	0.6377
	3	-1.1764	-0.0112	0.0195	0.6621	-1.3641	1.8700	-1.2747	-0.0471	-0.2101	-0.3590	0.2121	1.6787	-1.0804	0.0831	-0.1732	-0.0247	0.5090	0.6861
	4	-1.3165	-0.0209	0.0350	1.2185	-2.4451	2.5291	-1.4280	-0.0780	-0.3262	-0.5182	0.2949	2.0554	-1.1123	0.1129	-0.2433	-0.0056	0.6879	0.5603
1964	1	-1.2453	-0.0160	0.0271	0.9357	-1.8957	2.1942	-1.4424	-0.0809	-0.3371	-0.5332	0.3028	2.0908	-1.1223	0.1223	-0.2651	0.0001	0.7437	0.5211
	2	-1.2247	-0.0145	0.0248	0.8539	-1.7368	2.0973	-1.3992	-0.0722	-0.3043	-0.4884	0.2795	1.9846	-1.1044	0.1056	-0.2259	-0.0105	0.6436	0.5916
	3	-1.1615	-0.0102	0.0179	0.6031	-1.2496	1.8003	-1.2773	-0.0475	-0.2120	-0.3618	0.2136	1.6851	-1.0885	0.0906	-0.1908	-0.0200	0.5542	0.6545
	4	-1.2768	-0.0181	0.0306	1.0607	-2.1388	2.3424	-1.4332	-0.0790	-0.3301	-0.5238	0.2979	2.0682	-1.1286	0.1282	-0.2790	0.0038	0.7792	0.4962
1968	1	-1.1911	-0.0122	0.0211	0.7205	-1.4778	1.9395	-1.4115	-0.0746	-0.3136	-0.5013	0.2863	2.0148	-1.1608	0.1583	-0.3496	0.0230	0.9597	0.3694
	2	-1.1814	-0.0116	0.0201	0.6822	-1.4034	1.8942	-1.3907	-0.0704	-0.2979	-0.4798	0.2751	1.9637	-1.1319	0.1313	-0.2862	0.0056	0.7979	0.4832
	3	-1.1385	-0.0086	0.0154	0.5118	-1.0725	1.6924	-1.2897	-0.0500	-0.2214	-0.3750	0.2206	1.7156	-1.1101	0.1109	-0.2383	-0.0073	0.6757	0.5692
	4	-1.2187	-0.0141	0.0242	0.8300	-1.6908	2.0694	-1.4365	-0.0796	-0.3326	-0.5274	0.2999	2.0762	-1.1737	0.1703	-0.3779	0.0305	1.0320	0.3186
1972	1	-1.1686	-0.0107	0.0186	0.6307	-1.3041	1.8340	-1.4794	-0.0882	-0.3651	-0.5724	0.3235	2.1817	-1.2820	0.2717	-0.6157	0.0948	1.6296	-0.1084
	2	-1.1583	-0.0099	0.0175	0.5899	-1.2248	1.7856	-1.4526	-0.0828	-0.3448	-0.5446	0.3090	2.1158	-1.2079	0.2024	-0.4530	0.0505	1.2241	0.1838
	3	-1.1237	-0.0075	0.0137	0.4526	-0.9579	1.6229	-1.3140	-0.0548	-0.2398	-0.4007	0.2341	1.7753	-1.1559	0.1538	-0.3388	0.0194	0.9327	0.3887
	4	-1.1856	-0.0118	0.0205	0.6982	-1.4353	1.9139	-1.5282	-0.0981	-0.4021	-0.6232	0.3430	2.3017	-1.3429	0.3286	-0.7494	0.1312	1.9809	-0.3484
1976	1	-1.1475	-0.0092	0.0163	0.5474	-1.1421	1.7351	-1.5517	-0.1029	-0.4198	-0.6473	0.3624	2.3595	-1.4511	0.4297	-0.9870	0.1961	2.5871	-0.7749
	2	-1.1436	-0.0089	0.0159	0.5319	-1.1121	1.7168	-1.5198	-0.0964	-0.3956	-0.6143	0.3453	2.2809	-1.2770	0.2670	-0.6048	0.0919	1.6116	-0.0887
	3	-1.1163	-0.0071	0.0129	0.4234	-0.9015	1.5885	-1.3495	-0.0621	-0.2668	-0.4377	0.2534	1.8627	-1.1911	0.1867	-0.4161	0.0404	1.1301	0.2499
	4	-1.1736	-0.0110	0.0192	0.6507	-1.3431	1.8578	-1.6152	-0.1157	-0.4679	-0.7135	0.3970	2.5154	-1.5377	0.5180	-1.1772	0.2475	3.0729	-1.1164

Source: Oum and Gillen, 1983.

Note: Subscripts: 1 = air, 2 = bus, 3 = rail, 4 = goods, 5 = services, *y* = income.

Thus, the relative importance of using significantly more general functional forms in intercity demand modelling will remain difficult to determine until the sort of structured hypothesis testing represented by the Gaudry-Wills and Oum-Gillen efforts are repeated within richer data sets that permit appropriate market segmentation. This would appear to be an eminently worthwhile undertaking, particularly since it is so very rarely done with "normal" demand modelling efforts which all too typically make relatively arbitrary designs concerning model functional form at the outset of the modelling process and never revisit or test these assumptions anywhere within that process.

These important caveats concerning model aggregation/market segmentation aside, one should not lose sight of the very general, very consistent conclusion reached by these studies — that the simple multinomial logit model is likely to have difficulty in modelling "minority" (that is, small-share) modes and is almost certainly making overly strong assumptions concerning the nature of modal share elasticities, particularly cross-elasticities. As is made clear by the discussion of multinomial logit models of intercity mode choice in the next section, these concerns do seem to be verified by the experience gained with these models. As is also discussed in Section 4, other generalizations of the multinomial logit model are possible. In general, these consist of various forms of "nested" or "structured" decision structures that provide more complex elasticity structures while essentially retaining the analytical and computational simplicity of the logit model. As is discussed in subsequent sections, selection from among these various approaches (for example, generalized functional form versus generalized decisions structure versus market segmentation) may well require further systematic research similar in nature to that reviewed in this section, but applied within a broader conceptual context and within a richer empirical environment.

4. Disaggregate Intercity Travel Demand Models

4.1 INTRODUCTION

This section begins with a brief overview in subsection 4.2 of the disaggregate choice approach to modelling travel demand. Subsection 4.3 then presents a summary review of early non-Canadian disaggregate modelling efforts. All major Canadian disaggregate intercity mode-split models reported in the

literature during the past 20 years are then reviewed in some detail in subsections 4.4 and 4.5. These models can be grouped into two relatively distinct categories: multinomial logit models based on revealed preference data (that is, models are estimated using observations of actual mode choices made by actual travellers within the intercity passenger system) and "structured" binary logit models based on stated preference data (that is, models are estimated based on the choices of travellers who have been asked to state the choice they would make within a hypothetical but realistic situation). Subsection 4.4 discusses the first group (the revealed preference models), while subsection 4.5 discusses the second group (stated preference models). In addition to the Canadian models discussed in subsections 4.4 and 4.5, promising recent U.S. modelling efforts based on both revealed and stated preferences are also discussed.

4.2 OVERVIEW OF DISAGGREGATE CHOICE MODELS[13]

As discussed at some length in Section 2, modelling trip-making at the disaggregate level of the individual trip-maker has many conceptual advantages relative to the more traditional modelling of aggregate city-to-city flows. The dominant operational disaggregate mode choice model is the multinomial logit model which has the general functional form:

$$P_{it} = e^{V_{it}} / \sum_{j \in C_t} e^{V_{jt}} \qquad [4.1]$$

$$V_{it} = \beta' \boldsymbol{X}_{it} \qquad [4.2]$$

where:

P_{it} = probability that individual t chooses alternative i from the set of feasible alternatives, C_t

V_{it} = systematic utility of alternative i for individual t

$\boldsymbol{X}_{it}$ = vector of explanatory variables, consisting of attributes of alternative i (travel time, cost, etc.) and the decision-maker t (income, occupation, etc.)

β = vector of model parameters or coefficients (the prime indicates the transpose operator, thus $\beta' \boldsymbol{X}_{it}$ represents the dot product of two column vectors, β and $\boldsymbol{X}_{it}$)

The modal utility function V_{it} need not be linear in the parameters as shown in equation [4.2]. In practice, however, it generally is, since this assumption greatly simplifies the model estimation process. Note that X_{it} can include non-linear combinations of explanatory variables (for example, travel cost divided by income) as well as dichotomous "dummy" variables (for example, equal to one in value if individual t belongs to a given occupation group, equal to zero in value otherwise) without violating this linear-in-the-parameters property.

Given equations [4.1] and [4.2], the general forms of logit model mode share direct and cross-elasticities can be derived. These are:

$$e_{iik} = \beta_k X_{ik}(1 - P_i) \qquad [4.3]$$

$$e_{ijk} = -\beta_k X_{jk} P_j \qquad [4.4]$$

where X_{ik} is the value of the kth variable for mode i, and the subscript t denoting the individual has been dropped for simplicity of presentation from both the X and P variables.

These elasticities are similar to those found for the CTC model (see subsection 3.2), with the additional dependence on the level of the variable being changed (that is, X_{ik} or X_{jk}). As with the CTC model, cross-elasticities are constant across the "unchanged" modes, and both direct and cross-elasticities vary linearly with modal share (ignoring the effect of the X terms) in the same way as for the CTC model. Generally, the actual magnitudes of the elasticities will depend critically on the magnitude of the product $\beta_k X_{ik}$.

The constant cross-elasticity assumption inherent in multinomial logit models represents the single biggest weakness of the modelling method. It means that improvement in one mode (or the introduction of a new mode) will result in a diversion of trips to the changed or new mode in fixed proportions from all other modes available. This characteristic is often referred to as the "independence of irrelevant alternatives" (IIA) property. One of

the simplest ways of expressing the IIA property is to note that the ratio of choice probabilities for two modes *i* and *j* is given by:

$$P_{it}/P_{jt} = e^{V_{it}}/e^{V_{jt}} = e^{(V_{it} - V_{jt})} \qquad [4.5]$$

That is, the ratio of choice probabilities for any two modes in the choice set depends only on the systematic utilities of the two modes and not in any way on what other modes are in the choice set or on the characteristics of these other modes. In other words, these other modes are "irrelevant" to probability ratio (P_{it}/P_{jt}).

Thus, for example, if the rail mode is upgraded significantly the logit model will predict that trips diverted to the new rail service will consist of the same proportion of trips from each of the competing modes (presumably car, air and bus). The car/air and car/bus probability ratios (or any other ratio combination for these three modes) remain constant, as required by equation [4.5]. A real intercity travel market, however, is *not* likely to behave in this way. In practice, the improved rail mode will likely attract greater or fewer proportions of trips from competing modes, depending on the price/time/service combination offered. (A very high-speed, high-cost rail service presumably might divert air trips primarily; a high-speed, moderately priced service might divert car trips primarily; etc.) Hence the logit model predictions might significantly over- or underestimate modal diversion because of its IIA assumption (or equivalent constant cross-elasticity assumption).

If this IIA assumption proves to be untenable in a given application, then a more complex choice model is required. In current practice this typically means using some form of structured or nested logit model. These models are discussed in more detail in subsections 4.4.6 and 4.5, where examples of intercity passenger demand models of this form are presented.

Generally, coefficients or parameters of multinomial logit models are statistically estimated using maximum likelihood estimation (MLE) based on the observed choices made by a sample of actual trip-makers within the system being modelled. This sample can be drawn either from travellers found on each mode in the system (that is, a random sample is drawn from rail passengers, bus passengers, etc.) or from a set of households or individuals selected at random from the population at large (in which case information about the household's recent intercity travel behaviour is usually obtained).

In the first case, the choice-based sample must be weighted appropriately to obtain unbiased model parameter estimates (Manski and Lerman, 1977). Subsections 4.3 and 4.4 discuss various models which have been developed with revealed preference data and include ones developed using both choice-based and household-based survey data.

As briefly noted in the introduction, disaggregate models can also be estimated using stated preference data obtained by asking a selected group of people to make choices within hypothetical choice contexts. For example, "*if* the relative times, costs, etc., for these two modes were such and such, which mode *would* you choose?" Subsection 4.5 reviews the historical evolution of Canadian models based on stated preference data. These models are of particular importance to this report since they represent the main method used by VIA Rail during the past decade to project intermodal competition within the Canadian intercity travel market.

4.3 SUMMARY OF EARLY NON-CANADIAN MODELS[14]

Perhaps the earliest application of the multinomial logit model to intercity passenger mode choice modelling (although using aggregate data) is by Ellis et al. (1971). Watson (1972, 1974) developed a rail versus car, binary logit model for the Glasgow–Edinburgh corridor, while Leake and Underwood (1978) developed rail versus air, binary logit models for work and non-work purposes for the London–Manchester and London–Glasgow corridors. Parameter values for the two corridors were found to be quite similar, except that a positive rail bias existed in the London–Manchester corridor, whereas an air bias existed in the longer London–Glasgow corridor. Binary logit models comparing rail individually with car, bus and air were developed for the Buffalo–Albany–New York City corridor and then combined to estimate rail ridership impacts of energy-related transportation policies (Cohen et al., 1978).

One of the first applications of disaggregate multinomial logit models to intercity passenger mode choice was the Stopher and Prashker (1976) model, developed using the 1972 National Travel Survey (NTS). Although statistically significant, plausibly signed parameter estimates were obtained for business and non-business models. Counter-intuitive elasticities and very poor replication of mode shares in selected corridors were also obtained. These poor results were blamed on the data base used although, as Koppelman et al. (1984) pointed out, model specification problems also

appear to have existed. In particular, the level of service variables were all expressed as ratios relative to average values. This appears to be a holdover from some of the earlier aggregate model formulations, with little behavioural rationale, especially within a disaggregate logit model formulation.

Grayson (1981) achieved significantly improved results using the 1977 NTS data and an improved model specification (for example, inclusion of income and deletion of the relative value formulation of service variables).

Stephanedes et al. (1984) estimated a three-mode model (bus, plane and car) for the Twin Cities–Duluth corridor. This model had very high alternative specific constants and generated very high bus travel time and fare elasticities (–2.0 and –4.0, respectively). Again, these results can be attributed to methodological weaknesses in the model's development — in this case including the use of a very small, non-random sample and the mixing of reported and estimated service variables (for the chosen and unchosen modes respectively).

Finally, Morrison and Winston (1983) developed the first nested logit model of intercity passenger travel choice. It consisted of three stages: destination, mode and the decision to rent a car at the destination end. Data from the 1977 NTS were used to estimate the model. A more detailed discussion of the nested logit model is presented in subsection 4.4.6.

4.4 CANADIAN REVEALED PREFERENCE MODELS

Five multinomial logit models of Canadian intercity passenger mode choice have been developed and reported in the literature over the past 20 years. These are (in chronological order of model development):

- the TDA model, developed for the Ottawa–Montreal corridor using 1972 survey data specially collected for the project;
- the Ridout–Miller model, developed using the 1969 CTC data base for the Windsor–Quebec City corridor;
- the Wilson et al. model, developed using 1984 Canadian Travel Survey (CTS) data for Canada-wide intercity travel;

- the Abdelwahab et al. model, developed using the same 1984 CTS data base as the Wilson et al. model; and
- the PM model, developed by KPMG Peat Marwick for the Ontario/Quebec Rapid Train Task Force using 1988 VIA Rail data for the Windsor–Quebec City corridor.

Each of these models is discussed in the following subsections. In addition, Koppelman's work at Northwestern University through much of the 1980s is representative of recent U.S. efforts in this area and is reviewed in some detail subsection 4.4.6.

4.4.1 The TDA Model[15]

This model was developed specifically for the Ottawa–Montreal corridor using 1972 data specially collected for the study. The mode choice data were collected using on-board surveys for the air, rail and bus modes and a roadside interview for the car mode. The project report does not contain any discussion of the weighting procedure used to adjust logit model estimation results for the choice-based data collection approach used.

This study's most significant contribution relates to the very detailed statistical examination of the role a wide range of service variables play in explaining intercity mode choice (at least in 1972 in the Ottawa–Montreal corridor). The on-board and roadside surveys collected information on a wide range of modal service characteristics and the attitudes and perceptions of the trip-makers with respect to these characteristics. Prior to developing the multinomial logit model, an intensive analysis of these data was undertaken. This included factor analyses to determine the primary dimensions affecting modal choice; discriminant analyses to determine the variables which best discriminate, or identify, users of each mode; and contingency table, regression and linear programming analyses designed to check the discriminant analysis results. General conclusions from this study include the following:

- The self-reported rankings of how important 24 different variables were in influencing the choice of mode of travel, based on the percentage of "very important" or "fairly important" responses for each variable, are as shown in Table 4-1.

- Factor analyses of these importance rankings and the ratings of all four modes with respect to 18 modal attributes indicate that the two most important factors in modal choice are a "comfort" dimension and a "time and schedule" dimension.
- Discriminant analyses indicate that many of the variables which best discriminate between modal groups are often not ranked highly in importance by travellers. For example, by far the best three discriminating variables overall were "availability of car at destination" (ranked eighth overall in importance), "able to work en route" (ranked 18th overall in importance) and "availability of food" (ranked 11th overall in importance). Conversely, "safety" (ranked fourth in overall importance) had the fourth lowest discriminant coefficient of the 24 variables considered. (It is noted that 75 percent of the market analyzed belongs to the car mode, by far the least safe of the available modes.) Similarly, the three highest ranked variables, "schedule," "confidence of arriving on time" and "travel time," possessed discriminant coefficients which were half the magnitude of the three best discriminating variables listed above.
- Starting with the 18 modal attributes mentioned in the second point above, plus a wide range of time, cost and frequency measures, the best-fitting logit models for business and pleasure trips were found to consist of:
  - access plus egress time;
  - perceived door-to-door time (ranked from "very poor" to "very good" using a seven-point scale);
  - perceived convenience of departures (seven-point scale); and
  - door-to-door cost per person (for the "pleasure" model only).
- No attempt was made within the study to investigate the effect of traveller socio-economic characteristics (income, etc.) on intercity modal choice.

Table 4-1
RANK ORDERING OF FACTORS AFFECTING MODE CHOICE, TDA STUDY

Rank order	Variable no.	Variable definition
1	2	Schedule
2	3	Confidence of arriving on time
3	1	Travel time
4	7	Safety
5	8	Minimum of advance arrangements
6	11	Fatigue
7	5	Sitting comfort
8	23	Car at destination
9	14	Ability to relax en route
10	10	Freedom from noise
11	21	Food availability
12	4	Cost
13	12	Privacy
14	16	Ease of luggage handling
15	18	Pleasant interior
16	20	Credit card
17	24	Car left for family
18	13	Work en route
19	19	Special smokers' section
20	22	Personalized service
21	6	Terminal comfort
22	15	Amount of luggage
23	9	Seeing the country
24	17	Meeting interesting people

4.4.2 The Ridout–Miller Model[16]

This model was developed using the 1969 CTC data base for common carrier usage in the Windsor–Quebec City corridor. This choice-based data base required appropriate weighting of the observations during the parameter estimation process. As has been discussed, it did not contain any information on car trips (the single biggest component of the corridor's travel), thus limiting the ultimate policy sensitivity of the model. The main objectives of this modelling exercise were to gain experience in the application of disaggregate logit models to the intercity mode-choice problem and to investigate differences in travel behaviour across different trip purposes.

Given the latter objective, three models were developed, one for each of the following trip purposes: business, pleasure (combination of visit friends and relatives, vacation, and shopping and entertainment), and personal (combination of personal business and other). A wide variety of functional specifications were investigated for each model. Table 4-2 summarizes the results of this model estimation exercise in terms of the variables included in the final "best" specification of each model[17] and the model parameter values estimated for each variable in each model. Note that the systematic utility functions for each of the modes in the model can be recovered by adding together each of the variables defined in Table 4-2(a), multiplied by their associated parameter values in Table 4-2(b). Thus, for example, the rail-mode utility function in the business model is given by:

$$V_{rail} = 0.2032 - 0.01442*ACC - 0.004578*EG_R - 0.03507*(FARE/INC) - 0.04029*TIME_R + 0.6755*d_R \quad [4.6]$$

where all variables are as defined in Table 4-2.

Table 4-2 indicates that major differences exist in the functional forms found to best fit the data for the three models. These include the treatment of access and egress times, in-vehicle time, frequency and occupation. In other words, the way in which people evaluate the competing modes and hence how they choose a mode, appears to differ significantly from one trip purpose to another. This is over and above the differences in parameter values found for a given variable for each of the trip purposes.

Table 4-2 shows that the models best fit the data when a composite fare divided by income term is used. This implies that fare elasticities vary systematically with income level (certainly not an unreasonable proposition), given by the following modified versions of equations [4.3] and [4.4]:

$$e^p_{ii,\,fare} = (\gamma_p/Y)F_i(1 - P_i) \quad [4.7]$$

$$e^p_{ii,\,fare} = -(\gamma p/Y)F_jP_j \quad [4.8]$$

where p denotes the trip purpose, Y is the trip-maker's income (in this case represented by an index that ranges from 1 to 9 — low to high — in value), and F_i is the fare for mode i.

Given the estimated coefficient values, pleasure and personal travellers have cost elasticities which are 9.1 and 7.5 times greater, respectively, than the cost elasticity for business travellers for comparable values of modal service levels and traveller characteristics. These results are qualitatively consistent with prior expectations (that is, business travel should be much less cost sensitive than other types of intercity travel).

Again, consistent with prior expectations, pleasure and personal travellers are found to have travel time elasticities with a smaller order of magnitude than business travellers using the air mode (0.097 and 0.12 times smaller, respectively) and three to four times smaller than business travellers using the rail or bus modes (0.26 and 0.32 times smaller, respectively). This, in turn, implies that air business travellers have time elasticities that are over two and one half (2.67) times as large as bus and rail business travellers — again a reasonable result.

Table 4-2
Estimation Results, Ridout–Miller Model

(a) Independent variables		
Category	**Variable**	**Description**
Level of service variables (for mode *M*; *M* = A, R, B for air, rail bus)	ACC_M	Access distance (km)
	EG_M	Egress distance (km)
	$FARE_M$	Fare (dollars)
	$TIME_M$	In-vehicle time (h)
	$FREQ_M$	Frequency (vehicles/day)
	$DIST_M$	Air travel distance (km)
Socio-economic variables	INC	Household income
	AGE	Age of respondent
	SEX	Sex of respondent
	JOB	Occupation of respondent
	IND	Industry of respondent
	EDUC	Education level
Alternative-specific variables	D_A	1 for car mode; 0 otherwise
	D_R	1 for rail mode; 0 otherwise
	d_{A1}	1 if the traveller is in the manufacturing, construction, retail or wholesale industry, for the air mode; 0 otherwise
	d_{A2}	1 if the traveller is in the finance, insurance, real estate, or "other" industry, for the air mode; 0 otherwise
	d_{R1}	1 if the traveller is in the medical or government services, for the rail mode; 0 otherwise

Table 4-2 (cont'd)
ESTIMATION RESULTS, RIDOUT–MILLER MODEL

(b) The final models						
Variable	**Business**		**Pleasure**		**Personal**	
D_A	–0.3391	(1.79)	–1.918	(6.78)	–2.187	(8.13)
D_R	2.2032	(1.60)	0.1677	(1.41)	0.2332	(1.44)
ACC/DIST					–7.922	(4.49)
ACC	–0.01442	(2.41)				
ACC_A/DIST				(5.36)		
ACC_R/DIST			–14.42	(3.05)		
ACC_B/DIST			–5.283	(8.49)		
EG/DIST			–15.59	(1.49)		
EG_A	–0.06210	(6.85)	–2.052			
EG_R	–0.004578	(0.61)				
EG_B	–0.03612	(2.75)				
FARE/INC	–0.03507	(8.61)	–0.3201	(11.73)	–0.2616	(9.91)
TIME			–0.01044	(10.76)	–0.01275	(9.32)
$TIME_A$	–0.1075	(7.91)				
$TIME_R$	–0.04029	(15.60)				
$TIME_B$	–0.04029	(15.60)				
log (FREQ)			1.469	(6.43)	1.403	(4.62)
d_{A1}	0.8612	(6.99)				
d_{A2}	0.2715	(2.88)				
d_R	0.6755	(5.16)				
% right		83.9		48.4		45.6
ρ^2		0.703		0.171		0.176
No. of observations		2,497		2,551		1,082
No. of cases[a]		4,994		5,102		2,164

a. **The number of cases equals the number of unchosen alternatives for each observation, summed over the total number of observations.**

If α_{ip} and γ_{ip} are the time and fare parameters for mode i and purpose p, respectively, then the Ridout–Miller model implies that the value of time VOT_{ip} (1969 Can.$/hour) for travellers using mode i for purpose p is given by:

$$VOT_{ip} = (\alpha_{ip}/\gamma_{ip})*Y \qquad [4.9]$$

Given the model parameter values shown in Table 4-2, equation [4.9] implies VOTs that range from $3.07 to $27.63 for air business travellers (as income ranges from the lowest to highest category), $1.15 to $10.35 for rail and bus business travellers, and $0.03 to $0.30 and $0.05 to $0.44 for pleasure and personal purposes, respectively.

As discussed in Ridout and Miller (1989), the lack of a significant, correctly signed frequency term in the business model is certainly unexpected and disappointing, but is most likely due to the relative invariance in observed frequencies in the estimation data set, especially given the extent to which Toronto–Montreal trips dominate the business trip sample.

This lack of sufficient variability in observed modal service variables is a recurring problem in intercity demand modelling using revealed preference data, due to the relatively small number of origin-destination pairs in most corridor-oriented data sets. For example, the CTC data set originally had 34 city-pairs; Ridout and Miller were able to use only four of these city-pairs due to lack of disaggregate access/egress information for other cities in the corridor. This lack of variability is also due to the fact that most level-of-service attributes vary only on a city-pair basis and not from observation to observation for a given city-pair. The Koppelman model discussed in subsection 4.4.6 below manifests similar data-related problems in that the business model also fails to achieve a correctly signed, statistically significant frequency term. In addition, Koppelman was forced to estimate cost and travel time terms jointly by constructing a "generalized cost" term using assumed values of time to avoid collinearity problems largely caused by this lack of variability.

The final point to note from Table 4-2 is that despite extensive investigations, socio-economic variables (over and above income) play little role in explaining modal choice within the corridor. In particular, the occupation-related variables included in the business model appear to have little substantive theoretical rationale.

Two final points should be made with respect to this model. First, examination of prediction success tables constructed by comparing the model's expected predicted mode choices for the sample versus the actual observed choices indicates that the model is unable to distinguish effectively between the rail and bus modes, which typically have very similar travel times and costs for most observations. Indeed, on several links the rail service typically may cost more, provide less-frequent service and take approximately the same time and yet attract a higher modal share than the competing bus service. This phenomenon is reflected in the relatively large modal constant for rail in each of the models developed, which indicates that the observed rail mode split is underestimated by the model based on the variables included

in the model (essentially travel time, cost and frequency). This observed modal split depends in a systematic way on other factors not included in the model.

The second point is that the overall goodness-of-fit of these models is not very good. The expected percent right and ρ^2 values for the business model look very impressive, but ultimately these are a function of the fact that the air mode dominates the business market in this sample and hence the model can achieve a good fit by predicting that most trips go by air, without really discriminating between the various modal usages that actually occur. The corresponding statistics for the pleasure and personal modes are quite low by normal logit model standards (a comparable urban mode-split model — if such a comparison is meaningful — might be expected to have a ρ^2 of the order of 0.35 to 0.40 and an expected percent right of 60 or better[18]). These low goodness-of-fit statistics indicate that there is considerable uncertainty relating to modal choice which has been left unexplained by these models.

4.4.3 The Wilson et al. Model[19]

This model is estimated using 1984 Canadian Travel Survey (CTS) data. CTS is a home interview survey conducted periodically since 1977 by Statistics Canada to collect information on long-distance travel behaviour of Canadians. Wilson et al. used the 1984 survey data to develop four-mode models (air, rail, bus and car) for eastern and western Canada (Thunder Bay is the westernmost city included in the eastern region) for business and non-business trip purposes. (Models, however, are only reported for the eastern business and western non-business cases.)

Table 4-3 presents the variable definitions and coefficient estimates obtained for the eastern business model and the western non-business model, while Table 4-4 compares the results for the eastern business model with those obtained in the Ridout–Miller business model. From Table 4-4 it can be readily seen that the Wilson et al. model has a very different functional specification from the Ridout–Miller model.[20] Perhaps the most striking of these differences is that in the Wilson et al. model, income enters as an alternative-specific, stand-alone variable rather than interacting with travel cost (as in the Ridout–Miller model). These two approaches represent quite different hypotheses with respect to the effect of income on modal utilities

Table 4-3

ESTIMATION RESULTS, WILSON ET AL. MODEL

Variable No.	Variable definition	Model for business trips: eastern region			Model for non-business trips: western region		
		Parameter estimate	Asymptotic standard error	*t*-statistic value error	Parameter estimate	Asymptotic standard	*t*-statistic value
1	Constant specific to the bus utility function	16.596	3.965	4.18	1.362	1.902	0.72
2	Constant specific to the rail utility function	18.016	4.282	4.20	–0.593	2.137	–0.18
3	Constant specific to the air utility function	15.382	4.332	3.55	0.905	1.904	0.48
4	TTDT (travel time/distance)	–166.285	88.960	–1.86	–0.037	0.006	–6.47
5	TCDT (travel cost/distance)	–15.074	7.536	–2.00	–10.781	1.965	–5.48
6	TF (frequency of service)	0.018	0.004	4.27	0.0024	0.002	1.21
7	HH income specific to the bus utility function	–0.0000273	0.0000282	–0.96	–0.0000102	0.0000189	–5.36
8	HH income specific to the rail utility function	–0.0000488	0.0000350	–1.39	–0.0000333	0.0000328	–1.01
9	HH income specific to the air utility function	0.0000588	0.0000250	2.34	0.0000116	0.0000828	–0.14
		Log likelihood At equal share, $L(0) = -195.47$ At market share, $L(c) = -146.43$ At convergence, $L(\beta) = -90.29$ **Likelihood ratio** $-2[L(c) - L(\beta)] = 112.28$ **Likelihood ratio index** At zero, $\rho^2(0) = 0.538$ Adjusted, $\rho^2(a) = 0.492$ At constant, $\rho^2(c) = 0.383$ Number of cases = 141			**Log likelihood** At equal share, $L(0) = -1048.00$ At market share, $L(c) = -624.12$ At convergence, $L(\beta) = -446.32$ **Likelihood ratio** $-2[L(c) - L(\beta)] = -355.59$ **Likelihood ratio index** At zero, $\rho^2(0) = 0.574$ Adjusted, $\rho^2(a) = 0.566$ At constant, $\rho^2(c) = 0.285$ Number of cases = 756		

Note: **HH = household.**

and, hence, choices. The alternative-specific variable approach represents income as generating a "bias" between the various modes that varies with income (that is, it alters the values of the modal constants across individuals as a function of their incomes), but assumes that the traveller's sensitivity to travel cost per se remains unchanged as income varies.

Table 4-4
Comparison of Ridout–Miller and Wilson et al. Business Models

Variable	Model	
	Ridout–Miller	Wilson et al.
Constants		
Bus	—	16.592 (4.18)
Rail	0.2032 (1.64)	18.016 (4.20)
Air	–0.3391 (1.793)	15.382 (3.55)
Access distance (km)	–0.01442 (2.408)	—
Egress distance (km)		
Air	–0.0621 (6.846)	—
Rail	–0.004578 (0.612)	—
Bus	–0.03612 (2.75)	—
Travel time/distance (generic)	—	–166.285 (–1.86)
Travel time		
Air	–0.1075 (7.914)	—
Rail	–0.04029 (15.60)	—
Bus	–0.04029 (15.60)	—
Travel cost/distance	—	–15.084 (–2.00)
Fare/income	–0.3507 (8.606)	—
Frequency	—	0.018 (4.27)
Dummy for employment in manufacturing	0.8612 (6.992)	—
Dummy for employment in finances and other services	0.2715 (2.282)	—
Dummy for employment in medical and government services	0.6755 (5.163)	—
Household income		
Bus	—	–0.0000273 (–0.96)
Rail	—	–0.0000488 (–1.39)
Air	—	–0.0000488 (2.34)
ρ^2	0.7043	0.538

Notes: Numbers in parentheses are *t*-statistics values; dummy variables = 0 or 1.

This means that fare elasticities will still vary with income, but in a much less dramatic way than in the Ridout–Miller model. That is, equations [4.3] and [4.4] can be used to compute fare elasticities, with changes in income (holding all other factors constant) changing the P_i or P_j terms (that is, the modal share probabilities). The change in elasticity with respect to income

is much lower than in the Ridout–Miller model, for which equations [4.7] and [4.8] apply, and in which income enters the elasticity equation directly, as well as having an indirect effect via the probability terms. Further, the direction of the elasticity change with respect to income depends on the relative values of the modal income parameters. Given the parameter values shown in Table 4-3, rail and bus fare elasticities actually *decline* very slightly with increased income, while air fare elasticities increase slightly with income. While one cannot reject this result out of hand, it is not clear that it is consistent with reasonable a priori expectations concerning the effect of income on fare elasticities.

For similar reasons, the Wilson et al. model generates a constant value of time per model (that is, one which does not vary with income). These times, for the two reported models, are $11.02 (1984 Can.$) and $0.003 (1984 Can.$) for the eastern business model and western non-business model, respectively. The business value of time certainly cannot be rejected out of hand; however, the non-business result is surely unreasonable.

The Wilson et al. model interacts both travel time and cost with trip distance. In each case, the relative sensitivity to travel time and cost decreases as distance increases (although these sensitivities vary in the same way with distance so that the value of time does not vary with trip distance). This is not an illogical result, especially given that the *differences* between modal service characteristics determine logit model probabilities. In other words, the Wilson et al. model indicates that a given difference in the travel times or costs between two modes becomes less critical to the modal choice process as trip distance increases (that is, a five-minute travel time difference is less important for a trip of 1,000 kilometres than for a trip of 100 kilometres) — again, a not unreasonable result a priori.

As with the Ridout–Miller model, socio-economic variables other than income did not improve the model's fit of the observed data. Similarly, various "dummy" variables (that is, variables that equal either zero or one) designed to capture effects such as party size, weekend travel, trip duration and trip distance failed to make a significant contribution to the model (a not inconsistent result to the TDA model findings).

Table 4-5 presents prediction success tables for the two reported models. The results are similar to the Ridout–Miller findings in that the models do a very poor job of predicting rail and bus mode shares. In this case they appear

to predict virtually zero mode shares — a very poor result indeed! Rail and bus modes are clearly "minority" modes in both models (representing, collectively, only 11 percent and 6 percent of the trips in the two samples). It has often been found that multinomial logit models perform poorly in the prediction of minority modes.

This problem is compounded by the household-based sampling method used. It results, first, in relatively few usable observations overall (in this case 141 and 756, respectively) and, second, in very few observations of minority-mode choices. These problems can be alleviated through use of the choice-based, on-board survey approach (as used in collecting the CTC data base), since both larger samples can be efficiently gathered and minority modes can be oversampled in a statistically valid manner. As is clear from the Ridout and Miller results, sampling methodology alone is not sufficient to resolve the rail-bus prediction problem.

Table 4-5
PREDICTION SUCCESS TABLES, WILSON ET AL. MODEL

	Car[a]	Bus[a]	Rail[a]	Air[a]	Row total (observed trips)	Observed share (%)
Prediction success table for business trip model: eastern region						
Car[b]	56	—	—	13	69	48.94
Bus[b]	8	—	—	2	10	7.09
Rail[b]	4	—	—	2	6	4.25
Air[b]	6	—	—	50	56	39.72
Column total (predicted trips)	74	—	—	67	141	100.00
Predicted share (%)	52.48	—	—	47.52	100.00	
Percent correctly predicted	81.16	—	—	89.29		
Prediction success table for non-business trip model: western region						
Car[b]	473	—	—	25	498	65.87
Bus[b]	37	—	—	2	39	5.16
Rail[b]	5	—	—	2	7	0.93
Air[b]	106	—	—	106	212	28.04
Column total (predicted trips)	621	—	—	135	756	100.00
Predicted share (%)	82.14	—	—	17.86	100.00	
Percent correctly predicted	94.98	—	—	50.00		

a. Number of predicted trips by each mode.
b. Number of observed trips by each mode.

4.4.4 The Abdelwahab et al. Model[21]

This model, like the Wilson et al. model, was developed using the 1984 CTS data. Twelve models in all were developed — one model for each of recreational travel, business travel, short-distance travel (less than 960 kilometres) and long-distance travel, each estimated for eastern Canada (Thunder Bay and east), western Canada (west of Thunder Bay) and Canada as a whole (representing the combination of the first two models). A common set of variables was estimated in each of the 12 models, with the exception that the recreation and business purpose models had a dummy variable capturing the short/long-distance categorization, while the short- and long-distance models similarly had a dummy variable capturing the recreation/business purpose categorization. Table 4-6 defines the variables used in the final model specification adopted, while Table 4-7 presents the model estimation results obtained for the 12 models tested.

The primary purpose of developing these models was to test the spatial transferability of intercity, multinomial, logit mode choice models. Visual comparison of the various western and eastern models indicates that the parameter estimates for corresponding models between the two regions vary. More formal statistical tests show that, even after updating model parameters before applying them to another region,[22] eastern region models are not generally transferable to western region models, and vice versa.[23] These results are generally consistent with results found in the intra-urban case, in which transferability is rarely accomplished, except in the case of very similar cities possessing very similar transportation systems, etc.[24]

Table 4-6
Definition of Explanatory Variables, Abdelwahab et al. Model

Explanatory variable	Description
BUS-DUMMY	Dummy variable which is equal to 1 if bus is chosen and 0 otherwise
RAIL-DUMMY	Dummy variable which is equal to 1 if rail is chosen and 0 otherwise
AIR-DUMMY	Dummy variable which is equal to 1 if air is chosen and 0 otherwise
TD	Travel time (including terminal, wait and transfer times) in minutes divided by trip length in miles
CD	Travel cost (including overnight cost) in cents divided by trip length in miles
DISINC	Disposable income = household income ($000) divided by number of people contributing to household income
DD	Trip length dummy variable which is equal to 1 if the trip is short and chosen mode is bus or rail *or* the trip is long and chosen mode is car or air, and 0 otherwise
PD	Trip purpose dummy variable which is equal to 1 if the trip is recreational and chosen mode is car *or* the trip is business and chosen mode is bus, rail or air, and 0 otherwise
PASSNITE	Number of nights spent away from home divided by number of people on the trip
PCON	Number of working household members

Table 4-7
Estimation Results, Abdelwahab et al. Model

(a) Nationwide models								
Explanatory variable	**Recreational travel**		**Business travel**		**Short travel**		**Long travel**	
BUS-DUMMY	2.46	(2.17)	−0.68	(−0.35)	2.92	(2.21)	0.56	(0.00)
RAIL-DUMMY	−6.12	(−5.77)	—[a]	—	−6.41	(−6.18)	—[a]	—
AIR-DUMMY	−7.23	(−6.42)	−3.35	(−1.73)	−7.36	(−6.49)	−3.86	(−2.31)
TD	−2.94	(−7.72)	−3.08	(−4.55)	−3.57	(−7.66)	−3.50	(−5.12)
CD	−0.15	(−7.90)	−0.037	(−1.60)	−0.15	(−7.37)	0.01	(0.20)
DISINC	0.07	(3.79)	0.053	(1.40)	0.092	(5.04)	0.0009	(0.00)
DD	−2.03	(−3.17)	−2.46	(−3.39)				
PD					5.36	(3.16)	7.39	(3.00)
PASSNITE	0.31	(1.30)	0.46	(0.79)	−0.55	(−2.06)	1.62	(1.03)
PCON	0.41	(1.84)	0.20	(0.33)	0.40	(1.57)	−0.15	(−0.30)
$L(\hat{\beta})$	−326.46		−48.56		−318.68		−46.16	
χ^2	1,614.56		340.48		879.22		858.28	
ρ^2	0.7172		0.7780		0.5797		0.9029	
No. obs.	1,465		247		1,150		572	

Table 4-7 (cont'd)
ESTIMATION RESULTS, ABDELWAHAB ET AL. MODEL

Explanatory variable	Recreational travel		Business travel		Short travel		Long travel	
(b) Eastern region models								
BUS-DUMMY	2.08	(1.26)	−1.37	(−0.33)	3.20	(1.77)	−1.65	(−0.05)
RAIL-DUMMY	−6.53	(−3.89)	—[a]	—	−6.06	(−3.44)	—[a]	—
AIR-DUMMY	−9.41	(−5.06)	−6.27	(−1.91)	−8.71	(−4.36)	−6.41	(−1.94)
TD	−5.14	(−6.30)	−5.32	(−3.75)	−6.33	(−7.04)	−1.87	(−1.73)
CD	−0.11	(−3.85)	−0.02	(−0.46)	−0.09	(−3.00)	−0.14	(−1.00)
DISINC	0.11	(3.19)	0.08	(1.14)	0.10	(3.04)	0.12	(1.26)
DD	−1.97	(−2.84)	−2.91	(−1.93)				
PD					4.29	(2.62)	5.75	(2.00)
PASSNITE	−0.24	(−0.48)	0.20	(0.22)	−0.35	(−0.73)	2.29	(0.68)
PCON	0.14	(0.32)	0.86	(0.79)	−0.089	(−0.20)	0.55	(0.68)
$L(\hat{\beta})$	−101.64		−10.48		−116.02		−9.00	
χ^2	516.62		191.10		541.582		154.78	
ρ^2	0.7176		0.0012		0.7000		0.8958	
No. obs.	594		110		615		94	
(c) Western region models								
BUS-DUMMY	5.58	(2.57)	−2.20	(−0.63)	5.71	(2.54)	−0.17	(0.10)
RAIL-DUMMY	−8.89	(−5.01)	—[a]	—	−9.91	(−4.73)	—[a]	—
AIR-DUMMY	−9.44	(−5.13)	−3.11	(−0.89)	−10.20	(−4.83)	−6.10	(−2.10)
TD	−4.14	(−5.11)	−0.98	(−1.99)	−4.90	(−4.98)	−3.65	(−2.71)
CD	−0.23	(−7.19)	−0.096	(−3.11)	−0.26	(−6.19)	−0.015	(−0.17)
DISINC	0.081	(2.14)	0.11	(1.63)	0.16	(3.76)	0.056	(0.81)
DD	−2.71	(−3.09)	−1.89	(−2.65)				
PD					5.27	(3.14)	2.09	(1.91)
PASSNITE	−0.545	(−1.50)	−0.057	(−0.10)	−0.65	(−1.78)	0.85	(0.30)
PCON	0.82	(1.99)	0.71	(0.69)	0.51	(1.17)	0.54	(0.52)
$L(\hat{\beta})$	−103.22		−16.86		−80.68		−12.60	
χ^2	760.18		115.20		529.40		390.40	
ρ^2	0.7864		0.7736		0.7664		0.9394	
No. obs.	701		92		527		270	

a. **Sample size limitations — rail mode not included in this model.**

Note: **Numbers in parentheses are *t*-statistics values.**

4.4.5 The Peat Marwick (PM) Model[25]

This model was developed by KPMG Peat Marwick for the Ontario/Quebec Rapid Train Task Force using 1988 survey data collected by VIA Rail for its 1989 review. It used the same data base as the HORIZONS model discussed

in subsection 4.5 below. The model developed is a four-mode (air, rail, bus, car) multinomial logit model, disaggregated by business and non-business trip purposes.

Table 4-8 presents the model parameter estimates for the two models developed. As with the previous two models, one of the key features of the PM model is its treatment of income, which varies yet again from the two previous approaches. In this model, travellers were split into two groups on the basis of their income ("low," less than $30,000; "high," $30,000 or more), where income is expressed in 1988 Canadian dollars. Separate access/egress time and run-time parameters were then estimated for each income group for each trip purpose. This assumed that all other parameters in the utility functions (notably the cost parameter) were, on average, the same for the two income groups. Points to note concerning this approach to incorporating income effects within the model include the following:

- Value of time varies in this model with trip-maker income. Table 4-9(a) summarizes the values of time implied by this model by income level, trip purpose and time component.
- As with the Wilson et al. model, fare elasticities only vary with income (holding all other factors constant) to the extent that the modal probabilities change. This only occurs when income changes from below $30,000 per year to above. Otherwise, fare elasticities do not change with income.

Table 4-8
Estimation Results, PM Model

Variable	Parameter value	
	Business	**Non-business**
Rail constant	–1.6600	–1.6150
Air constant	–0.2206	–1.4846
Bus constant	–4.8370	–1.6700
Cost	–0.0317	–0.0416
Access time — low income	–0.0256	–0.0152
Access time — high income	–0.0393	–0.0255
Run time — low income	–0.0037	–0.0022
Run time — high income	–0.0134	–0.0088
Frequency	0.0992	0.0635
Large city — rail	1.0440	1.2230
Large city — air	0.4999	0.6338
Large city — bus	1.1360	1.1910
Group	—	–1.3330

Tables 4-9(b) and 4-9(c) provide point elasticities calculated from the PM model for the Toronto–Montreal route, based on data provided in Peat Marwick (1990). Points to note from Table 4-9 include the following:

- As with all simple logit models, the cross-elasticities shown are constant across competing modes. The numbers shown represent this constant cross-elasticity for a change in fare or run time for the mode shown. For example, the PM model implies a cross-elasticity for a change in rail fare for low-income business trips of 0.44. As is discussed at length throughout this review, the constant cross-elasticity assumption of the logit model renders these cross-elasticities somewhat suspect.
- All modes are cost-elastic in this model except the car mode for non-business purposes.
- This model indicates that low-income travellers are time-inelastic across all modes. High-income travellers are time-elastic for the rail and bus modes (regardless of trip purpose) and for the car mode for business trips.
- Air-based business trips have fare cross-elasticities considerably greater than 1.0, a result which appears somewhat counter-intuitive.
- Car-based fare cross-elasticities tend to be near 1.0 in magnitude. Non-business, high-income car-based time cross-elasticities also tend to be greater than 1.0 in magnitude. If these cross-elasticities can be trusted, they imply that common carrier usage on the Toronto–Montreal route is sensitive to both car time and cost.
- With the exception of the above-mentioned two cases, the cross-elasticities shown are quite small in magnitude.

Two factors not found in the Wilson et al. model that are designed to explain the car/common carrier competition over and above travel time, cost and frequency effects are the large city and group dummy variables.[26] While the parameters for these variables have expected signs, the statistical and, more importantly, numerical significance of these terms implies that further categorization of the market may well be required to properly specify the decision processes at work in the Windsor–Quebec City corridor.

Table 4-9

VALUES OF TIME AND ELASTICITIES, PIM MODEL

(a) Values of time ($/hour)			
Trip purpose	**Income level**	**Access/egress**	**Line-haul**
Business	High	75	25
	Low	48	7
Non-business	High	37	13
	Low	22	3

(b) Fare elasticities, Toronto–Montreal[a]								
	Direct elasticities				**Cross-elasticities**			
	Rail	**Air**	**Bus**	**Car**	**Rail**	**Air**	**Bus**	**Car**
Business								
Low income	–1.76	–3.51	–1.04	–2.08	0.44	2.61	0.02	1.14
High income	–2.06	–1.57	–1.06	–2.60	0.14	4.55	0.00	0.62
Non-business, non-group								
Low income	–1.49	–4.31	–1.11	–0.52	0.27	0.30	0.24	0.81
High income	–1.56	–3.95	–1.25	–0.44	0.19	0.66	0.10	0.90
Non-business, group								
Low income	–1.66	–4.50	–1.26	–0.19	0.10	0.11	0.09	1.14
High income	–1.69	–4.38	–1.31	–0.15	0.07	0.23	0.03	1.18

(c) Run time elasticities, Toronto–Montreal[a]								
	Direct elasticities				**Cross-elasticities**			
	Rail	**Air**	**Bus**	**Car**	**Rail**	**Air**	**Bus**	**Car**
Business								
Low income	–0.90	–0.15	–1.34	–0.80	0.22	0.11	0.03	0.44
High income	–3.82	–0.24	–4.94	–3.63	0.25	0.69	0.01	0.86
Non-business, non-group								
Low income	–0.57	–0.14	–0.67	–0.29	0.10	0.01	0.14	0.45
High income	–2.38	–0.52	–3.02	–0.96	0.30	0.09	0.24	1.99
Non-business, group								
Low income	–0.63	–0.15	–0.76	–0.11	0.04	0.00	0.05	0.63
High income	–2.57	–0.58	–3.17	–0.33	0.10	0.03	0.08	2.61

a. **Calculated using data provided in Exhibit II-11 (Peat Marwick 1990).**

In particular, note that the net bias of non-business group travellers to non-large cities is virtually –3.0 for both the rail and bus modes (–2.948 for rail, –3.003 for bus; obtained by adding the modal constant to the group variable parameter). This means that for either of these modes to be preferred to the car mode they would have to be $72 *cheaper* than the car mode or *save* 341 minutes in run time relative to the car (assuming the high-income case; multiply by 4.0 for the low-income case) or some combination of these two cases. Similar comparisons can be constructed for access time and frequency effects. Since rail and bus costs and frequencies are worse than car costs and frequencies (especially for a group), and rail and bus times are comparable, one can expect rail and bus choice probabilities for this group to be approximately $e^{-3.0}$ or 0.05 times the car choice probability value. Although more difficult to evaluate in the abstract, the air mode is likely to be similarly uncompetitive for this category of travellers, given that the very high cost of the mode (especially on a group basis) is likely to more than compensate for its smaller run times and slightly smaller modal constant. (Also note that air access times are likely to be larger than rail and bus access times for this category of traveller as well.)

Given this result it may well be the case that the non-business travel market should be further divided into group trip-makers and individual trip-makers. At a minimum, these two categories of travellers likely have quite different utility functions in terms of relevant variables and their parameter values. Even more fundamentally, they may also have very different *choice sets* from which they are making their choices. In particular, group non-business travellers between smaller cities may be effectively "captive" to the car mode (at least those with access to a car), regardless of whether or not common carrier modes are objectively available to them. If this is the case, then group travellers should be separately analyzed from other types of travellers, and their inclusion in the overall non-business market simply obscures and confounds the relationships which exist within this market.

Similar points can be made with respect to the large-city dummy variables, which significantly reduce the magnitudes of the net bias for each mode. (In the case of the air mode for business travel, the large-city variable actually changes the sign of the air bias term from negative (relative to car) to positive.) This might again point to the existence of two travel markets: a large-city market, in which the four modes compete on a more even basis, largely as a function of their relative modal service characteristics, and a small-city

model, in which people are predisposed to the use of the car for reasons that go beyond the measured time, cost and frequency values of the competing modes. Alternatively or in combination with this, it might imply some problem in the definition of modal service variables and/or modal availability (that is, choice sets) for small cities within the model's data base.

This discussion of the implications of large dummy variable and modal bias parameter values raises the more general question of the role of modal bias (or alternative-specific constant) terms in models such as the multinomial logit model. They are intended to capture the "all else being equal," *systematic* preferences shown by travellers for the various modes; that is, they capture systematic effects of modal or personal characteristics that affect mode choice but which are not otherwise explicitly captured within the model (typically these factors might include comfort and convenience effects, safety, reliability, etc.). Such bias terms *must* be included in "ranked alternatives" models such as logit mode choice models to avoid creating a bias in other parameters in the model.[27] Ideally, one hopes these terms prove to be numerically small in value, even if they are statistically significant. In practice, however, they are often numerically large, relative to other terms in the modal utility functions.

The PM model is typical in this respect, with all but the air business constant being both statistically significant and numerically large. Referring back to Tables 4-2 and 4-3, it is clear that both the Ridout–Miller model and the Wilson et al. model can be similarly criticized. For example, the rail business constant implies that the rail mode would have to be $52 cheaper than the car mode to nullify the impact of this constant on business travellers' utility calculations. The presence of these large constants raises several important concerns about the use of such models in forecasting. These include the following:

- To the extent that such terms dominate the utility functions, changes in level-of-service associated with alternatives under consideration result in small predicted changes in mode choice.
- The presence of such large constant terms generally implies that important variables affecting mode choice have been omitted from the systematic utility function.

- As noted above, such large bias terms may, in fact, be indicative of a mis-specification of the market in terms of the choice sets actually or perceived to be available to travellers. The bus business constant of –4.84 implies that the model would significantly overpredict bus usage by business travellers on the basis of cost, time and frequency alone. Over and above the omitted variables effect, it may well be that most business travellers simply do not consider the bus as a viable mode for most business trips. If this is the case, then inclusion of the bus mode in the choice sets for these travellers represents a mis-specification of the problem.
- The impact of introducing a new or dramatically upgraded mode such as high-speed rail is very difficult to forecast when large bias terms are present, since it is not at all clear what the new mode's bias term should be. In particular, a persuasive argument can be made that the rail bias terms of –1.660 and –1.615 in the PM business and non-business models should *not* be retained if the current corridor service is replaced with a significantly upgraded (or, one might well argue, entirely new) high-speed rail service. It is also unclear, based on the historically observed behaviour in the system, what the new mode's bias term value should be.[28]

This last point obviously lies at the heart of much of the debate concerning intercity travel demand model specification and application, especially when such models are frequently motivated by the need to study the impact of new modes on corridor flows and mode splits.[29] The elimination of this problem is the primary motivation of the abstract mode modelling approach characteristic of early aggregate modelling efforts. It is clear, however, that our intercity models, theories and data bases are such that abstract mode models (either aggregate or disaggregate) are not likely to be achieved in practice, leaving modellers to deal with the existence of the modal constants the best way possible. Approaches include:

- Leaving the constants "as is." This is appropriate for minor system changes or, perhaps, for short-run impacts of major system changes. It is likely, however, to be overly conservative with respect to the long-run impact of major service improvements.
- Judgementally changing the constants, perhaps based on experience with similar changes observed in other similar corridors. This approach can provide useful sensitivity testing of the model's forecasts. It also opens the technical demand-forecasting process up to charges from critics of

"tinkering" with the model to generate more desirable results. Further, similar changes in similar corridors are much more difficult to find in practice than many planners would like to admit.

- Developing an alternative model or modelling approach which permits a more sensitive treatment of the changes being considered while at the same time being "objectively defensible." Examples of such approaches are presented in subsection 4.5 below.

4.4.6 The Koppelman Model[30]

Koppelman has used 1977 Nationwide Personal Transportation Study (NPTS) data to develop four-stage nested logit models of U.S. intercity business and non-business passenger travel. The four stages are: trip frequency choice, trip destination choice, mode choice and service class choice. Each model stage is conditional upon higher-level decisions (for example, service class choice is conditional upon the mode chosen) and affects these higher-level decisions through the use of "inclusive value" terms in the upper-level utility functions to represent the expected utility associated with the lower-level decision. For example, consider the lowest two levels of the Koppelman model: mode and service class choice. In the nested logit model formulation, service class choice, conditional upon mode choice, is represented as an ordinary logit model of the form:

$$P_{c|m} = e^{V_{c|m}/\phi} / \sum_{c'} e^{V_{c'|m}/\phi} \qquad [4.10]$$

where $P_{c|m}$ is the probability of choosing service class c given mode choice m, $V_{c|m}$ is the systematic utility of service class c for mode m, and ϕ is a scale parameter which must lie between zero and one for a properly specified model. The upper level mode choice model is then given by:

$$P_m = e^{(V_m + \phi I_m)} / \sum_{m'} e^{(V_m + \phi I_m)} \qquad [4.11]$$

where V_m is the systematic utility of mode m (excluding factors relating to service class choice) and I_m is the inclusive value associated with the lower-level service class choice for mode m. This inclusive value is the expected maximum utility associated with the service class choice given that mode is selected. For logit models, this expected maximum utility can be shown to be:[31]

$$I_m = \log_e(\Sigma_c e^{V_{c|m}/\phi}) \quad [4.12]$$

This four-stage nested approach is intended to provide a theoretically consistent and sound approach to modelling intercity travel demand.[32] Specific advantages of the approach include the following:

- It permits multistage models to be built which are internally consistent (that is, with respect to scale, modelling assumptions, etc.).
- It provides an explicit, theoretically sound expression for the "trip induction" term to be included in the trip generation/distribution stage(s) of the model; that is, the inclusive value term constructed using the mode-choice model utilities for inclusion in the higher-level trip distribution model. For example, if this model is expressed as the probability of choosing destination d given a known origin zone o, then the corresponding inclusive value ("trip induction") term is given by:

  $$I_{d|o} = \log_e[\Sigma_m e^{(V_m+\phi I_m)/\delta}] \quad [4.13]$$

  where δ is the scale parameter for the mode choice level in the nested structure (that is, it will lie between zero and one in value and will be the parameter multiplied by $I_{d|o}$ in the destination choice model). Use of $I_{d|o}$ in the destination choice model to represent the impact of service changes on trip generation/distribution will result in theoretically consistent direct and cross-elasticities and should result in plausible levels of trip induction occurring (something that most ad hoc trip induction terms traditionally used do not often achieve).
- The nested approach at least partially circumvents the IIA or constant cross-elasticity assumption discussed in subsection 4.2. From the point of view of the overall joint choice process, correlation is permitted among alternatives sharing common upper-level components. For example, at the mode choice level, air mode service class combinations possess correlation because they share the air mode component of the "choice bundle." These air-related alternatives, however, are still assumed to be uncorrelated with the other alternatives at this level — the rail, bus and car modes. Similarly, at the destination choice level, the mode-destination "bundles" are correlated for a given destination because they share this common destination choice, but alternative destinations (and mode choices across these alternative destinations) remain uncorrelated. Thus, the nested logit model permits a significant relaxation of the very strict IIA assumption

of the ordinary logit model, although it still incorporates a fairly rigid covariance structure among the alternatives which may or may not be acceptable within a given application. This point is discussed further in subsection 4.5.

The disaggregate approach is motivated by the concerns raised in Section 2 of this report concerning aggregation bias. Unfortunately, a truly disaggregate data set of intercity passenger demand was not available for this model's development. The NPTS data set was used because it was the most disaggregate available, but it lacked sufficient spatial detail to allow access and egress times and costs to be computed. Thus, the primary purpose of this model is to demonstrate the feasibility of the disaggregate, nested logit modelling approach rather than develop a definitive model for policy testing.[33]

Given the emphasis within this paper on mode-split modelling, only the service class and mode choice models of the Koppelman model are discussed. Table 4-10 presents the air mode service-class model developed. This is a three-alternative model (first class, coach and discount class). Data limitations prevented the development of a comparable rail service-class model. Similarly, the data base was not large enough to support the development of separate air service-class models for business and non-business trips. Trip purpose, therefore, was incorporated into the model using dummy variables. The results indicate that business travellers are much less likely to use the discount class (presumably due to the various booking and scheduling constraints associated with this fare class), but show little preference between the coach and first-class alternatives, as indicated by the numerically small and statistically insignificant parameter for the business trip first-class dummy variable.

Cost, total daily departures (which typically vary by fare class) and income all enter the model in statistically significant ways with expected signs. In particular, higher-income people are less likely to choose discount class and more likely to take first class, relative to coach. Travel time is not included in this stage of the model since it is invariant across service classes for a given origin-destination pair.

Table 4-10
FARE/SERVICE CLASS CHOICE, KOPPELMAN MODEL

Variable	Estimate	*t*-statistics value
Alternative-specific constant		
Discount class	–0.311	0.6
First class	–0.889	1.2
Level of service		
Fare cost ($)	–0.010	2.7
Daily departures	0.555	4.1
Income ($10,000)		
Discount class	–0.263	1.3
First class	0.350	2.1
Business trip		
Discount class	–1.605	3.7
First class	–0.160	0.3
Goodness-of-fit measures		
Log likelihood		
At equal shares	–258.2	
At market shares	–205.3	
At $\hat{\beta}$	–172.3	
Likelihood ratio index (ρ^2)		
Equal share base	0.333	
Market share base	0.161	
Number of cases	235	

Table 4-11 presents the business mode choice model developed. This is a very simple model relative to the other models reviewed in this section. Points to note from this table include the following:

- The time-cost structure of the specification is the same as the Wilson et al. structure (that is, generic cost term, travel times categorized by income level), and hence the same comments concerning time and cost elasticities apply.
- Values of time cannot be deduced from this model since they were assumed prior to model estimation to be $60/hour and $20/hour for high- and low-income travellers, respectively, and then used to construct a "generalized cost" term to use in the model estimation. This approach was adopted to circumvent the high collinearity between travel time and cost that was found in the data — a common problem in intercity travel demand modelling.

Table 4-11
BUSINESS TRIP MODE CHOICE, KOPPELMAN MODEL

Variable	Estimate	*t*-statistics value
Alternative constant		
Car	–0.883	1.5
Bus	–1.703	2.2
Rail	–2.227	2.8
Level of service		
Cost ($)	–0.0046	3.0[a]
Travel time — high income (minutes)	–0.276	3.0[a]
Travel time — low income (minutes)	–0.092	3.0[a]
Distance less than 250 miles	0.324	1.5
Car	2.263	4.3
Bus and rail	1.994	2.9
Goodness-of-fit measures		
Log likelihood		
At equal shares	–359.1	
At market shares	–193.7	
At $\hat{\beta}$	–136.3	
Likelihood ratio index (ρ^2)		
Equal share base	0.623	
Market share base	0.304	
Number of cases	259	

a. **Travel time and cost variables were estimated as part of generalized cost with value of time set at $60/hour for high-income and $20/hour for low-income travellers.**

- The inclusive value parameter lies between zero and one in value. This implies that the nested model structure assumed cannot be rejected.
- As in the Ridout–Miller model, bus and rail frequencies were found to be statistically insignificant for business trip mode choice. Air frequency, as represented within the inclusive value term, does have a significant, albeit indirect, impact on business mode choice.
- All three surface modes are more attractive than the air mode for short trip distances, as indicated by their numerically large and statistically significant parameters on the dummy variables for trips of less than 250 miles in length.

Table 4-12 presents Koppelman's non-business mode choice model. Points to note from this table include the following:

- Values of time were fixed within the model prior to estimation to permit a generalized cost to be computed. In this case, $45/hour and $15/hour were assumed for high- and low-income travellers, respectively.
- As in the Ridout–Miller model, frequency is correctly signed and significant for non-business trips.

Table 4-12
Non-Business Trip Mode Choice, Koppelman Model

Variable	Estimate	*t*-statistics value
Alternative constant		
Car	1.687	4.0
Bus	0.386	0.6
Rail	0.137	0.2
Level of service		
Cost ($)	–0.00257	3.8[a]
Travel time — high income (minutes)	–0.1154	3.8[a]
Travel time — low income (minutes)	–0.0385	3.8[a]
Bus and rail frequency	0.0399	1.9
Composite air class utility	0.456	4.0
Income ($10,000)		
Car	0.0746	0.6
Bus and rail	–0.4539	2.4
Distance less than 250 miles		
Car	1.703	3.8
Bus and rail	0.8565	1.5
Distance less than 500 miles		
Car	1.796	3.5
Bus and rail	–0.816	1.3
Goodness-of-fit measures		
Log likelihood		
At equal shares	–495.6	
At market shares	–347.3	
At $\hat{\beta}$	–265.0	
Likelihood ratio index (ρ^2)		
Equal share base	0.465	
Market share base	0.235	
Number of cases	356	

a. **Travel time and cost variables were estimated as part of generalized cost with value of time set at $45/hour for high-income and $15/hour for low-income travellers.**

- In addition to the categorization of the travel time term by income, income enters this model directly. The results indicate that increasing income results in lower bus and rail utilities and has a small and statistically weak positive impact on car utilities, relative to the air mode.
- Distance effects are again captured by dummy variables. As in the business model, short (less than 250 miles) trips exhibit a surface mode bias relative to the air mode. This bias reverses in the case of the rail and bus modes for long-distance trips (greater than 500 miles), but remains strongly positive for the car mode.

4.5 STATED PREFERENCE MODELS

Considerable research has been undertaken to develop and assess stated preference based models for travel demand modelling applications. Early work in this area includes Louviere et al. (1981) and Louviere and Hensher (1982). For more recent reviews see Hensher et al. (1988) and Ben-Akiva et al. (1990). In general, the main advantages of the stated preference approach include the following:

- It provides the analyst with far greater control over the range and combination of service factors to which respondents are exposed, allowing for the investigation of a greater variability in travel times, fares, etc., than is often possible in revealed preference contexts (in which only the times, fares, etc., experienced on the relatively few number of origin-destination pairs sampled can be used). It also means that the high correlation between time and cost which often exists in observed systems can be "broken" by using uncorrelated combinations of these variables.
- It allows hypothetical or not-yet-existing modes or service levels to be tested for trip-maker responses. This is especially valuable for high-speed rail applications in which revealed preference data may "misrepresent" the modal biases expected for such services.

The major disadvantages of the approach are, first, that it requires very careful survey designs in order to ensure that valid responses are obtained. Second, questions still exist among some travel demand modellers concerning the overall validity of the technique; that is, can stated preference data be trusted to provide useful estimates of what people will actually do when faced with real, rather than hypothetical, choices? Full investigation of

this issue is well beyond the scope of this paper. The operating assumption of this review is that the answer to this question is provisionally yes, especially given the relatively promising field results discussed below.[34]

Several stated preference based models relating to intercity travel demand have been developed. Louviere and Hensher (1982), for example, investigated both intercity air–bus competition in the U.S. midwest and destination/fare class choice for leisure air travel from Australian origin cities. Morikawa et al. (1991) discussed combining revealed preference and stated preference data in a model of Japanese intercity mode choice among rail, bus and car modes. This latter study is of particular interest because it may represent a practical method for using the strengths of both revealed preference and stated preference data while minimizing the weaknesses of both approaches. More research is required before the overall utility of this approach can be assessed.

The remainder of this section focusses on two major, operational applications of the stated preference approach to intercity passenger travel demand modelling. The first is the COMPASS model, the successor to the SIGNALS and HORIZONS models developed for and used by VIA Rail in its 1984 and 1989 high-speed reviews, respectively. COMPASS has also been used in several U.S. high-speed rail studies. The second is a model developed by Charles River Associates (CRA), which has also seen application in several U.S. high-speed rail corridor studies. These two modelling approaches are reviewed in subsections 4.5.1 and 4.5.2.

4.5.1 SIGNALS, HORIZONS, COMPASS: Evolution of a Stated Preferences Approach and Overall Modelling System

SIGNALS, HORIZONS and COMPASS represent three generations of essentially the same model design. The first-generation model is SIGNALS, the property of Transmark (the consulting wing of British Rail), which was used, along with PERAM, by VIA Rail in its 1984 high-speed rail study. SIGNALS is the least well documented of the three models[35] and has been largely superseded for Canadian modelling applications by the other two models. Hence, the remainder of this section will focus on HORIZONS and COMPASS.

HORIZONS is the second-generation model, developed by Cole, Sherman and Associates Ltd. for use in VIA's 1989 review. COMPASS is the third and most recent version of this modelling system. It is the property of

Transportation and Economic Management Systems, Inc. (TEMS) and has been applied in several recent U.S. high-speed rail corridor studies. The common thread through this evolutionary process is that all three models have the same primary designer (Dr. Alex Metcalfe), and each succeeding model has built on the experience gained in the previous model, both in terms of the evolution of improved methods and in terms of incorporating data and empirical relationships from previous models into the succeeding versions.

COMPASS:[36] COMPASS is a multimode (air, rail, bus, car) modelling system which contains four major components dealing with total travel demand by all modes (as a function of socio-economic factors); induced demand (generated by changes in modal service levels); modal split; and "economic rent" (dealing with the impact of modal service changes on property values, income, employment, etc., in areas served by the intercity transportation system). It is a PC-based software system written in C, which has been designed to provide a modelling platform within which a range of models and modelling assumptions can be tested against a common data base relating to socio-economic and transport network characteristics.

The mode-split model used within this overall modelling stystem is a hierarchical decision structure in which total demand is first split between car and common carrier modes. The common carrier demand is then split between air and surface modes. The surface mode demand is then split between rail and bus. At each stage a binary logit model is estimated that has the form:

$$P_{ijp1} = 1/(1 + \exp\{-[B_{0p} + B_{1p} f(GC_{ijp1}, GC_{ijp2})]\}) \qquad [4.14]$$

where:

P_{ijp1} = probability of choosing alternative "1" from the set of alternatives {1, 2} (where, for example, alternative 1 might be auto and alternative 2 would then be common carrier) for origin-destination pair *ij* for trip purpose *p*

GC_{ijpm} = "generalized cost" of travel by mode *m* for purpose *p* for origin-destination pair *ij*

$f(\ )$ = either the difference of the generalized costs for the two alternatives (alternative 1 minus alternative 2) or the ratio of the generalized costs (alternative 1 divided by alternative 2)

B_{0p}, B_{1p} = model parameters for trip purpose p

The model parameters are estimated through regression analysis of the linearized form of equation [4.14]:

$$\log_e (P_{ijp1}/P_{ijp2}) = B_{0p} + B_{1p} f(GC_{ijp1}, GC_{ijp2}) \qquad [4.15]$$

The generalized cost terms for composite alternatives (for example, for surface common carrier modes) are constructed by weighting the generalized costs of the individual alternatives constituting the composite.

The modal generalized cost is defined as:

$$GC_{ijpm} = TT_{ijm} + TC_{ijpm}/VOT_{pm} + (VOF_{mp}*OH)/(VOT_{pm}*F_{ijm}) \qquad [4.16]$$

where:

TT_{ijm} = total travel time from i to j by mode m, with "out-of-vehicle" time components (access/egress, waiting, etc.) weighted by a factor of 2 to represent the additional disutility associated with these aspects of the trip

TC_{ijm} = total travel cost for the trip (including access/egress costs) for mode m from i to j

F_{ijm} = frequency from i to j for mode m (departures per week)

OH = operating hours per week

VOT_{pm} = value of time for mode m for purpose p

VOF_{pm} = value of frequency for mode m for purpose p

Values of time and frequency are derived through data gathered from an attitudinal survey of intercity travellers, segmented by trip purpose, mode, distance (short/long) and income (high/low), designed to elicit the respondents'

stated preferences with respect to mode choice as a function of modal attributes. Two methods are used to compute VOT and VOF from these data. The first method (Method 1) is called the "comparison method," in which the VOT (VOF) at which an individual switches from preferring the higher cost, lower travel time (higher frequency) alternative to preferring the lower cost alternative is used to define the VOT (VOF). The second method (Method 2) involves estimating binary logit models, with VOT and VOF values derived from the logit model coefficients.

Table 4-13 presents results from these two methods of calculating value of time for the tri-state corridor (Chicago–Milwaukee–Twin Cities). Table 4-14 compares the tri-state VOTs with those found in other corridor studies. It is suggested by the model developers that the higher tri-state VOTs reflect the longer trip distances in this corridor. Table 4-15 provides additional VOT/VOF information from the tri-state study.

Table 4-13
Comparison of VOT Results, Tri-State COMPASS Model

Mode/purpose	No. of valid surveys		VOT (1990$/hour)	
	Method 1	Method 2	Method 1	Method 2
Air/business	183	77	64.8	66.6
Air/other	270	97	34.0	41.9
Rail/business	63	24	39.9	45.1
Long trips[a]	14[b]	8[b]	44.5	39.8
Short trips	49	16[b]	38.6	47.7
Rail/other	207	115	28.0	32.8
Long trips	149	101	31.0	37.4
Short trips	58	14[b]	20.1	30.0
Bus/other	145	64	21.8	31.7
Long trips	72	48	28.5	34.2
Short trips	73	16[b]	15.1	24.1
Car/business	54	36	43.0	44.2
Long trips	35	23[b]	46.3	47.4
Short trips	19[b]	13[b]	37.1	38.5
Car/commuting	142	50	21.3	30.3
Long trips	6[b]	3[b]	25.7	47.4
Short trips	136	47	20.9	29.3
Car/other	377	200	25.8	37.4
Long trips	221	145	32.3	37.4
Short trips	156	55	16.9	37.1

a. **Long trips are over 100 miles, and short trips are 100 miles or less.**
b. **Less than 30 valid surveys.**

Table 4-14

COMPARISON OF VOT RESULTS, SELECTED CORRIDOR STUDIES[a]

	Tri-State (430 miles)	New York[b] (310 miles)	Ontario–Quebec[c] (180–300 miles)	Illinois[d] (200–300 miles)
	Value of time (1990$/hour)			
Air				
Business	64	51	58	54
Non-business	34	32	32	19
Rail				
Business	40	26	25	28
Non-business	28	21	19	13
Car				
Business	43	26	25	23
Non-business	26	26	18	13
Bus				
Business	25	—	17	—
Non-business	22	32	12	—
	Value of frequency (1990$/hour)			
Air				
Business	33	24	31	11
Non-business	22	3	21	7
Rail				
Business	18	11	15	6
Non-business	16	8	11	4
Car				
Business	—	17	18	7
Non-business	—	14	12	6
Bus				
Business	16	—	13	—
Non-business	13	10	9	—

a. To facilitate comparison with the tri-state study, values derived for the other three corridors were inflated to 1990$.
b. Rensselaer Polytechnic/Cole, Sherman Inc.
c. Consumer Contact Ltd/Cole, Sherman Inc.
d. British Rail.

Table 4-15

DETAILED VOT AND VOF RESULTS, TRI-STATE COMPASS MODEL

(a) Summary of VOT and VOF trade-off results				
	Air	Rail	Bus	Car
	Value of time (1990$/hour)			
Business	64.8	39.9	25.4[a]	43.0
Commuting	50.9[a]	27.0	13.7[a]	21.3
Other	34.0	28.0	21.8	25.8
	Value of frequency (1990$/hour)			
Business	33.4	17.7	16.5[a]	—
Commuting	27.7[a]	16.1	10.9[a]	—
Other	22.0	16.1	13.0	—

(b) VOT and VOF trade-off results by trip length[b]							
		Rail		Bus		Car	
	Air	Long	Short	Long	Short	Long	Short
	Value of time (1990$/hour)						
Business	64.8	44.5 (14)[d]	38.6	NI[c]	NI	46.3	37.1 (7)
Commuting	NI	NI	27.0	NI	13.7 (7)	25.7	20.9
Other	34.0	31.0	20.1	28.5	15.1	32.3	16.9
	Value of frequency (1990$/hour)						
Business	33.4	18.2 (13)	17.5	NI	NI	—	—
Commuting	NI	NI	16.1 (21)	NI	10.0 (7)	—	—
Other	22.0	17.8	11.7	15.1	10.9	—	—

Table 4-15 (cont'd)
DETAILED VOT AND VOF RESULTS, TRI-STATE COMPASS MODEL

(c) VOT and VOF trade-off results by income group[a]								
	Air		Rail		Bus		Car	
	High	Low	High	Low	High	Low	High	Low
	Value of time (1990$/hour)							
Business	73.7	55.6	45.7 (27)[d]	35.0	NI[c]	21.6	44.9 (22)	41.7
Commuting	NI	NI	NI	NI	26.8 (28)	20.2	NI	NI
Other	36.4	32.0	30.2	26.9	29.5	25.5	21.8 (21)	22.4
	Value of frequency (1990$/hour)							
Business	35.4	32.6	19.9 (27)	NI	NI	NI	—	—
Commuting	25.9	20.3	14.5	11.4 (20)	11.4 (20)	13.5	—	—

a. **Quota cells not originally identified for analysis.**
b. **"Long" indicates long-distance trips of more than 100 miles, and "short" indicates short trips of 100 miles or less.**
c. **"NI" stands for "not included" and indicates quote cells deliberately excluded from the quota survey and trade-off analysis as they were too small a sample group to be effectively analyzed.**
d. **Quota cells with numbers in parentheses had less than 30 valid surveys; the number given in parentheses is the actual number of surveys.**
e. **"High" stands for high household income of $60,000 or more per year, and "low" indicates low household income of less than $60,000 per year.**

Finally, Table 4-16 presents the estimation results for the three binary logit models developed for the tri-state corridor. Note that with the exception of the business air-surface model, the difference formulation consistently generates a higher r^2 value than the corresponding ratio formulation. This is presumably an encouraging result in that the difference formulation is consistent with the random utility theory derivation of the model,[37] whereas the ratio formulation is much more ad hoc in rationale. Also note that the bias column in each part of this table indicates the percentage of trips predicted by the model to take the indicated mode when the generalized costs of the mode and its alternative are equal. The extent to which this percentage is less than 50 percent is indicative of factors other than generalized cost

which affect the given choice but which are not explicitly captured within the model except in the constant terms (for example, convenience and privacy with respect to the car).

Table 4-16
ESTIMATION RESULTS, TRI-STATE COMPASS MODEL

Purpose	Model	B_{0p}	B_{1p}	r^2	Bias (%)
(a) Public versus car mode-split model coefficients (car bias)					
Business	Ratio	1.747 (12)	−1.941 (−17)	0.58	5
	Difference	−0.120 (−14)	−0.010 (−14)	0.79	3
Commuting	Ratio	0.922 (8)	−1.203 (−12)	0.36	7
	Difference	−0.161 (−13)	−0.007 (−18)	0.60	4
Other	Ratio	−0.378 (−14)	−0.866 (−16)	0.35	12
	Difference	−0.279 (−17)	−0.002 (−23)	0.63	7
(b) Surface versus air mode-split model coefficients (air bias)					
Business	Ratio	2.795 (20)	−3.894 (−17)	0.67	25
	Difference	−0.840 (−15)	−0.006 (−14)	0.62	20
Commuting	Ratio	4.122 (—)	−4.241 (—)	—	3
	Difference	−0.000 (—)	−0.005 (—)	—	0
Other	Ratio	4.122 (11)	−4.241 (−10)	0.66	3
	Difference	−0.000 (−16)	−0.005 (−13)	0.70	0
(c) Rail versus bus mode-split model coefficients (rail bias)					
Business	Ratio	7.858 (7)	−6.019 (−3)	0.66	36
	Difference	2.110 (5)	−0.009 (−7)	0.82	39
Commuting	Ratio	7.739 (4)	−7.066 (−4)	0.36	16
	Difference	0.631 (5)	−0.008 (−5)	0.41	15
Other	Ratio	5.583 (4)	−5.137 (−6)	0.62	11
	Difference	0.382 (7)	−0.007 (−4)	0.64	9

HORIZONS:[38] Two versions of HORIZONS were developed during the 1989 Rail Passenger Review Study. The interim model (HORIZONS I) used the COMPASS mode-split modelling method described above; that is, sequential binary models, with weighted average generalized cost terms used at each level to represent the level of service associated with the next lower level in the decision tree. Table 4-17 presents the parameter estimates for the Windsor–Quebec City corridor obtained for this version of the model, using the 1988 data base developed as part of this study.

This use of weighted average generalized costs to represent lower-level service attributes, however, can be criticized in that it is ultimately an ad hoc formulation which is not consistent with random utility theory. Further,

random utility theory provides an explicit specification of what a representative service term should consist. It is the so-called "inclusive value" or "logsum" term of the nested logit model discussed in subsection 4.4.6. The adoption of the inclusive value formulation for the representative lower level service measure necessitates the use of the difference formulation of the model, since this is the only version which is mathematically and theoretically consistent with this term's use.

Table 4-17
ESTIMATION RESULTS, INTERIM HORIZONS MODEL

(a) Public versus car mode-split model coefficients						
Purpose	**Model**	B_0	B_1		r^2	**% public**
Business	Ratio	1.0708	−1.3478	(131)	0.77	43
	Diff.	0.3130	−0.0134	(89)	0.60	57
Commuting	Ratio	0.5268	−1.1301	(85)	0.70	35
	Diff.	−2.4101	−0.0074	(18)	0.09	8
Tourist/others	Ratio	−0.1851	−0.8144	(82)	0.52	27
	Diff.	−1.8588	−0.0030	(17)	0.04	13
(b) Surface versus air mode-split model coefficients						
Purpose	**Model**	B_0	B_1		r^2	**% surface**
Business	Ratio	3.7857	−4.2083	(159)	0.80	40
	Diff.	−0.4648	−0.0082	(142)	0.76	39
Commuting	Ratio	7.1092	−5.7602	(60)	0.72	79
	Diff.	1.0696	−0.0117	(73)	0.81	74
Tourist/others	Ratio	7.4604	−6.1532	(87)	0.60	79
	Diff.	1.3436	−0.0076	(90)	0.61	79
(c) Rail versus bus mode-split model coefficients						
Purpose	**Model**	B_0	B_1		r^2	**% rail**
Business	Ratio	6.6641	−4.6676	(16)	0.51	88
	Diff.	1.6535	−0.0085	(14)	0.45	84
Commuting	Ratio	5.2573	−5.0354	(9)	0.39	56
	Diff.	−0.1252	−0.0073	(9)	0.40	47
Tourist/others	Ratio	3.8830	−3.9581	(30)	0.43	48
	Diff.	−0.0169	−1.0070	(34)	0.49	49

In addition, it was felt that structural intra- and interprovincial differences in modal usage could be captured through the use of two provincial dummy variables, I_O and I_Q, defined equal to one if the trip was an intra-provincial trip within Ontario and Quebec, respectively. Introduction of these provincial dummy variables, plus the use of the inclusive value terms described

above resulted in the final version of the model or HORIZONS II. Table 4-18 presents the estimation results for the final model version. Points to note from this table include the following:

- In comparing the final model r^2 values with those of the interim model (Table 4-17), it is seen that the goodness-of-fit has improved considerably relative to the interim difference models (which generally had rather poor goodness-of-fit values), as well as relative to the interim ratio models (which tended to out-perform the interim difference models but which had consistently lower values relative to the final model).
- A few of the ϕ values estimated are greater than 1.0. This indicates that the decision structure is possibly mis-specified. Ideally, alternative decision structures should be investigated in such cases. For example, perhaps bus should be first split off from the other two higher-quality modes (that is, air and rail). It does not appear that such alternative structures were investigated.
- It is interesting to note that the public versus private commuter model ϕ value is essentially 1.0. This implies that a joint model could replace the assumed nested model; that is, that a simpler multinomial logit model defined across the car and common carrier modes would work as well. Given that the commuter market is presumably approaching the intra-urban market in characteristics, and given that the multinomial logit model often is found to work quite well in the intra-urban case, this perhaps provides some validation of the approach adopted.

Table 4-19 presents a comparison of the interim model forecast results versus the final model forecasts (with and without the provincial dummy variables) for one test case. From this table it is seen that the replacement of the weighted average generalized costs with the logsum terms results in a significant shift in predicted usage away from the car mode to the common carrier modes, with the majority of this shift going to the rail mode. The impact of the provincial dummy variables is less dramatic but still noticeable. In this case, it deflates the predicted rail mode share by roughly 10%. The net effect of these two changes is a final model mode split for the inter-provincial Toronto–Montreal market which is not overly different from the interim model results (for example, 45.4% rail mode share versus 42.0%), whereas the final model intra-provincial results are considerably different from the interim model values (for example, 28.5% final rail mode share for Toronto–Ottawa versus the interim value of 20.5%).

Table 4-18

ESTIMATION RESULTS, FINAL HORIZONS MODEL

		r^2
I. Mode-split equations		
	Rail versus bus level	
Business	$\ln(P_{rail}/P_{bus}) = 3.092 + 0.420I_O - 1.620I_Q - 0.00541GC_{rail} + 0.00286GC_{bus}$ (2.3) (12) (7) (4)	0.73
Commuter	$\ln(P_{rail}/P_{bus}) = 1.594 - 0.00724GC_{rail} + 0.00724GC_{bus}$ (36) (24)	0.91
Other	$\ln(P_{rail}/P_{bus}) = -0.249 + 0.442I_O - 1.588I_Q - 0.00241C_{rail} + 0.00227GC_{bus}$ (6) (27) (14) (14)	0.69
	Surface versus air level	
Business	$\ln(P_{sur}/P_{air}) = -10.177 + 0.220I_O + 3.328I_Q + 1.444U_{sur} + 0.0171GC_{air}$ (9) (143) (145) (92)	0.87
Commuter	$\ln(P_{sur}/P_{air}) = -8.867 + 1.585I_O + 1.511I_Q + 0.582U_{sur} + 0.0199GC_{air}$ (11) (10) (52) (101)	0.90
Other	$\ln(P_{sur}/P_{air}) = -4.850 + 1.983I_O + 2.710I_Q + 1.677U_{sur} + 0.00807GC_{air}$ (54) (93) (54) (42)	0.72
	Public versus private level	
Business	$\ln(P_{pub}/P_{car}) = -8.105 + 0.698I_O + 2.306I_Q + 0.893U_{pub} + 0.0146GC_{car}$ (13) (42) (74) (128)	0.83
Commuter	$\ln(P_{pub}/P_{car}) = -6.782 + 1.134I_O + 0.849I_Q + 1.079U_{pub} + 0.0291GC_{car}$ (11) (9) (58) (101)	0.84
Other	$\ln(P_{pub}/P_{car}) = -3.957 + 0.143I_O + 1.958I_Q + 0.722U_{pub} + 0.0101GC_{car}$ (3) (49) (78) (99)	0.69
II. Total demand equations		
Business	$\ln(\text{trips}) = -15.775 - 0.230I_O + 2.013I_Q + 1.647U_{TOT} + 1.036 \ln(\text{emp*inc})$ (0.9) (7) (21) (11)	0.77
Commuter	$\ln(\text{trips}) = -15.756 - 0.346I_O - 0.369I_Q + 0.732U_{TOT} + 1.077 \ln(\text{emp*inc})$ (0.9) (0.8) (19) (8)	0.75
Other	$\ln(\text{trips}) = -14.759 + 0.306I_O + 1.731I_Q + 0.907U_{TOT} + 1.043 \ln(\text{emp*inc})$ (1.6) (8) (27) (12)	0.85

Note: **Values for *t*-statistics in parentheses.**

Table 4-19

COMPARISON OF FORECAST RESULTS, INTERIM AND FINAL HORIZONS MODELS

Model	Projected market shares (%)							
	Toronto–Ottawa				Toronto–Montreal			
	Rail	Air	Car	Bus	Rail	Air	Car	Bus
Base year	4.1	22.7	64.0	9.2	14.8	39.2	41.3	4.7
Interim HORIZONS model (with base year weighting)	20.5	14.5	61.4	3.7	42.0	21.6	33.2	3.2
Logsum utility approach	31.6	15.6	46.1	6.7	51.1	22.7	23.0	3.1
Logsum model (enhanced with provincial indicators)	28.5	15.5	47.2	8.8	45.4	25.0	26.2	3.5

Notes: Strategy: Rail frequency of 24 one-way trains daily. Rail in-vehicle time cut in half. Other modes unchanged. Implementation year — 1987.

Finally, Table 4-20 presents value of time, frequency and "reliability" computed for the Windsor–Quebec City corridor from the attitude survey/trade-off analysis approach described under the COMPASS model. Averaging over the two computation methods yields the values shown in Table 4-21, which are compared with similar results obtained for other North American intercity travel corridors.

Table 4-20

VALUE OF TIME, FREQUENCY AND RELIABILITY, HORIZONS MODEL (1988 CAN.$/HOUR)

	Business		Commuter		Tourist		Other purpose	
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
	Value of time							
Rail	25.2	40.4	17.8	24.0	19.6	26.9	16.0	23.4
Air	62.4	69.0	53.5	44.7	26.4	32.3	24.1	30.4
Bus	21.6	14.8	12.8[a]	11.0[a]	11.1	18.2	10.5	15.9
Car	25.8	30.2	13.2	24.1	16.4	23.4	15.6	21.4
	Value of frequency							
Rail	15.0	18.7	8.2	13.5	12.0	15.5	8.5	14.5
Air	31.7	36.4	9.8	30.7[a]	20.7	25.8	17.2	23.9
Bus	13.1	15.4	7.0[a]	7.4[a]	7.9	12.7	7.0	13.4
Car	14.5	18.0	7.8	13.8	10.2	13.6	9.0	13.5
	Value of reliability							
Rail	49.8	64.9	29.4	48.4	30.6	61.2	30.6	57.5
Air	72.0	86.0	39.0	31.9[a]	46.2	58.7	42.6	56.4
Bus	46.8	56.0	23.4[a]		31.8	53.4	26.4	50.1
Car	44.4	58.6	27.0		30.0	51.0	31.8	53.7

a. **Less than five valid surveys in each cell.**

Table 4-21

COMPARISON OF VALUES OF TIME AND FREQUENCY (1988 CAN.$/HOUR)

	Value of time				Value of frequency			
	Ont./Que.	N.Y.	Ill.	Ohio	Ont./Que.	N.Y.	Ill.	Ohio
Rail								
Business	27.8	28.9	30.6	—	16.9	11.9	6.4	—
Non-business	21.3	23.4	14.6	—	12.0	8.0	4.6	—
Air								
Business	65.7	56.6	59.8	29.6	34.1	26.5	12.6	10.9
Non-business	35.2	35.7	20.4	24.4	23.0	3.1	7.9	8.2
Bus								
Business	18.2	—	—	—	14.3	—	—	—
Non-business	13.3	35.7	—	11.9	9.2	10.6	—	5.5
Car								
Business	28.0	29.5	25.5	17.8	16.3	14.6	6.3	6.2
Non-business	19.0	29.5	15.3	14.8	11.3	13.3	4.1	4.5

Note : Values from previous studies were adjusted for inflation using published CPI figures and, where necessary, converted to Canadian dollars using U.S.$1.00 = Can.$1.23.

4.5.2 The CRA Model[39]

The starting point for the development of the Charles River Associates (CRA) model consists of the following observations:

- As has been noted several times in this report, the constant cross-elasticity (IIA) assumption of the simple multinomial logit model appears overly strong and unrealistic for the intercity mode choice case. Introduction of high-speed rail, for example, is unlikely to divert travellers in equal proportions from the competing modes.
- Brand et al. (1991) argue that nested logit models do not satisfactorily resolve this problem, since they still assume constant cross-elasticities *within* a given level of the decision structure (for example, in the HORIZONS/COMPASS formulation, between air and surface modes).
- Given that car, air and bus users are observed to possess very different values of time, frequency, etc. (compare Tables 4-11, 4-20, etc.), it can be expected that current users of *each* of these modes will divert to rail at various rates with respect to various types of rail service changes (that is, time-cost-frequency combinations). Further, the nature of travellers' values of times, elasticities, etc., are revealed through the fact that they are observed (or, in forecast mode, predicted) to have chosen a given mode. Thus, for example, we know that current car users will be quite cost sensitive but relatively time insensitive (as well as sensitive to factors such as departure flexibility, ability to carry luggage, etc.), and hence more likely to divert to moderately priced rail options than more expensive options. Conversely, air travellers are generally more time sensitive and less cost sensitive and hence will be more responsive to changes in rail travel times than fares. Presumably, therefore, an approach which directly captures these trade-offs within these different sub-markets will perform better than one which only captures the average response of the aggregated market.

Given these observations, the CRA model uses "direct" demand models to predict the origin-destination flows by mode for each of the air, car and (if available) bus modes, in the absence of high-speed rail. Bimodal logit models are then used to predict the diversion from each of these modes to high-speed rail, given the introduction of this mode (induced high-speed rail trips are generated as a separate calculation, making use of the behavioural

relationships identified in the direct demand and mode-split models). In other words, three separate logit models are used to estimate the rail–car, rail–air and rail–bus competition.

Choice-based, stated preference survey methods are used to elicit the trade-offs between car, air and bus users' current modal attributes and high-speed rail attributes required to estimate the logit models' parameters. Model estimation results are shown in Table 4-22. As indicated, each model consists of cost and time terms plus a high-speed rail constant. Thus, the probability $P_{HSR|mp}$ of a traveller choosing high-speed rail in this model, given original mode m and trip purpose p, is given by:

$$P_{HSR|mp} = \frac{\exp(\alpha_{mp} + \beta_{mp}C_{HSR} + \gamma_{mp}T_{HSR})}{\exp(\alpha_{mp} + \beta_{mp}C_{HSR} + \gamma_{mp}T_{HSR}) + \exp(\alpha_{mp} + \beta_{mp}C_m + \gamma_{mp}T_m)} \quad [4.17]$$

where:

C_k = travel cost, mode k (k = HSR, m)

T_k = composite travel time, mode k (k = HSR, m)

α_{mp} = high-speed rail constant, original mode m, trip purpose p

β_{mp}, γ_{mp} = cost and time coefficients, original mode m, trip purpose p

Table 4-22
ESTIMATION RESULTS, C/A MODEL

	Coefficients			
	Air		Car	
	Business	Non-business	Business	Non-business
Cost (1990$)	−0.0379	−0.0609	−0.0283	−0.0321
	(−4.5)	(−4.2)	(−2.2)	(−3.3)
Composite time (h)[a]	−1.3444	−1.723	−0.5636	−0.2817
	(−6.4)	(−5.3)	(−3.4)	(−2.5)
HSR constant[b]	−0.0599	0.3326	−0.771	−1.1967
	(−0.4)	(1.7)	(−1.2)	(−2.3)

a. **Composite travel time = line-haul time + 0.667(access + egress time) + 0.5(wait time).**
b. **HSR = high-speed rail.**
Note: **Numbers in parentheses are *t*-statistics.**

Tables 4-23 and 4-24 present values of time and high-speed rail direct elasticities computed from the model as recently calibrated for the "Texas triangle" (Dallas–Houston–Austin–San Antonio). Table 4-23 presents values of time disaggregated by current mode (air and car; bus is not a factor in this market), trip purpose (business, non-business) and time component (line-haul and access/egress). It is interesting to note that access/egress time values are *less* than the line-haul values, contrary to the typical urban case (as well as contrary to the HORIZONS/COMPASS assumption). Brand et al. (1991) observe that the intercity case differs from the urban case in that competition exists at the line-haul level. In addition, a significant difference in scale exists between the two time components, especially as trip lengths increase. Both of these factors, it is argued, contribute to travellers placing a higher value on line-haul than access/egress time.

Table 4-23
Implied Values of Travel Time by Mode and Trip Purpose in Texas, CRA Model (1990 U.S.$/hour)

Current mode	Trip purpose			
	Business		Non-business	
	Line-haul time	Access/egress time	Line-haul time	Access/egress time
Air				
Value of time ($/h)	35	24	28	19
Fraction of hourly wage rate	(1.3)	(0.9)	(1.5)	(1.0)
Car				
Value of time ($/h)	20	13	$9	6
Fraction of hourly wage rate	(1.0)	(0.7)	(0.5)	(0.3)

The elasticities presented in Table 4-24 are computed for the Houston–Dallas route, based on proposed downtown stations and a rail fare set at two-thirds the air fare. The air business elasticity of –0.86 rises to over 1.0 in magnitude as rail fares are set equal to air fares. Similarly, the air non-business elasticity rises to a value of –1.0 at a rail fare of about 90 percent of the non-business air fare. Conversely, car users are already marginally fare-elastic at the two-thirds air fare value. Finally, note the relative inelasticity of rail access/egress time in this model for the 240 mile (380 km) trip being analyzed.

Table 4-24
HIGH-SPEED RAIL ELASTICITIES BY MODE AND TRIP PURPOSE IN TEXAS, CRA MODEL

Mode and trip purpose	HSR elasticities[a]		
	Line-haul time	Access/egress time	Fare
Air			
Business	–0.86	–0.36	–0.81
Non-business	–0.85	–0.37	–0.74
Car			
Business	–0.61	–0.21	–1.02
Non-business	–0.38	–0.14	–1.05

a. **Calculated for characteristics between Houston and Dallas assuming that high-speed rail fares are two-thirds the air fare.**

5. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 INTRODUCTION

This section summarizes the material presented in the previous sections with respect to the following key issues:

- findings concerning demand elasticities, values of time, etc. and their implications with respect to modal substitutability;
- findings concerning selection of functional form and modelling approach; and
- suggestions concerning fruitful directions for future Canadian intercity passenger travel demand modelling.

5.2 VALUE OF TIME, DEMAND ELASTICITIES AND MODEL SUBSTITUTABILITY

5.2.1 Value of Time

A comparison of values of time (VOTs) estimated by the HORIZONS/COMPASS modelling system (Table 4-14) and those estimated by the CRA model (Table 4-23) are generally comparable, despite differences in methodology. In particular, note that the CRA air and car line-haul VOTs are both derived from binary logit models involving rail as the second mode. This implies

that the average rail mode VOT presumably lies in the range of $20 to $35 for business travel and $9 to $28 for non-business travel — values which nicely bracket the HORIZONS/COMPASS rail mode VOTs reported in Table 4-14. Similarly, the CRA line-haul VOTs for both car business and non-business and air non-business are reasonably consistent with the HORIZONS/COMPASS values, especially given the relatively short travel distances within the Texas corridor (that is, VOTs generally increase with trip length). The CRA air business VOT is, however, low relative to the HORIZONS/COMPASS values. Note that these comparisons are based on VOTs expressed in 1990 U.S. dollars.

Tables 4-9(a) and 4-21 provide a comparison of VOTs calculated using the PM multinomial logit model and the HORIZONS model calibrated to the same 1988 data base for the Windsor–Quebec City corridor (the "Ont./Que." column in Table 4-21). In this case, both tables express VOTs in 1988 Canadian dollars. Comparison between these two tables is difficult to make, given that the PM model disaggregates VOT by income (in addition to trip purpose) while the HORIZONS model disaggregates VOT by mode and purpose. In general, however, it appears that the PM model generates line-haul VOTs that are significantly lower than the HORIZONS model values. For example, the PM business line-haul VOTs range from $7 for low-income travellers to $25 for high-income travellers, whereas the range in VOTs across modes in the HORIZONS model is $18.20 (for bus) to $65.70 (for air), with both car and rail VOTs being higher than the PM upper bound of $25 (that is, $28.00 and $27.80, respectively). Similarly, the PM non-business range is from $3 to $13, which lies entirely below the HORIZONS non-business range of $13.30 to $35.20. Conversely, the PM access/egress VOTs are three to seven times higher than the line-haul VOTs, compared to the assumed ratio of two in the HORIZONS model.

Even without adjusting for inflation, it is clear that the Wilson et al. model VOTs (expressed in 1984 Canadian dollars) of $11.02 (business) and $0.03 (non-business) are significantly low relative to all three of the other models just discussed. Similarly, the Ridout–Miller non-business VOTs are significantly low relative to these models. The Ridout–Miller business VOTs, however, are not inconsistent with the more recent results discussed above, especially for higher income levels characteristic of business travellers (see subsection 4.4.2).

Thus, a certain degree of consistency in VOT estimates can be found across the models reviewed, especially if one "adjusts" for the various methodological differences (and differences in strengths and weaknesses) which exist in these models. In particular, the HORIZONS VOTs reported in Tables 4-14 and 4-21[40] possess considerable face validity in that they not only compare well with values estimated using the same modelling method in other corridors, but they also generally compare well with values generated by applying significantly different methods to the modelling of Windsor–Quebec City mode choice behaviour.

5.2.2 Demand Elasticities

In general, elasticities could not be computed from the information provided in the papers and reports reviewed. The one notable exception to this rule is the PM model, for which sufficient information was provided to compute elasticities for the Toronto–Montreal route based on 1987 operating conditions. Otherwise, this review is dependent on elasticities reported in the papers and reports reviewed. Unfortunately, only one model reviewed in Section 4 (the CRA model) has any reported elasticities (see Table 4-24). All other elasticities reported are for the aggregate models discussed in Section 3. Table 5-1 summarizes these aggregate results, plus the PM model calculations. Points to note from Tables 4-24 and 5-1 include the following:

- Both the Gaudry–Wills and the Oum–Gillen models indicate that intercity direct fare elasticities are greater than 1.0 in magnitude (that is, that demand is fare elastic). The Gaudry–Wills results indicate a much higher fare elasticity for bus and rail modes than do the Oum–Gillen results.
- The Gaudry–Wills results indicate that the car and air modes are time-inelastic, while the rail and bus modes have larger-magnitude elasticities that may marginally exceed 1.0, depending on the model assumed.
- The CTC elasticities are somewhat consistent with (although generally higher than) the Gaudry–Wills results, despite the much simpler modelling method used in the former model.
- The CRA model time elasticities are not inconsistent with the Gaudry–Wills results, especially given the aggregate nature of the latter.

Table 5-1

INTERCITY MODE SHARE ELASTICITIES, SELECTED MODELS

(a) Direct fare elasticites[a]				
Mode	**Gaudry–Wills[b]**	**Oum–Gillen[c]**	**CTC[d]**	**PM[e]**
Air	1.87–1.53	1.16	2.71–2.97	1.57–4.50
Rail	2.84–2.45	1.25	1.66–2.64	1.49–2.06
Bus	2.91–2.50	1.44	2.91–3.87	1.04–1.31
Car	1.25–1.08	—[f]	—[f]	0.15–2.60
(b) Direct time elasticities[a]				
Mode	**Gaudry–Wills[b]**	**Oum–Gillen[f]**	**CTC[d]**	**PM[e]**
Air	0.86–0.83	—	0.46–0.62	0.14–0.58
Rail	0.77–1.06	—	0.35–1.52	0.57–3.82
Bus	0.77–1.08	—	1.44–2.15	0.67–4.94
Car	0.38–0.53	—	—[f]	0.11–3.63

a. **For convenience of presentation, the negative signs on these elasticities have been deleted.**

b. **Obtained from Table 3-4. The first number shown is the CLCS-1 model elasticity; the second is the TLCS-1 value. Elasticities are based on 1972 data and are evaluated at the sample average.**

c. **Obtained from Table 3-5 by averaging the 1972 elasticities across the four quarters. Note that these are *expenditure* share elasticities rather than true mode share values.**

d. **Obtained from Table 3-1. Range indicates highest and lowest elasticities reported in this table.**

e. **Obtained from Table 4-9. Range indicates the highest and lowest elasticities obtained across the trip purpose–income level combinations considered in Table 4-9.**

f. **Elasticity not estimable from this model.**

- The CRA model, however, indicates that rail fare elasticities tend to be inelastic for current air users, contrary to the aggregate model results. These fare elasticities do, however, increase in magnitude as the rail fare rises towards the air fare level with the cross-over into the elastic range occurring at rail fares which are somewhat less than the competing air fare. Current car users in the CRA model have virtually unit rail fare elasticities, given an assumed rail fare equal to two-thirds the competing air fare.

- The PM model common carrier fare elasticities are reasonably consistent with the aggregate model results, although the variation in both the air and auto mode values seems large relative to the other findings (particularly the CRA model results).

- The variation in PM model time elasticities for the non-air modes seems to be very high relative to the other findings, although the low end of the PM model time elasticities are generally consistent with the other models' values.

Drawing generalized conclusions from such scattered results obtained from such different models is clearly hazardous at best and quite possibly foolish to undertake. Nevertheless, the following hypotheses, which appear to be consistent with the findings of this review, are advanced:

- Intercity travel demand tends to be time-inelastic. Time elasticities tend to vary from approximately –0.40 to –0.85 depending on the model used and the mode involved. The lower magnitude tends to be characteristic of car-related travel, while the upper level tends to be characteristic of air-related travel.
- Car-related intercity travel demand tends to be slightly cost-elastic, particularly in the rail fare ranges likely to be associated with high-speed rail operations.
- Air-related intercity travel demand tends to be slightly cost-inelastic, unless rail fares approach those for air, in which case demand may become unit-elastic or even slightly elastic.

These hypotheses obviously lean heavily on the CRA model results (in particular with respect to the air-related fare elasticities) and tend to be couched in the CRA model terms. This approach is adopted based on the following considerations:

- Based on the VOT comparisons discussed in subsection 5.2.1, the CRA model appears to yield similar results to currently operational Canadian models of somewhat similar design (for example, HORIZONS). Hence, in the absence of more complete information, the reported CRA elasticities are taken as being representative of this generation of models.
- As noted above, with the exception of the air-related fare elasticity case, the CRA results are reasonably compatible with the earlier Gaudry–Wills results.

- The disaggregate modelling approach is viewed as a theoretically stronger basis for modelling than the very aggregate, statistical/empirical approach represented by the Gaudry–Wills model. Hence, when in doubt, the disaggregate model results will be favoured.

5.3 FUNCTIONAL FORM AND MODELLING APPROACH

As is clear from the final point made in the previous section, judgements concerning a modelling approach are inherently dependent upon evaluations of modelling results. This is why so much of this review focusses on methodological considerations: the validity of empirical results cannot be assessed independently of the means by which these results are obtained. In terms of intercity passenger travel demand modelling methods, some fairly clear directions with respect to the evolution of these methods have emerged from this review. These can be summarized by the following observations:

- A minimum level of disaggregation is required to achieve behaviourally plausible, policy-sensitive models. This disaggregation must include the development of separate models for business and non-business purposes (with further disaggregation of the non-business category, as appropriate). The model also must be sufficiently disaggregated spatially to permit reasonable calculations of access and egress travel times and costs by mode.[41]

- With the notable exception of income, few socio-economic variables have been found to affect intercity mode choice in a consistently significant way. While this may partially reflect data deficiencies and/or lack of appropriate model testing procedures, the consistency of this result across every disaggregate model reviewed does seem to indicate some robustness in the finding.[42] This is good news for modellers, in that it reduces the amount of model disaggregation required for model specification and, correspondingly, simplifies the model aggregation/forecasting problem.

- Both aggregate and disaggregate modelling results indicate that the simple multinomial logit model is not an appropriate model formulation for intercity mode choice. The constant cross-elasticity (IIA) assumption of the multinomial logit model is untenable, based both on theoretical principles and empirical observations. Various forms of structured logit-based models are typically used to circumvent the problems inherent in the multinomial logit model.[43] These include:

- the nested logit model (typified by the Koppelman model, subsection 4.4.6);
- the sequential application of hierarchical binary logit models (typified by the HORIZONS/COMPASS family of models, subsection 4.5.1);[44] and
- the use of pairwise (rail versus a competing mode) binary logit mode choice models applied to competing mode travel volumes (the CRA model, subsection 4.5.2).

All three approaches possess various strengths and weaknesses, while the latter two, at least, are representative of the current operational state-of-practice.

- Choice-based survey methods have generally emerged as the survey method of choice for mode-choice modelling, given the greater control which such methods give over survey design as well as the greater efficiency in sample collection that can be achieved.
- The use of stated preference techniques is becoming commonplace as a means of determining plausible values of time, etc., for predicting travellers' responses to the introduction of essentially new services such as high-speed rail.

5.4 DIRECTIONS FOR MODEL DEVELOPMENT

Despite the considerable improvements in the intercity demand modelling state-of-the-art which has occurred over the past 20 years, several issues remain which require further investigation if this state-of-the-art is to continue to develop and if the contribution of these models to intercity passenger policy formulation and decision making is to be maximized. In general terms, these issues relate to the need for more systematic, general investigations into alternative model specifications and into the practical as well as statistical performance of these models. More specifically, these issues include the following:

- A need exists to explore intercity travel market segmentation in a more detailed, systematic way than has generally been undertaken. Ridout and Miller (1989) and Abdelwahab et al. (1991) represent examples of very partial attempts to explore this issue, but much more comprehensive

investigations involving more detailed data bases are required. The role of trip distance and income as categorizing rather than explanatory variables, car ownership (an almost totally unexplored variable in the intercity context) and seasonal variations in travel choices (again, almost totally unexplored but surely of significant interest in the Canadian context) all require considerable additional investigation.

- A need exists to explore in a consistent way (that is, using the same data base, etc.) the various options for structuring the intercity mode choice process discussed in the previous section. Additional options also exist, including alternative orderings of choices within the binary choice hierarchy.
- A need exists to apply the use of very generalized functional forms (typified by the work of Gaudry) within the context of the structured (partially) disaggregate models characteristic of current operational methods. Computational complexities undoubtedly exist with respect to this approach.[45] Nevertheless, use of such generalized functional forms typically widens the range of "testable" model assumptions and provides useful insights into the extent to which the more restricted functional forms (and, hence, typically the underlying theory generating these restricted functional forms) are adequately capturing observed behaviour.

In general, these identified needs point to the more basic need for treating intercity travel demand modelling as a research task; that is, as a (typically interactive) process of hypothesis formulation and testing designed to improve our understanding of intercity travel behaviour in general and our practical capabilities for predicting future travel behaviour in particular. This approach can be contrasted with the all too common approach adopted in this field in which models are treated as proprietary tools that are designed to promote a particular point of view and that are not open to peer scrutiny and professional, informed debate. Without such scrutiny and debate, however, the modelling state-of-the-art will inevitably fail to achieve its potential, will suffer from a general lack of credibility and, hence, inevitably fail the policy formulation process it is intended to serve.

ENDNOTES

1. No attempt to review single mode demand models (such as air demand forecasting models) has been made in this study, since these provide little or no information concerning intermodal substitutability. For a recent review of air demand forecasting models, see Hutchinson (1991).

2. For more detailed criticisms of aggregate models, see Rice et al. (1981) and Koppelman et al. (1984).

3. Technically, one rarely observes the probability (frequency) of an individual's modal choice, but rather the choice of a single mode in a single-choice situation. This complicates the model estimation process somewhat but does not alter the basic argument being made here.

4. As is discussed further in subsection 4.4.6, most so-called disaggregate models actually still retain some level of spatial aggregation, primarily due to data limitations. The overall methodological approach, however, is essentially disaggregate in nature.

5. There are exceptions to this generalization. See, for example, subsection 4.4.6, which discusses the Koppelman model. This model involves extension of disaggregate choice theory to the entire intercity travel demand modelling process.

6. See Hartgen and Cohen (1976), Rice et al. (1981) and Koppelman et al. (1984).

7. VIA Rail (1984a, 1984b).

8. Transport Canada/Ministry of Transportation and Communications (1979), Transport Canada (1979).

9. This section is based on Canadian Transport Commission (1970).

10. For a more complete description of the dogit model, see Gaudry and Dagenais (1979).

11. Gaudry and Wills (1979), p. 165.

12. Oum and Gillen (1983), pp. 184-85.

13. For more detailed discussion of disaggregate choice modelling theory, methodology and applications see, for example, Domencich and McFadden (1975), Hensher and Johnson (1981), Kanafani (1983), Manski and McFadden (1984) and Ben-Akiva and Lerman (1985).

14. For more detailed reviews of these models, see Hartgen and Cohen (1976), Rice et al. (1981) and Koppelman et al. (1984).

15. This section is based on Transportation Development Agency (1976).

16. This discussion is based on Ridout (1982) and Ridout and Miller (1989).

17. "Best" is defined in terms of statistical significance and agreement with a priori expectations of the parameter estimates, overall goodness-of-fit of the model and explicit statistical tests comparing the goodness-of-fit of competing model specifications.

18. See, for example, Miller and Cheah (1991).

19. This discussion is based on Wilson et al. (1990).

20. One of the most important differences is the lack of access/egress terms in the Wilson et al. models. This is due to a lack of sufficient information in the CTS data base to compute such terms. This is the single biggest weakness of the CTS data base and is, in fact, the reason why Ridout and Miller did not use it for their modelling work.

21. This discussion is based on Abdelwahab et al. (1991) and Abdelwahab (1990).

22. Updating involves statistically adjusting parameters estimated for one region using (typically limited) information concerning the new region to which the model is to be applied. For discussion of these methods see, for example, Atherton and Ben-Akiva (1976) and Koppelman and Wilmot (1982, 1986).

23. See Abdelwahab (1990) for details of these transferability tests.

24. See, in particular, McCoomb (1983) for a detailed examination of the transferability of urban mode choice models within Canada.

25. This discussion is based on Peat Marwick (1990), Ontario/Quebec Rapid Train Task Force (1991) and Ellis (1990).

26. Such terms are also absent from the Ridout–Miller model, but in this case by definition, since the car mode is excluded from this model.

27. These bias terms are the equivalent of the constant or "*y*-intercept" term in a linear regression equation. In linear regression, the regression line always passes through the point defined by the mean values of the dependent and independent variables. If the *y* intercept is forced through zero, then the other coefficient(s) in the model (representing the slope(s) of the line with respect to the independent variable(s)) will be correspondingly biased. In MLE estimation of logit models, the model always reproduces the aggregate modal shares observed in the estimation sample. If the bias terms are omitted, then the other parameters in the model will be biased, analagous to the regression example.

28. A similar argument might be made for other model parameters, but it is a much less persuasive one. In particular, the new mode's travel times and costs are likely to fall within the overall range of times and costs already experienced by travellers within the system. Thus, the model, if otherwise "properly" constructed, should be capturing these modal service trade-offs adequately. The constants, however, have buried within them the particular set of unobserved characteristics that exist within the current modes. A significantly upgraded or new mode is likely to have quite a different set of these unobserved characteristics and, hence, quite a different modal constant.

29. This can be contrasted with the urban case in which new mode introduction is rarely the issue. Rather, urban models are used to examine alternative expansions of existing modal networks (that is, road and transit), a situation in which the transferability of historical model parameters — including the modal constants — into future contexts is a more readily acceptable assumption.

30. This discussion is based on Koppelman (1989).

31. For more detailed discussions of the nested logit model and its derivation from random utility theory, see, among others, Ben-Akiva and Lerman (1985). In general, the nested logit can be viewed as a generalization of the ordinary logit model that permits complex decision structures to be modelled in a theoretically consistent yet practical way.

32. Koppelman and Hirsh (1986). For a similar discussion of these issues, see Rice et al. (1981).

33. Similar problems exist with publicly available data sets in Canada, as indicated by the difficulties encountered by Ridout and Miller, and Wilson et al. in their modelling efforts. For a detailed discussion of data-related issues in intercity passenger travel demand modelling in Canada, see Miller (1985).

34. This issue exists despite the extensive experience in the market research field with stated preference methods (see, for example, Green and Srinivasan (1978) and Cattin and Wittink (1982)). For a detailed discussion of the strengths and weaknesses of both revealed and stated preference data in travel demand models, see Ben-Akiva et al. (1990).

35. See VIA Rail (1984a, 1984b).

36. This discussion is based on Transportation and Economic Management Systems, Inc. (undated) and Transportation Management Systems, Inc./Benesch (1991).

37. See any text dealing with disaggregate logit models, for example, Ben-Akiva and Lerman (1985).

38. This discussion is based on VIA Rail (1989a, 1989b, 1989c, 1989d).

39. This section is based on Brand et al. (1991).

40. These are the same VOTs. In Table 4-14 they are expressed in 1990 U.S. dollars, while in Table 4-21 they are expressed in their original units of 1988 Canadian dollars.

41. Although not explicitly discussed within the model review chapters, current models typically involve the use of a zone system for each urban area that is sufficiently detailed to permit reasonably accurate access/egress times/costs to be calculated for each intercity travel mode. See, for example, VIA Rail (1989a) and Transportation Economic Management Systems, Inc. (undated). This can be contrasted with earlier, aggregate models in which a single set of average access/egress times/costs for each city-pair would be used.

42. In a West German study not previously referenced in this review, Brog (1982) similarly reported "surprisingly little impact" of socio-demographic variables on personal intercity mode choice behaviour. This result was obtained from a situational approach to the problem, based on detailed attitudinal survey results, rather than on econometric models such as the ones reviewed in this paper. Thus, this result also appears to be relatively robust across analysis methodology.

43. The one significant exception to this statement involves the continuing investigations of Gaudry on the use of aggregate, very generalized functional forms (Gaudry 1989, 1990). This approach does not appear to be popular with the developers of operational intercity models, probably due to the econometric and computational complexities involved. As discussed further in the next section, however, much of this work is potentially transferable to a more disaggregate, operational environment within the structured modelling approach discussed here.

44. As discussed in subsection 4.5.1, this system of models may or may not be consistent with the nested logit model formulation, depending on the model application.

45. Theoretical problems may also exist. In particular, use of these generalized functional forms sometimes implies loosening the ties between the empirical model and micro-economic utility theory which usually underlies the empirical model and which provides the empirical model with much of its a priori plausibility.

REFERENCES

Abdelwahab, W. M. 1990. "Transferability of Intercity Disaggregate Mode Choice Models in Canada." *Canadian Journal of Civil Engineering* 18, pp. 20–26.

Abdelwahab, W. M., J. D. Innes and A. M. Stevens. 1991. "Development of Disaggregate Mode Choice Models of Intercity Travel in Canada." Unpublished manuscript. Fredericton: Department of Civil Engineering, University of New Brunswick.

Atherton, T. and M. E. Ben-Akiva. 1976. "Transferability and Updating of Disaggregate Travel Demand Models," *Transportation Research Record* 610, pp. 12–18.

Ben-Akiva, M. E. and S. R. Lerman. 1985. *Discrete Choice Analysis.* Cambridge, Mass.: MIT Press.

Ben-Akiva, M. E., T. Morikawa and F. Shiroishi. 1990. "Analysis of the Reliability of Stated Preference Data in Estimating Mode Choice Models." *Proceedings of the Fifth World Conference on Transport Research.* Evanston, Ill.: World Conference on Transport Research Society.

Brand, D., T. E. Parody, P. S. Hsu and K. Tierney. 1991. "Forecasting High Speed Rail Ridership." Paper presented at the 71st Annual Meeting of the Transportation Research Board, National Research Council, January 1992. (NCHRD Report 1341.)

Brog, W. 1982. "The Application of the Situational Approach to Depict a Model of Personal Long-Distance Travel." Paper presented at the 61st Annual Meeting of the Transportation Research Board, Washington, D.C., January 1982.

Cattin, P. and D. R. Wittink. 1982. "Commercial Use of Conjoint Analysis: A Survey." *Journal of Marketing* 46, pp. 44–53.

Cohen, G. S., N. S. Erlbaum and D. T. Hartgen. 1978. "Intercity Rail Travel Models." *Transportation Research Record*, no. 673, pp. 21–25.

Canadian Transport Commission. 1970. *Intercity Passenger Transport Study*. Ottawa: Canadian Transport Commission, Research Branch. September 1970.

Domencich, T. and D. McFadden. 1975. *Urban Travel Demand — A Behavioral Analysis*. Amsterdam: North-Holland.

Ellis, R. 1990. "Intermodal Substitution and High-Speed Rail: Evidence from the United States and Lessons for Canada." In *Canadian Transportation Policy*. Edited by D. W. Gillen. Policy Forum Series, no.18. Kingston: Queen's University, John Deutsch Institute for the Study of Economic Policy, pp. 51–60.

Ellis, R. H., P. R. Rassam and J. C. Bennett. 1971. "Consideration of Intermodal Competition in the Forecasting of National Intercity Travel." *Highway Research Record*, no. 369, pp. 253–61.

Gaudry, M. J. I. 1989. "Asymmetric Shape and Variable Tail Thickness in Multinomial Probabilistic Response: Three Model Type Families in a Quasi-Direct Format Application to Intercity Travel Demand with Aggregate Canadian Data." Draft in progress. Montreal: University of Montreal, Centre for Transportation Research, June 1989.

—. 1990. "Three Families of Mode Choice Models Applicable to Intercity Travel Demand with Aggregate Data." In *Canadian Transportation Policy*. Edited by D. W. Gillen. Policy Forum Series, no.18. Kingston: Queen's University, John Deutsch Institute for the Study of Economic Policy, pp. 43-50.

Gaudry, M. J. I. and M. G. Dagenais. 1979. "The Dogit Model." *Transportation Research* 13B, pp. 105–11.

Gaudry, M. J. I. and M. J. Wills. 1978. "Estimating the Functional Form of Travel Demand Models." *Transportation Research* 12, pp. 257-89. (Figure 3-1 and Tables 3-2, 3-3, 3-4 reprinted with permission from Pergamon Press Ltd.)

—. 1979. "Testing the Dogit Model with Aggregate Time-Series and Cross-Sectional Travel Data." *Transportation Research* 13B, pp. 155-66. (Table 3-5 reprinted with permission from Pergamon Press Ltd.)

Grayson, A. 1981. "Disaggregate Model of Mode Choice in Intercity Travel." *Transportation Research Record,* no. 835, pp. 36–42.

Green, P. E. and V. Srinivasan. 1978. "Conjoint Analysis in Consumer Research: Issues and Outlook." *Journal of Consumer Research* 5, pp. 103–23.

Hartgen, D. T. and G. S. Cohen. 1976. *Intercity Passenger Demand Models: State of the Art.* Preliminary Research Report 112. Albany, N.Y.: New York State Department of Transportation, Planning Research Unit. December 1976.

Hensher, D. A. and L. W. Johnson. 1981. *Applied Discrete-Choice Modelling.* London: Croom Helm.

Hensher, D. A., P. O. Barnard and T. P. Thruong. 1988. "The Role of Stated Preference Methods in Studies of Travel Choice." *Journal of Transport Economics and Policy* 12, pp. 45–58.

Hutchinson, B. G. 1991. *Analyses of Canadian Air Travel Demands.* Unpublished manuscript. Waterloo: University of Waterloo, Department of Civil Engineering.

Kanafani, A. 1983. *Transportation Demand Analysis.* New York: McGraw-Hill.

Koppelman, F. S. 1989. "Models of Intercity Travel Choice Behavior." Paper presented at the 68th Annual Meeting of the Transportation Research Board, National Research Council, Washington, D.C., January 1989. (NCHRD Report 1241.)

Koppelman, F. S. and M. Hirsh. 1986. "Intercity Passenger Decision Making: Conceptual Structure and Data Implications." *Transportation Research Record*, no. 1085, pp. 70–74.

Koppelman, F. S., G. K. Kuah and M. Hirsh. 1984. *Review of Intercity Passenger Travel Demand Modelling: Mid-60's to the Mid-80's.* Evanston, Ill.: Northwestern University, The Transportation Center.

Leake, G. R. and J. R. Underwood. 1978. "Comparison of Intercity Bi-Modal Split Models." *Transportation Planning and Technology* 5, pp. 55-69.

Louviere, J. J. and D. A. Hensher. 1982. "On the Design and Analysis of Simulated Choice or Allocation Experiments in Travel Choice Modelling." Working Paper 50. Iowa City: University of Iowa, The Institute of Urban and Regional Research.

Louviere, J. J., D. H. Henly, G. Woodworth, R. J. Meyer, I. P. Levin. J. W. Stones, D. Curry and D. A. Anderson. 1981. "Laboratory-Simulation Versus Revealed-Preference Methods for Estimating Travel Demand Models." *Transportation Research Record*, no. 794. pp. 42–51.

Manski, C. and S. R. Lerman. 1977. "The Estimation of Choice Probabilities from Choice-Based Samples." *Econometrica* 45, pp. 1977–88.

Manski, C. and D. McFadden, eds. 1984. *Structural Analysis of Discrete Data with Econometric Applications.* Cambridge, Mass.: MIT Press.

McCoomb, L. A. 1983. "Analysis of the Transferability of Disaggregate Demand Models Among Ten Canadian Cities." *Transportation Forum* 3, 1, pp. 19–32.

Miller, E. J. 1985. "The Data Base for Intercity Passenger Travel Analysis: Current State and Future Prospects." Paper presented at the 20th Annual Conference of the Canadian Transportation Research Forum, Toronto, May 1985.

Miller, E. J. and L. S. Cheah. 1991. "An Operational, Integrated, Mode Choice and Route Assignment Model." Paper presented at the 71st Annual Meeting of the Transportation Research Board, Washington, D.C., January 1992.

Morikawa, T., M. E. Ben-Akiva and K. Yamada. 1991. "Forecasting Intercity Rail Ridership Using Revealed Preference and Stated Preference Data." Paper presented at the 70th Annual Meeting of the Transportation Research Board, Washington, D.C., January 1990.

Morrison, S. A. and C. Winston. 1983. "An Econometric Analysis of the Demand for Intercity Passenger Transportation." Draft manuscript.

Ontario/Quebec Rapid Train Task Force. 1991. *Ontario/Quebec Rapid Train Task Force Final Report.* Ontario/Quebec Rapid Train Task Force. May 1991.

Oum, T. H. and D. W. Gillen. 1983. "The Structure of Intercity Travel Demands in Canada: Theory, Tests and Empirical Results." *Transportation Research* 17B, 3, pp. 175–91. (Table 3-6 reprinted with permission from Permagon Press Ltd.)

Peat Marwick. 1990. *Analysis of the Market Demand for High Speed Rail in the Quebec/Ontario Corridor.* Final report to Ontario/Quebec Rapid Train Task Force. Vienna, Va.: KPMG Peat Marwick. June 1990.

Rice, R. G., E. J. Miller, G. N. Steuart, R. Ridout and M. Brown. 1981. *Review and Development of Intercity Passenger Travel Demand Models.* Research Report no. 77. Toronto: University of Toronto/York University Joint Program in Transportation. June 1981.

Ridout, R. 1982. *Development of Disaggregate Intercity Passenger Mode Split Models.* M.Eng. project report. Toronto: University of Toronto, Department of Civil Engineering.

Ridout, R. and E. J. Miller. 1989. "A Disaggregate Logit Model of Intercity Common Carrier Passenger Modal Choice." *Canadian Journal of Civil Engineering* 16, pp. 568-75.

Stephanedes, Y. S., V. Kummer and B. Padmanabhan. 1984. "A Fully Disaggregate Mode-Choice Model for Business Intercity Travel." *Transportation Planning and Technology* 9, p.1.

Stopher, P. R. and J. N. Prashkter. 1976. "Intercity Passenger Forecasting: The Use of Current Travel Forecasting Procedures." *Transportation Research Forum Proceedings 17th Annual Meeting* Vol. XVII, No. 1, pp. 67-75.

Transport Canada. 1979. *SOMPS: An Outline of PERAM Forecasts for Domestic Travel by Mode.* Draft preliminary working paper. Ottawa: Transport Canada, Strategic Planning Group.

Transport Canada/Ministry of Transportation and Communications. 1979. *Southern Ontario Multimodal Passenger Studies.* Ottawa: Transport Canada and Minstry of Transportation and Communciations, Ontario. September 1979.

Transportation Development Agency. 1976. *Mode Choice for Intercity Passenger Travel Montreal–Ottawa.* Working paper, TP 234, TDA 75. Montreal: Transportation Development Agency. August 1976.

Transportation and Economic Management Systems. Undated. *The COMPASS System, Demand Forecasting and Strategic Economic Analysis System.* Hamilton: Transportation and Economic Management Systems, Inc.

Transportation Management Systems Inc./Benesch. Undated. *Tri-State High Speed Rail Study Chicago–Milwaukee–Twin Cities Corridor.* Final project report to the Illinois, Minnesota and Wisconsin departments of transportation. Great Falls, Va.: Transportation Management Systems, Inc.

VIA Rail. 1984a. *High-Speed Passenger Rail in Canada.* Summary report. Montreal: VIA Rail Canada. February 1984.

—. 1984b. *High-Speed Passenger Rail in Canada.* Vol. II, Appendices. Montreal: VIA Rail Canada. April 1984.

—. 1989a. *Passenger Rail Demand Forecasting Study.* Final report. 1989 Rail Passenger Review working paper, T015-6-32, Toronto: Cole, Sherman and Associates, Ltd. July 1989.

—. 1989b. *Review of Passenger Rail Transportation in Canada.* Montreal: VIA Rail Canada. July 1989.

—. 1989c. *Review of Passenger Rail Transportation in Canada.* Supplementary vol. I. *Approach.* Montreal: VIA Rail Canada. September 1989.

—. 1989d. *Review of Passenger Rail Transportation in Canada.* Supplementary vol. III. *Corridor Services.* Montreal: VIA Rail Canada. September 1989.

Watson, P. L. 1972. "The Value of Time and Behavioural Models of Mode Choice." Ph.D. dissertation. Edinburgh: University of Edinburgh.

—. 1974. "Comparison of the Model Structure and Predictive Power of Aggregate and Disaggregate Models of Intercity Mode Choice." *Transportation Research Record,* no. 524.

Wills, M. J. 1981. "On the Equivalence of Common Rearrangements for Sum Constrained Travel Estimation Problems." Ottawa: Transport Canada, Strategic Planning Group.

Wilson, F. R., S. Damodaran and J. D. Innes. 1990. "Disaggregate Mode Choice Models for Intercity Passenger Travel in Canada." *Canadian Journal of Civil Engineering* 17, pp. 184–91.