ISSN 0250-7625

IIASA

REPORTS

International
Institute for
Applied
Systems
Analysis

Volume 4 Number 2 October-December 1981
Completing Volume 4

IIASA
REPORTS

Volume 4 (1981)
Completing Volume 4

Cover

Photograph: Robert A. Duis

International Standard Serial Number 0250-7625

IIASA Reports presents research conducted at IIASA that has been independently reviewed before publication. However, the views and opinions it expresses are not necessarily those of the Institute or the National Member Organizations that support it or of the institution that sponsored the research.

AN ECONOMIC ANALYSIS OF SUPPLEMENTARY IRRIGATION IN SKÅNE

Robert J. Anderson, Jr.
International Institute for Applied Systems Analysis, Laxenburg, Austria

SUMMARY

This report analyzes the water demand for supplementary irrigation in Skåne. Using water balance models, recent IIASA studies of Skåne demonstrated that agricultural water use could be a critical factor in future water management decisions in this region, and raised questions about possible economic effects:

- *What is the potential demand for irrigation water at current (1978) crop prices and irrigation costs?*
- *What effect would this level of irrigation have on the market for irrigated crops, and how would the changed market conditions in turn affect the demand for irrigation?*
- *What effect would a significant increase in the cost of irrigation have on the quantity of water used for that purpose?*

This report responds to these questions. However, for several reasons, the answers are tentative. The data on which they are based are seriously incomplete. The quantitative analysis determines only the demand for water per hectare of crop area; the analysis that determines the land areas planted in different crops is strictly qualitative. The estimates presented cover only two of the crops irrigated in Skåne: table potatoes and sugar beets.

Subject to these important caveats, the analysis shows that, at roughly current crop prices and water costs, irrigation demands may indeed be as great as those calculated using simple water balance models. The analysis thus supports the conclusion of related IIASA studies that potential water supply and demand in this region could become seriously out of balance.

The analysis also shows that irrigation would have little effect on the market for irrigated crops. Hence, there would be little hope that this kind of feedback effect would contribute much to closing the gap between potential demand and supply.

The sole remaining options for balancing supply and potential demand are reallocating water from other users and/or expanding the capacity of the water supply system. In all

cases, the result would be a substantial increase in the opportunity cost of irrigation. The analysis shows that the demand for irrigation probably would be reduced substantially if irrigation costs were increased to reflect the opportunity costs of reallocating existing supplies or of expanding capacity.

The conclusions of this report thus reinforce those from other studies conducted as a part of IIASA's analysis of regional water management in Skåne. Agricultural water demand is an important – perhaps even the critical – factor in future planning and management of the water supply system in this region.

---

1 INTRODUCTION

Recent work at IIASA (Arthur 1980, and Strzepek 1981) raises the possibility that future use of water for supplementary irrigation in Skåne will severely stress the current water resources of that region. Arthur showed that the irrigation rules now being recommended to farmers in Skåne would result in average irrigation water usage on irrigated acreage of from 86 to 194 millimeters per hectare (mm/ha), depending upon the crop. Strzepek converted these average figures into totals for the regions of the Kävlinge River Basin, added estimates of water demands for other purposes, and compared these totals to the estimated water yield of the basin. He found that the water supply system frequently did not yield enough water to satisfy all demands simultaneously.†

These calculations portend a serious imbalance of water supply and potential demand in Skåne. They certainly indicate that a thorough analysis of potential irrigation water demands in this region is in order. Any such analysis should include an examination of the effects of economic factors such as crop prices and irrigation costs.

Figure 1 shows the major relationships and variables that should be considered in a complete investigation. Broadly speaking, irrigation demands are derived from market demand for crops which can be produced using irrigation water, and other inputs such as fertilizer and seed.†† In outlining the system, it is useful to think of price and output determination as a cyclical series of four steps, as shown in Figure 1.

1. Demands for crops, along with relationships that describe the ways in which inputs can be combined to produce crop outputs, result in demands for inputs.
2. Demands for inputs, together with input supplies determine input prices and quantities of inputs used for crop production. For example, these relationships determine the quantities of land planted in each crop, and the quantities of seed, water, and fertilizer applied to each hectare.

---

†In a simulation covering 75 years, Strzepek (1981) found that, in 83 percent of the years, the yield of the basin would be insufficient to meet the sum of potential irrigation usage [calculated by multiplying the usages per hectare reported in Arthur (1980) by corresponding crop areas in 1976, and multiplying this result by the Malmöhus County Board of Agriculture's estimates of the percentages of crop areas that potentially will be irrigated in Malmöhus County], 1976 levels of municipal and industrial demand, and water-quality-related stream-flow regulations.

††Section 2 gives a more complete explanation of the economic relationships shown in Figure 1.

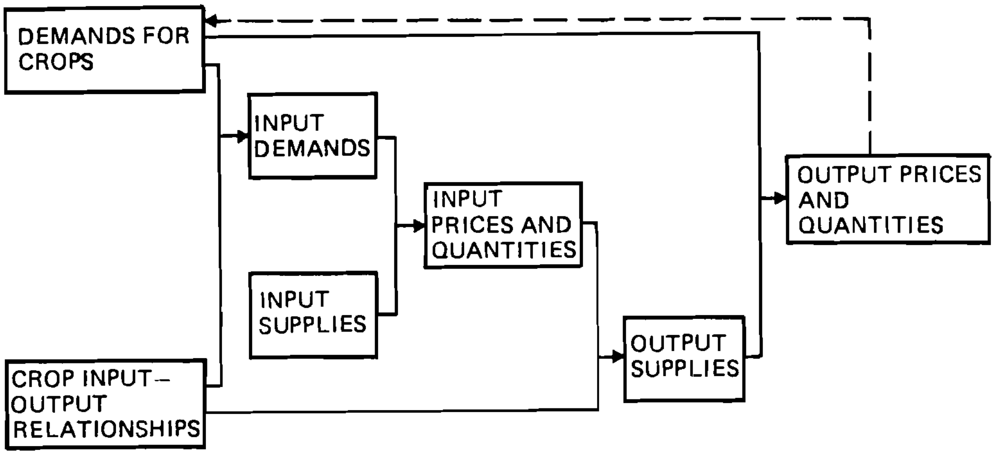


FIGURE 1 Economic determinants of irrigation usage.

3. Input quantities and prices, together with crop production input–output relationships, determine crop supplies.
4. Crop supplies and demands interact to produce market prices and quantities of crops.

An important implication of Figure 1 is that input and output prices and quantities in related markets are interdependent. In general, effects occurring in one market have ramifications in all other markets. For example, increased use of irrigation water to produce, say, potatoes would increase the yield and production of potatoes; this in turn may lead to a decrease in the price of potatoes, which would reduce the land area planted in potatoes and the demands for nonland inputs, including irrigation water. In the absence of additional shocks, the final outcome would be a readjustment of prices and quantities in all markets. Any complete investigation on the balance of water supply and demand must attempt to account for all important market adjustment mechanisms that could help this balance.

In spite of its seeming comprehensiveness, Figure 1 simplifies the market adjustment process in two important respects. First, some adjustment mechanisms have been omitted. For example, we have ignored the role that prices play in influencing the state of the technical arts for transforming inputs into outputs. In general, when scarcities arise, the search begins for technologies that will economize on the use of scarce resources. If water were scarce, agronomists would seek crop varieties less critically affected by water. We have ignored this type of linkage mechanism in Figure 1, since it generally occurs only over relatively long periods of time whereas our concern here is with the relatively immediate future.

A second simplification is that Figure 1 does not explicitly show the many nonmarket factors that affect the determination of prices and quantities. One might imagine these factors as being represented by the white space on the page that engulfs the forces explicitly represented. This image would be appropriate. Nonmarket factors, such as price supports, or restrictions on the quantity of land planted in a particular crop, modulate and in some instances overwhelm the market forces represented in the diagram. These nonmarket factors are particularly important in Swedish agriculture today, as we see in subsequent sections of this report.

Although Figure 1 is a simplified representation of market adjustment processes, an examination of all of the factors indicated in it is well beyond our means. Given the information at our disposal, the most that can be attempted is a partial analysis of the influence of economic and selected nonmarket factors on the demand for irrigation water in Skåne.

In particular, this report attempts to answer three questions. First, what would be the level of potential demand for irrigation water at current crop prices and irrigation costs? The estimates presented by Arthur (1980) are based on application of irrigation rules that are designed to maximize yield. These rules do not take into account economic factors such as the cost of irrigation, incremental yields due to irrigation, and additional farm income associated with incremental yields. When these other factors are accounted for, what level of demand for irrigation would be expected?

Second, what effect would irrigation have on the market for irrigated crops, and how would these altered market conditions affect the demand for irrigation? The range of possible market effects, depending upon particular conditions in the relevant markets and the agricultural policies that apply to them, includes expanded crop production accompanied by constant or falling prices, and possible increases in the cost of crop price support programs. The nature and magnitude of these market effects are extremely important to the balance of water supply and demand. Depending upon the form they take and the size they assume, crop market effects could either tend to moderate or to intensify the demand for irrigation water.

Third, what effect would a significant increase in the cost of irrigation have on the quantity of water demanded? Balancing supply and demand may well require that existing water supplies be reallocated and/or that capacity be expanded. In all instances, the cost of additional water may be substantially above current water costs, and this increase in cost, if allowed to affect irrigation decisions, could also help to balance supply and demand.

Our answers to these questions are tentative for several reasons. The data on which our estimates and analyses are based are seriously incomplete. We analyze quantitatively only the determination of the demand for water per hectare of crop area; our analysis of the determination of land areas planted in different crops is strictly qualitative. Our quantitative estimates cover only two of the crops – table potatoes and sugar beets – that are irrigated in Skåne.

Subject to these important caveats, our analysis shows that the levels of irrigation water demand projected by Arthur (1980) are consistent with the levels of demand one would project based upon an economic model of irrigation water demand, assuming 1978 crop prices and irrigation costs.† Our analysis thus supports the conclusion that, at current (1978) crop prices and water costs, water supply and potential water demand in Skåne could become seriously out of balance.

Our analysis also shows that the effects of irrigation on crop markets probably would contribute little to closing the gap between water supply and potential water demand. We

†1978 is the latest year for which the relevant published economic data were available at IIASA.

show that even if crop prices fell as a result of increased production, the resulting price decreases probably would not have much of an effect on irrigation water demand. If crop price supports restrained any tendency of prices to fall, this adjustment mechanism would be of no help in rebalancing water supply and demand.

The sole remaining options for balancing supply and demand for water are reallocation of water from other users (e.g., residential, commercial, industrial, environmental, and recreational users), and/or expansion of the capacity of the water supply system. In all cases, the result is likely to be a substantial increase in the opportunity cost of irrigation. Where balancing is effected through reallocation, these costs may take the form of inconvenience or even hardship on other water users as they reduce their water usage. Although total costs of system operation may not seem to go up, this loss of opportunity to use water – or opportunity cost – is a very real and probably substantial cost, and should be considered in analyzing this alternative for balancing supply and demand. Of course, the costs of expanded capacity are easier to identify and obviously are large.

Our analysis suggests that the demand for supplementary irrigation probably would be reduced substantially if irrigation costs were increased to reflect the opportunity costs of reallocation of existing supplies or capacity expansion. This conclusion is strengthened if possible irrigation-induced crop price decreases are considered simultaneously.

It is tempting to reach beyond these conclusions concerning the prospective demand for irrigation water and to make conclusions about appropriate public policy. For example, some readers might conclude: "These results show that potential water demand will exceed water supply. Therefore we must reallocate water or expand capacity." Or other readers might conclude: "These results show that if prices reflecting the full opportunity cost of resources were to prevail, there would be no imbalance between water supply and potential demand. Reallocation to agriculture or capacity expansion are economic wastes, and should not be undertaken."

Readers should resist making conclusions about public policy: conclusions concerning what should or should not be done depend upon the objectives of Swedish public policy. We make no attempt here either to identify these objectives or to reach any conclusions concerning appropriate policy.

In Section 2 of this report, the basic model for our analysis is developed. This model involves two important, and perhaps controversial, approximations. The first is an approximation of the relationship between water input and crop yield via a function that relates yield to total seasonal water inputs. The second is an approximation of irrigation decision-making under uncertainty. Section 2 explains the basis for these approximations and their effects on the results of our analysis.

Section 3 presents empirical estimates of the parameters required to estimate irrigation water demands for two crops, table potatoes and sugar beets. These crops are considered by Swedish agricultural experts to be the best candidates for expanded use of supplementary irrigation in Skåne. Three kinds of parameter estimates are presented. First, parameters of empirical distributions of precipitation over the growing seasons for these crops are estimated from historical data on precipitation measured at Lund. Our statistical analyses of these data show that the probability distribution of precipitation during the growing season for each crop can be approximated conveniently and satisfactorily by the Weibull distribution. Second, empirical estimates of seasonal water input–yield relationships are presented for table potatoes and sugar beets. These relationships are based upon the reported results of irrigation experiments conducted in southern Sweden. Very few data points were available for this purpose, so the estimated yield relationships presented here are subject to a great, although unquantified, amount of uncertainty. Third, estimates of the fixed and variable costs of irrigation are reported. The derivation of these estimates is described in detail in Appendix A.

Section 4 presents estimates of the demand for irrigation water for table potatoes and sugar beets using the model presented in Section 2 and the parameter estimates reported in Section 3. Expected demands of approximately 87 mm/ha for both crops are obtained, assuming that 1978 crop prices and irrigation costs prevail. These estimates are reasonably close to those reported in Arthur (1980).

Section 4 also examines the contribution of irrigation to farm income. This contribution, again assuming 1978 crop prices and irrigation cost levels, is found to be more than sufficient to cover the fixed costs associated with irrigation. However, for reasons that are explained in Section 2, our estimates of contribution to farm income tend to overstate the contributions that could in fact be expected. Nonetheless, when adjustments are made for this overstatement, the conclusion stands that irrigation is profitable at 1978 crop price and irrigation cost levels.†

In Section 5 we consider the possible effects of irrigation on crop markets and "feedback" effects on irrigation demand. As previously noted, increased yields, other things being equal, could result in decreased crop prices. These price decreases, in turn, could moderate the demand for irrigation water. Section 5 presents some very rough estimates of the extent to which crop prices might fall as a result of irrigation in Skåne, assuming other factors remain constant. Our analysis shows that even sizeable price decreases probably would not result in a substantial decrease in the quantity of irrigation water demanded.

Section 6 examines the effect of markedly higher variable costs of irrigation on the quantity of irrigation water that would be demanded. Our calculations suggest that increases in costs to levels that would reflect either the opportunity cost of reallocating existing water supplies or the costs of capacity expansion would reduce substantially the quantity of water demanded per hectare. This result is reinforced if the simultaneous effects of increased crop yields on crop markets and an increase in the cost of water are considered. We conclude that increasing irrigation costs, e.g., through imposition of a charge on the use of water for irrigation, to reflect the opportunity cost of the water resources involved would make a substantial contribution to redressing the potential imbalance between water supply and demand in Skåne.

Section 7 offers some general but nonetheless qualified conclusions. Our analysis supports the conclusion that use of water for supplementary irrigation in Skåne is profitable at roughly current prices and costs. Growing awareness of this undoubtedly accounts, in part, for the recent rapid adoption of irrigation techniques among farmers in the region. Our results on the combined effects of crop price changes in response to increased yields and increases in the cost of irrigation suggest that economic factors could come into play that would reduce or eliminate altogether the opportunities to employ supplementary irrigation profitably, and thereby markedly reduce the potential demand for irrigation water.

†Anderson (1980) developed an approximation that may be used to adjust estimates of the contribution of irrigation to farm income to eliminate, approximately, the overstatement.

2 THE MODEL

As noted in Section 1, water is demanded for irrigation because irrigation increases crop yields and, consequently, farm income. Thus, one important factor in calculating how much water is likely to be demanded is the quantitative relationship between water inputs, yield increases, and farm income increases.

Another important factor in estimating irrigation demand is the extent to which precipitation satisfies the water requirements of crops. There are three important aspects of the relationship between precipitation and irrigation demand. First, precipitation varies randomly. During some periods, precipitation is relatively great, and the need to supplement it with irrigation is correspondingly reduced; during other periods precipitation is relatively low, and the need to supplement it is great. As a consequence, irrigation demand also varies randomly. Second, farmers do not know exactly how much crop-usable water precipitation will yield. Thus, irrigation decisions must be made in the presence of uncertainty about the quantity of water that will be supplied by precipitation. Third, the effectiveness of precipitation in promoting crop growth varies, depending upon a number of other conditions.

This section explains how we model the relationship between water inputs and crop yields, and how we treat the various aspects of precipitation as a source of water input. The actual relationships between these variables are complex and our modeling of them is therefore at best approximate.

2.1 The Relationship between Water Input and Crop Yield

In general, the effect of water inputs on crop yields depends upon the crop variety, the type of soil, solar radiation, and upon the temporal pattern of application of the water. It also depends upon other soil and climatic factors, and upon subtle genetic differences in plants.

Several detailed models of crop–water relationships that attempt to incorporate one or more of these factors have been developed. In most of these models, the fundamental premises are that each crop variety has a genetically determined maximum potential yield (denoted by Y_M), and that actual yields below this maximum potential yield are the results of water stresses on the crop. The models differ primarily in the mathematical form given the water stress–crop yield relationship, and in the variables chosen to characterize this relationship.

Four basic concepts have been found to be useful in describing and modeling the effect of water on plant growth. The first concept is *permanent wilting point*. This is the moisture content of a given soil at which the leaves of a given type of plant growing in that soil become permanently wilted. This happens when the moisture in the soil falls to levels so low that the rate of transpiration exceeds the rate at which the plant is able to extract water from the soil.

The second concept is *field capacity*. This is defined as the quantity of water held in the root zone by the soil against gravity when the soil is allowed to drain freely. Clearly field capacity also depends upon both soil type and crop.

The third concept is *soil moisture tension*. This is the force with which water adheres to soil particles. The higher the moisture tension, the greater the force with which moisture is "bound" to the soil.

The fourth and final concept is *evapotranspiration*. This is the evaporation of water from soil and the transpiration of water by plants.

These four concepts are related to one another. For example, the closer a given soil layer is to field capacity, the lower the soil moisture tension in this layer, and the greater the rate of evapotranspiration. The higher the soil moisture tension, the greater the difficulty plants have in making use of this moisture and the lower the rate of evapotranspiration. The permanent wilting point is reached when plants are no longer able to overcome the forces that bind moisture to the soil.

There are two main theories concerning the precise nature of the relationship between water availability and plant growth. The first theory, which has been called the *equal availability theory*, holds that variations in soil moisture between the permanent wilting point and field capacity have no effect on yield (Veihmayer and Hendrickson 1955). This theory implies that the aim of irrigation, ignoring cost factors and other constraints, should be to maintain just enough water in the field to insure that available water does not fall below the permanent wilting point.

The second theory holds that the rate of plant growth is inversely related to the level of soil moisture tension in the root zone of the plant (Hagan, Vaadia, and Russel 1959). High levels of tension retard plant growth, and completely terminate it at the permanent wilting point.

One of the earliest formulations of production functions for irrigated agriculture based upon these theories is due to Moore (1961). Moore noted that the agronomic theories described above imply that there is a relationship between the percentage of the maximum growth rate that is attained by a plant and the percentage of available moisture (i.e., the moisture between field capacity and the permanent wilting point) that is depleted in the field in which the plant is growing. Figure 2 illustrates the form of the relationship posited

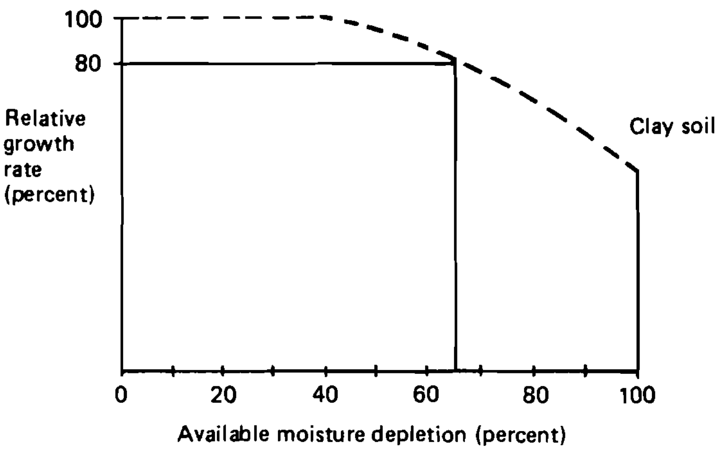


FIGURE 2 Relationship between soil moisture and plant growth.

by Moore in an hypothetical case. This figure shows that at 65 percent depletion of available moisture in clay soil, the growth rate of the plant is 80 percent of the maximum attainable by the plant. The rate of decrease of the growth rate percentage curve depends upon the effect of variations in water availabilities between field capacity and the permanent wilting point. Under the equal availability hypothesis, the curve would be flat at 100 percent of the maximum growth rate until 100 percent depletion is reached. Under the hypothesis that growth rates decline with depletion, the curve would begin to decline at lower depletion percentages.

Based upon variants of Moore's theory, several investigators have specified and/or estimated empirical relationships between water inputs – as measured by one or more of the concepts discussed earlier – and crop yield. For example, Hall and Butcher (1968) developed a model of the water–yield relationship that distinguishes between different stages in a plant's development, with overall growth being determined by multiplication of growth rates at different stages. The form of their model is given in eqn. (1)

$$Y(q) = \prod_{k=1}^{n} \alpha_k(\theta_k) Y_M(q_M) \tag{1}$$

where

- q is the total amount of water applied per unit area
- $Y(q)$ is the actual yield corresponding to application of q units of water
- q_M is the water required to maintain soil at field capacity
- $Y_M(q_M)$ is the maximum yield that can be obtained with an unlimited quantity of water
- θ_k is the available soil moisture during stage k
- k is the index of stages of growth
- n is the number of stages of growth
- $\alpha_k(\theta_k)$ is the function representing the effect of moisture deficiency during stage k on total yield

Minhas, Parikh, and Srinivasan (1974) specified the relationship between water and yield shown in eqn. (2)

$$\frac{Y}{Y_M} = \sum_{k=1}^{n} \left\{1 - \left[1 - \left(\frac{E}{E_p}\right)_k^2\right]\right\}^{b_k} \tag{2}$$

where

- $\left(\frac{E}{E_p}\right)_k$ is the ratio of actual evapotranspiration to potential evapotranspiration in stage k
- b_k is the coefficient measuring crop sensitivity to water deficits
- Y is the actual yield

Fitting this equation by regression methods, they determined that over 98 percent of the variation in experimental yields of wheat in India could be explained by the model.†

†Other papers that develop models of the water–yield relationship are Flinn and Musgreave (1967), Jensen (1968), Hiler and Clark (1971), and Hanks (1974).

In general, the literature establishes that it is possible, using yield experiment data, to obtain quite satisfactory empirical relationships between measures of water availability (e.g., available soil moisture depletion, the ratio of actual to potential evapotranspiration) and crop yield. Moreover, this literature holds out the promise that substantial improvements can be made in field-level irrigation management. The use of detailed field-specific and growth stage-specific water–yield relationships to improve irrigation management through sequential control of water inputs is amply illustrated in Burt and Stauber (1971) and Córdova and Bras (1979).

The ideal type of water–crop yield relationship for our examination of irrigation water demand in Skåne would be one that relates total seasonal water inputs (i.e., water inputs over the growing season) to yield. Strictly speaking, this can be done legitimately only if the intraseasonal distribution of water inputs is held fixed. (See Yaron 1971.) Nonetheless, it is possible to derive an approximate relationship between seasonal water input and yield even in cases where the intraseasonal distribution of water inputs is not held strictly fixed. Indeed, this is by far the most common practice in studies of the effects of water inputs on yields. (For example, see Hallgren 1971, and Hexem and Heady 1978.) We shall follow this common practice in further development of the model and in empirical investigations in subsequent sections of the report.

2.2 Precipitation and the Demand for Supplementary Irrigation

As noted at the beginning of Section 2, some portion of the water requirements of crops in Skåne is met by precipitation, with the balance to come from supplementary irrigation. We also noted that three aspects should be considered in examining the effects of precipitation on the demand for irrigation: (1) randomness in precipitation; (2) uncertainty about precipitation at the time irrigation decisions must be made; and (3) randomness in the effectiveness of precipitation in supplying water requirements to crops.

Let us consider each of these aspects in turn. Since we conduct our analysis in terms of the relationship between total water inputs over the growing season and crop yields, we are interested primarily in interseasonal randomness in precipitation. However, as has been discussed and is discussed further, intraseasonal randomness is an important determinant of the effectiveness of precipitation in supplying crop water requirements and, thus, requires some consideration.

Let us consider the implications of interseasonal randomness of precipitation. Let us suppose that there is an optimal (by some criterion as yet unspecified) quantity of total water input for a crop season, denoted by i^*. Let us also suppose that in each crop season, farmers know in advance exactly how much of this optimal level of water input will be supplied by precipitation. Then the quantity of irrigation water demanded would be either the difference between the optimum water input level and the level of precipitation, or zero, whichever is greater.

This situation is easily illustrated. Consider Figure 3, which depicts a hypothetical probability distribution of total precipitation for the season relevant to production of the crop under consideration. As illustrated, precipitation typically may vary over a wide range. The particular shape and position of the distribution depend upon the local climate. In the absence of irrigation, the distribution of the quantity of water input for agriculture and the distribution of precipitation are identical.

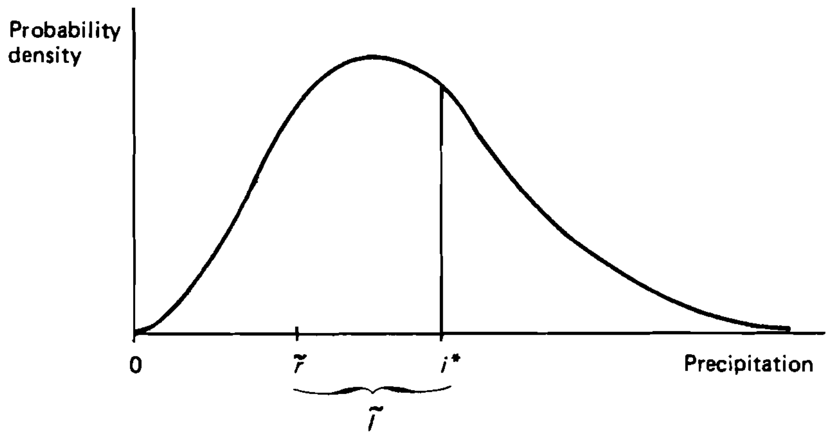


FIGURE 3 The effect of irrigation on the distribution of water.

If irrigation is undertaken, the situation is somewhat different. In any crop season in which precipitation (denoted by $\tilde{r}$) is less than i^*, irrigation water (denoted by $\tilde{I}$) is added to bring total water input up to the level i^*.

Two conclusions emerge concerning the effects of irrigation on water input, assuming perfect foreknowledge of precipitation. First, the quantity of irrigation water demanded will vary randomly from crop season to crop season depending upon the level of precipitation.† This follows from the fact that the quantity of irrigation water applied is adjusted to make up any deficit between the optimal total water input for the crop season i^* and the (random) level of total precipitation for the crop season $\tilde{r}$. Assuming perfect foreknowledge, the quantity of irrigation water applied is always exactly the correct amount needed to make up any gap between the optimal level of water input and precipitation. Second, the effect of irrigation under the perfect foreknowledge assumption is to alter the distribution of water input to the crop by chopping off the left-hand tail of the distribution. For example, if irrigation water is added to insure that available water is always at least i^* mm per hectare, then the probability density of water input if supplementary irrigation were practiced would be the right-hand tail of the precipitation distribution beyond i^*, scaled appropriately to possess the usual properties of a probability density function.

In practice, the irrigation plans applied in Skåne also shift the right-hand tail of the distribution of seasonal water inputs. Arthur (1980) shows that the irrigation operating rules currently being recommended to farmers in Skåne reduce the dispersion of the probability distribution of seasonal water inputs, and, in varying degrees, shift the entire distribution to the right in the direction of increased water inputs. As Arthur (1980) explains,

†The quantity of irrigation water demanded conditional on precipitation is deterministic. However, since precipitation varies from year to year, irrigation is a random variable with a probability distribution that may be derived from the probability distribution of precipitation. This is explained in more detail in Section 2.3.

the reason that this shift occurs is because irrigation decisions must be taken before precipitation is known with certainty.

In the completely general case, it is impossible to say precisely what effect uncertainty about the water input supplied by precipitation would have in an economic model of the quantity of irrigation water demanded. The specific effect would depend upon assumptions about the way in which randomness enters the model (e.g., multiplicative, additive), the information available to farmers at the time irrigation decisions must be made (e.g., precipitation forecasts), and the behavior of the farmers under uncertainty. In the absence of perfect foreknowledge, all that can be said with confidence is that the quantity of irrigation water applied would not in general be exactly the correct amount needed to make up a deficit between a target water input level and the actual level of precipitation. In some cases, more water than required would be added, in other cases, less water than required would be added.

In this analysis, we proceed as if precipitation were known with certainty at the time irrigation decisions are made. Since irrigation operating rules involve sequential control of irrigation water inputs in response to observed precipitation, and since short-term forecasts of precipitation are available, this assumption may be accepted as an approximation.

A third important consequence of the fact that precipitation is random is that its effectiveness as a source of water input varies. This effect is a result of intraseasonal randomness. For example, precipitation that occurs when soil moisture is already at field capacity contributes nothing to the water input of the crop. Indeed, it may cause injury through erosion with the run-off or, in extreme cases, waterlogging. Because precipitation does not come in carefully controlled doses, its effectiveness as a source of water input is generally lower than the effectiveness of irrigation.

Detailed field level models of sequential intraseasonal irrigation management, such as in Córdova and Bras (1979), incorporate hydrological balance models that represent the varying effectiveness of precipitation as a source of water input. Since our analysis is conducted in terms of total water input for the crop season, it is necessary to adopt a slightly different approach to incorporating this effect into our model. In particular, we allow for the difference in effectiveness of irrigation and precipitation as sources of water input by introducing a relative effectiveness parameter B.† This parameter always takes a value in the unit interval, reflecting the fact that precipitation, which is uncontrolled, is no more effective as a source of water input than irrigation. In general, it will be less effective. We shall then compute total effective water input as $W = I + Br$, where W is effective water input, I is irrigation, and r is precipitation.

To summarize, we shall make our calculations as though farmers had perfect foreknowledge about precipitation at the time irrigation decisions must be made. The only source of randomness in our model is thus natural variation in precipitation. While year to year variations in precipitation will influence the quantity of irrigation water demanded, the quantity applied in any year is – assuming perfect foreknowledge – always exactly the correct amount. This has the effect of biasing our demand estimates, although the direction of this bias cannot be determined without more information. However, since perfect foreknowledge also implies that irrigation water is never used needlessly, this assumption

†This parameter is not to be confused with "irrigation" efficiency, which measures the ratios between water inputs and water outputs, along the links in an irrigation system.

also has the effect of making our estimates of the contribution of irrigation to farm income too high.

2.3 The Demand for Irrigation

With these preliminaries, we can now analyze the demand for irrigation. Let us suppose that the relationship between yield per hectare and seasonal effective water input is quadratic as given in eqn. (3)

$$Y = a_1 W - a_2 W^2 \tag{3}$$

This equation is concave in water, and reaches a maximum at $a_1/2a_2$. It is similar in form to many of the functions obtained in empirical studies of water input–crop yield relationships. (For example, see Yaron 1971, Hexem and Heady 1978, and Hallgren 1947.)

Let us also suppose that the variable cost of irrigation is equal to C monetary units per hectare-millimeter (ignoring fixed costs for the moment), and that the net proceeds to the farmer from sale of one unit of crop are equal to p monetary units. If farm operators incurred costs of C monetary units per hectare-millimeter of irrigation water used, and if they sought to maximize expected farm income, the optimum level of irrigation water usage (assuming that the fixed costs were covered) would be

$$I \geqslant \frac{pa_1 - C}{2a_2 p} - Br = i^* - Br; \qquad I(i^* - Br) = 0 \tag{4}$$

This is the necessary condition of the level of irrigation water input that would maximize farm income (before fixed charges).†

Equation (4) leads to a simple rule for deciding when and how much to irrigate. In particular, if precipitation (adjusted for effectiveness) yielded less water than the amount i^* shown in eqn. (4), profits could be increased by "purchasing" irrigation water until the sum of precipitation (adjusted for effectiveness) plus irrigation water equaled i^*.

Our irrigation rule may be expressed thus

$$\begin{aligned} &\text{if } i^* \geqslant Br, \qquad I = i^* - Br \\ &\text{if } i^* < Br, \qquad I = 0 \end{aligned} \tag{5}$$

where I is irrigation, r is precipitation, B is the precipitation effectiveness parameter, and i^* is the optimality parameter in eqn. (4).

To examine the effect of irrigation on expected farm income, let us denote the probability density function of precipitation by $f(r)$. Then we may represent expected farm income when we follow the irrigation rule given in eqn. (5) above by the eqn. (6):

†Recalling that total water input is $W = I + Br$, and that r is assumed to be known at the time the irrigation decision is made, eqn. (4) is obtained by finding the maximum of the profit function, $\Pi = p[a_1 (I + Br) - a_2(I + Br)^2] - CI$ with respect to I, requiring that $I \geqslant 0$. Note that i^* in eqn. (4) is defined as $pa_1 - C/2a_2p$.

$$\Pi = F(i^*/B)\left\{p(a_1 i^* - a_2 i^{*2}) - C \int_0^{i^*/B} \frac{If(r)\,\mathrm{d}r}{F(i^*/B)}\right\}$$
$$+ [1 - F(i^*/B)]\left\{p\left[a_1 B \int_{i^*/B}^{\infty} \frac{rf(r)\,\mathrm{d}r}{1 - F(i^*/B)} - a_2 B^2 \int_{i^*/B}^{\infty} \frac{r^2 f(r)\,\mathrm{d}r}{1 - F(i^*/B)}\right]\right\} \quad (6)$$

where $F(x)$ is the probability that $r \leqslant x$. This equation, as clearly can be seen, has two main terms. The first term gives expected farm income in the event that effective precipitation (i.e., precipitation multiplied by the efficiency parameter B) is less than i^*. In this event, irrigation water is drawn bringing total available water supply to the level i^*, contributing $p(a_1 i^* - a_2 i^{*2})$ to expected income, and

$$C \int_0^{i^*/B} \frac{If(r)\,\mathrm{d}r}{F(i^*/B)}$$

to expected costs (again ignoring fixed costs). If precipitation exceeds i^*, then no irrigation water is drawn, and the contribution to expected farm income in this event is given by

$$p\left[a_1 B \int_{i^*/B}^{\infty} \frac{rf(r)\,\mathrm{d}r}{1 - F(i^*/B)} - a_2 B^2 \int_{i^*/B}^{\infty} \frac{r^2 f(r)\,\mathrm{d}r}{1 - F(i^*/B)}\right]$$

Each term is multiplied by the corresponding probability that the event indicated occurs (i.e., by $F(i^*/B)$ and $1 - F(i^*/B)$ respectively), and the two terms are added together to give expected farm income if irrigation is practiced.

If irrigation is not practiced, the corresponding expression for expected farm income is simply

$$\Pi_0 = F(i^*/B)\left\{p\left[a_1 B \int_0^{i^*/B} \frac{rf(r)\,\mathrm{d}r}{F(i^*/B)} - a_2 B^2 \int_0^{i^*/B} \frac{r^2 f(r)\,\mathrm{d}r}{F(i^*/B)}\right]\right\}$$
$$+ [1 - F(i^*/B)]\left\{p\left[a_1 B \int_{i^*/B}^{\infty} \frac{rf(r)\,\mathrm{d}r}{1 - F(i^*/B)} - a_2 B^2 \int_{i^*/B}^{\infty} \frac{r^2 f(r)\,\mathrm{d}r}{1 - F(i^*/B)}\right]\right\} \quad (7)$$

We have split the integrals in eqn. (7) into ranges in order to facilitate comparison of expected income with and without irrigation.

The increment to expected farm income from irrigation is determined by taking the difference between eqns. (6) and (7). When this is done, the following expression for incremental expected farm income due to irrigation is obtained.

$$\Delta\Pi = F(i^*/B)p\left\{(a_1 i^* - a_2 i^{*2}) - \left[a_1 B \int_0^{i^*/B} \frac{rf(r)\,\mathrm{d}r}{F(i^*/B)} - a_2 B^2 \int_0^{i^*/B} \frac{r^2 f(r)\,\mathrm{d}r}{F(i^*/B)}\right]\right\}$$
$$- C \int_0^{i^*/B} If(r)\,\mathrm{d}r \quad (8)$$

We have thus far ignored fixed costs in our analysis. Equation (8) gives the excess of expected revenues over expected variable costs obtained through irrigation according to the rule described by eqn. (5). Whether or not irrigation would make a net contribution to farm income after fixed costs are deducted thus depends upon the size of the increment calculated according to eqn. (8) relative to the size of fixed costs. If the increment calculated by eqn. (8) is greater than fixed costs, then irrigation would add to expected farm income. If not, then it will not add enough to cover fixed costs, and presumably would not be undertaken.

The model described here shows how irrigation demand per hectare of crop area and the contribution of irrigation to expected farm income may be calculated for any given value of net farm price p and variable cost of irrigation C. From the individual farmer's perspective, our calculations approximate the expected values of optimal irrigation water quantities and farm income assuming constancy in these parameters.

However, as our discussion of Figure 1 suggests demand and income may be different when a regional perspective is adopted. Two factors are important to mention here. First, irrigation increases expected yield of crops. These yield increases may, for example, affect crop market prices and quantities. That is, when the actions of all farmers in the region taken together are considered, we must allow for the possibility that the crop market conditions assumed in the derivation of individual farmer's irrigation demands change. In Section 5 we show how crop market effects could alter the results of the analysis.

A second factor mentioned in our discussion of Figure 1 is the possibility that expansion of the demand for some factor of production (irrigation water in this case) could necessitate an increase in the price charged for this factor. This possibility is discussed in Section 6, where the effects of increases in irrigation costs on irrigation water demand are considered.

3 PARAMETER ESTIMATES

Three types of parameter estimates are required in order to use the relationships developed in Section 2 to estimate the demand for irrigation water. These are estimates of the parameters of the probability distributions of precipitation over relevant time periods, estimates of the parameters of function relating seasonal water input to yield, and estimates of the fixed and variable costs of irrigation. Estimates of these parameters are presented in this section.

3.1 Precipitation Distributions

Probability density functions for precipitation over relevant time periods (see Table 1) were fitted to 75 years of data on precipitation at Lund. Inspection of the precipitation

TABLE 1 Periods for seasonal water input distributions.

Crop	Period
Potatoes	16 June–31 Aug
Sugar beets	1 July–15 Sept

SOURCE: Arthur 1980.

data (keeping in mind the expressions in Section 2 that require numerical calculation) suggested fitting Weibull densities to the data.

Using this form of density, the results reported in Table 2 were obtained. As can be seen from the table, the estimated densities fit the data well. The coefficients of determination are all relatively high, and chi-square tests fail to reject the hypotheses that the data were generated by Weibull densities with the parameter values reported. For example, in the case of the distribution of "potato season" precipitation, the probability of obtaining a chi-square statistic less than or equal to that obtained, when the null hypothesis is true, is 0.6080. This means that we could reject the null hypothesis only at significance levels of about 40 percent.

TABLE 2 Estimates of parameters of Weibull densities $g(r) = \gamma_0\gamma_1 r^{(\gamma_1-1)}e^{-\gamma_0 r^{\gamma_1}}$

Crop	Density parameters	R^2	Chi-square $\bar{\chi}^2$	Probability $P_6(\chi^2 \leq \bar{\chi}^2)$
Potatoes	$\gamma_0 = 0.96 \times 10^{-10}$ $\gamma_1 = 3.9818$	0.9498	14.8(14)	0.6080
Sugar beets	$\gamma_0 = 3.1 \times 10^{-10}$ $\gamma_1 = 4.1827$	0.9714	11.2(14)	0.3297

Taken together, the results reported in Table 2 indicate that our empirically estimated Weibull densities provide a good approximation to the observed distributions of precipitation.

3.2 Seasonal Water Input–Yield Functions

Seasonal water input–crop yield functions were fitted to experimental data (Swedish University of Agricultural Sciences 1966–1979; and Johansson and Linnér 1977). Data used to fit the parameters of water input–yield functions were taken from experiments conducted in southern Sweden. Very few experiments were available that could be used for this purpose. The seasonal water input–yield functions reported here therefore should be interpreted as rough approximations at best.

The procedure used in fitting the functions was as follows. Only data from experiments whose aim was to maximize yield were used.† Effective water inputs and yields

†In experiments designed to maximize yield, complementary inputs (e.g., fertilizer) frequently are applied in greater quantities than would be economical. In cases in which this occurs, estimates of the contribution of the treatment to output tend to be biased upward. It is probable, therefore, that our estimates of the incremental output due to irrigation overstate the increments that would actually be observed under normal farm operating conditions. However, we do not believe that this bias is very large in the present case.

were computed from experimental data. If several experiments were available, water inputs and yields were averaged. The parameters of a quadratic seasonal water input–yield function were then estimated by solving the following two equations for a_1 and a_2

$$Y = a_1 W - a_2 W^2$$

$$0 = a_1 - 2a_2 W$$

The efficiency parameter value B was selected for each crop by trial and error to approximate reported average yields with and without irrigation. The results obtained from these calculations are given in Table 3. The first three columns of the table report estimated

TABLE 3 Estimated parameters of water input–yield functions.

				Expected yield (dt/ha)	
Crop	a_1	a_2	B	Without irrigation	With irrigation
Potatoes	3.4826	0.0087	0.65	270.97	349.67
Sugar beets	4.1707	0.0091	0.75	373.72	477.86

values of a_1, a_2, and B. The last two columns report expected yield without irrigation and expected yield with irrigation. Thus, our estimates imply average yields in the absence of irrigation of approximately 270 dt/ha and 375 dt/ha for potatoes and sugar beets, respectively, and average yields with irrigation of approximately 350 dt/ha and 475 dt/ha, respectively.

3.3 Irrigation Costs and Crop Prices

Two types of economic parameters enter into the calculation of the demand for irrigation water. These are irrigation costs and net farm prices.

In Skåne today, there are no charges levied directly on the withdrawal of water from groundwater or surface water sources for irrigation. The water itself is free.† However, this does not mean that irrigation is free to the farmer. The withdrawal and application of irrigation water require investment in equipment and outlays for its operation.

Estimates of the investment and operating costs associated with water withdrawal and application are given in Table 4. As can be seen from the table, investment costs are estimated to be about 735 Swedish Kroner (skr) per hectare irrigated per year, and variable costs are estimated to be about 4 skr per hectare-millimeter.

†Swedish water law stipulates that water may be withdrawn only in amounts that will not harm the public's right to water, and establishes certain general controls on quantities of water that may be withdrawn for specific purposes without special permission. For example, withdrawals from groundwater for irrigation purposes are limited to 300 m^3 per 24 hours without special permission.

TABLE 4 Estimated irrigation costs.

	Estimated cost
Investment	735 (skr per hectare per year)
Operating	4 (skr per mm per hectare)

See Appendix A for explanation of cost estimates.

The second type of economic parameter required for our calculations is the net farm price of the crops we consider. Net farm price, or producer's price, is equal to the wholesale price for the crop less the cost of harvest, drying, sorting, and transporting the crop to the market. Table 5 reports average wholesale prices, preparation and delivery costs, and net farm prices in 1978 for table potatoes and sugar beets. The net farm prices shown are the prices used in our calculations of irrigation demand in Section 4.

TABLE 5 Wholesale and net farm prices, and preparation and transport costs of table potatoes and sugar beets.

Crop	Wholesale price (skr/dt)	Preparation and transport cost (skr/dt)	Net farm price (skr/dt)
Table potatoes	90	10	80
Sugar beets	16	1	15

4 ESTIMATES OF THE DEMAND FOR IRRIGATION WATER IN SKÅNE

All of the data needed to make the calculations explained in Section 2 are now available. To estimate irrigation demand, we now only require to calculate the expressions developed in Section 2, and certain auxiliary expressions, using the parameter estimates given in Section 3.

4.1 Partial Expectations

Evaluation of the expressions in Section 2 requires that the expectations of certain random variables be taken over a subset of their range. Fortunately this can be done with relative ease given the form of the precipitation densities (Weibull) and water input–yield functions (quadratic) employed here.

Consider first the partial expectation of the first-order term in the seasonal water input–yield relationship,

$$a_1 B \int_0^{i^*/B} r f(r) \mathrm{d}r = a_1 B \int_0^{i^*/B} \gamma_0 \gamma_1 r^{\gamma_1} e^{-\gamma_0 r^{\gamma_1}}$$

Let $t = \gamma_0 r^{\gamma_1}$, then

$$a_1(1/\gamma_0)^{1/\gamma_1} \int_0^{\gamma_0 (i^*/B)^{\gamma_1}} t^{1/\gamma_1} e^{-t} dt$$

But the integral in this expression is simply the incomplete gamma function with parameters $1/\gamma_1 + 1$ and $\gamma_0(i^*/B)^{\gamma_1}$. Thus, the partial expectations of first-order terms may be evaluated as

$$a_1(1/\gamma_0)^{1/\gamma_1} \Gamma_I[1/\gamma_1 + 1, \quad \gamma_0(i^*/B)^{\gamma_1}]$$

where $\Gamma_I(\cdot)$ is the incomplete gamma function.

The partial expectations of second-order terms may be reduced to simple expressions involving incomplete gamma functions by analogous reasoning. The resulting expression for the value of the partial expectation of second-order terms is

$$-a_2(1/\gamma_0)^{2/\gamma_1} \Gamma_I[2/\gamma_1 + 1, \quad \gamma_0(i^*/B)^{\gamma_1}]$$

4.2 Estimates of Quantities of Water Demanded

Our basic estimates of quantity of water demanded per irrigated hectare are presented in Table 6. These results are computed using the parameter values reported in Section 3.

Column (1) of Table 6 reports the value of the optimal irrigation parameter i^*, corresponding to 1978 prices and costs, determined according to eqn. (4) in Section 2. The value taken on by this parameter represents the quantity of effective water that maximizes the net contribution of the water input to farm income. At the price and cost combinations used in calculation of the results in Table 6, the quantity of effective water input that maximizes this contribution is only slightly less than the quantity of effective water input that would maximize yield.

Column (2) of Table 6 reports the probability that seasonal precipitation will yield less than the optimal quantity of effective water shown in Column (1). Thus, for example, our estimates imply that in more than 99 percent of the years, irrigation water would have to be applied to both potatoes and sugar beets in order to bring the water inputs up to optimal levels. In less than one percent of the years will precipitation supply the full amount of the optimal water inputs.

Column (3) of Table 6 reports the increases in expected yields that would result from irrigation, computed according to the relationships derived in Section 2. As can be seen by comparing these figures with corresponding estimates of expected yield in the absence of irrigation presented in Table 3, the expected increases are substantial. In the case of both crops, increases in expected yields amount to more than 25 percent of expected yields without irrigation.

Column (4) of Table 6 reports the expected quantity of water demanded for irrigation. This is obtained by evaluating the partial expectation of $i^* - Br$ over the interval $(0, i^*/B)$. Our model implies that expected irrigation demand at 1978 price and cost levels would be about 87 mm/ha for both crops.

TABLE 6 Analysis of water demands per hectare irrigated.

Crop	Period	Production	(1) i^* (mm/ha)	(2) $F(i^*/B)$	(3) $\Delta\bar{Y}$ (dt/ha)	(4) $\bar{I}$ (mm/ha)	(5) $c\bar{I}$ (skr/ha)	(6) $\Delta\bar{\Pi}$ (skr/ha)
Potatoes†	16 June–31 Aug	$a_1 = 3.4826$ $a_2 = 0.0087$	197.28	0.9989	78.63	87.03	348.12	5,942.31
Sugar beets††	1 July–15 Sept	$a_1 = 4.1707$ $a_2 = 0.0091$	214.51	0.9971	102.19	86.65	346.60	1,186.24

Parameters used in calculation:
† Farm price = 80 skr/dt; rainfall effectiveness = 0.65
†† Farm price = 15 skr/dt; rainfall effectiveness = 0.75
Variable cost of irrigation = 4 skr/mm/ha

These estimates are reasonably close to those reported for the same crops in Arthur (1980). This is not really surprising as the irrigation rules simulated by Arthur were derived in part from the same experimental data that were used to fit the water input–yield relationships employed in our calculations. Nonetheless, it is important to note that when current (1978) crop prices and water costs are taken into account, per hectare water demands are approximately the same as those estimated in Arthur's analysis.

Column (6) of Table 6 reports the expected contribution of irrigation to farm income (gross of fixed costs) per hectare. For example, our estimate of the expected contribution of irrigation of potatoes to farm income is about 5,940 skr per hectare irrigated. Since estimated fixed costs of irrigation are less than 800 skr per hectare irrigated (see Table 4), the estimates presented in Table 6 suggest that irrigation of both sugar beets and potatoes would be profitable at 1978 crop prices and irrigation costs.

It should be noted that the contribution to farm income of any single crop need not exceed fixed costs for irrigation to be profitable. This is because crops are grown in rotation, and in different plots on the same farm at the same time. For irrigation to be profitable, it is sufficient if the contribution to farm income from irrigating the mixture of crops is large enough to cover the fixed costs of irrigation.

4.3 Sensitivity of Water Use to Water Cost and Crop Prices

The sensitivity of the basic results presented in Table 6 to changes in crop prices and variable irrigation costs is investigated in Tables 7a and 7b. Table 7a reports estimated water quantity demanded at various combinations of net farm price for potatoes and variable costs of irrigation. The cell of the table corresponding to a net farm price of 75 skr per dt and a variable irrigation cost of 5 skr per mm/ha approximates the assumed values for crop price and irrigation cost used in calculating the results in Table 6.

Table 7a shows two interesting and important patterns. First, at low variable costs of irrigation (i.e., 5 skr/mm/ha), we see that the quantity of water demanded does not respond very much to changes in crop prices. At a net farm price of 35 skr/dt and variable irrigation cost of 5 skr/mm/ha, the estimated per hectare demand for water is 81.7 mm. At a price of 95 skr/dt and variable irrigation cost of 5 skr/mm/ha, demand for water is only about 5 mm/ha greater.

Demand is somewhat more sensitive to crop prices at higher irrigation costs. For example, at a cost of 25 skr/mm/ha, quantity demanded increases from about 49.5 mm/ha at a net farm price of 35 skr/dt to 74.8 mm/ha at a net farm price of 95 skr/dt. The increase in quantity demanded at this level of irrigation cost is thus over 25 mm/ha.

The second important pattern reflected in Table 7a is that sensitivity of quantity demanded to cost is greater at low net farm prices than at high prices. This can clearly be seen by comparing the columns of the table.

Table 7b shows the same general patterns as does Table 7a. Indeed, in some cases in which low crop prices are combined with high irrigation costs, the optimum irrigation parameter i^* is zero [see eqn. (4) in Section 2].

In Tables 7a and 7b, a broken line separates the price–cost combinations which yield a contribution to farm income of less than 800 skr/ha from those which yield this amount or more. Our estimate of fixed costs of irrigation per hectare is approximately 800 skr.

TABLE 7a Analysis of water demands for irrigation of potatoes.

Variable cost of irrigation water† (skr/mm/ha)	Net farm price (skr/dt) 35		55		75		95	
	$\bar{I}$ (mm/ha)	$\Delta\bar{\Pi}$ (skr/ha)	$\bar{I}$ (mm/ha)	$\Delta\bar{\Pi}$ (skr/ha)	$\bar{I}$ (mm/ha)	$\Delta\bar{\Pi}$ (skr/ha)	$\bar{I}$ (mm/ha)	$\Delta\bar{\Pi}$ (skr/ha)
5	81.70	2325.57	84.68	3892.16	86.07	5462.73	86.88	7034.76
25	49.47	1016.54	63.92	2406.62	70.80	3894.19	74.81	5417.90
45	21.97	316.41	43.90	1330.73	55.75	2629.14	62.84	4041.66
65	5.96	58.85	26.27	634.85	41.36	1659.62	51.11	2902.85

†Note that 1 hectare-millimeter is equal to 10 m^3. Variable cost figures may be converted to skr/m^3 by dividing by 10.

TABLE 7b Analysis of water demands for irrigation of sugar beets.

Variable cost of irrigation water† (skr/mm/ha)	Net farm price (skr/dt) 5		10		15		20	
	$\bar{I}$ (mm/ha)	$\Delta\bar{\Pi}$ (skr/ha)	$\bar{I}$ (mm/ha)	$\Delta\bar{\Pi}$ (skr/ha)	$\bar{I}$ (mm/ha)	$\Delta\bar{\Pi}$ (skr/ha)	$\bar{I}$ (mm/ha)	$\Delta\bar{\Pi}$ (skr/ha)
5	47.52	150.41	73.91	603.54	83.00	1101.42	92.13	2640.64
25	0	0	2.83	7.80	19.75	127.59	56.06	1161.28
45	0	0	0	0	0.46	1.33	25.08	365.39
65	0	0	0	0	0	0	6.92	69.90

†Note that 1 hectare-millimeter is equal to 10 m^3. Variable cost figures may be converted to skr/m^3 by dividing by 10.

Price–cost combinations falling below the broken line fail to yield a contribution to farm income sufficient to cover fixed costs. Price–cost combinations falling above the line do yield a large enough contribution to cover fixed costs. Although it is not necessary for all crops to yield a contribution sufficient to cover irrigation costs in order for irrigation to be profitable, it is necessary that the crop mix yields sufficient income to offset this amount. This means that at least one of the crops irrigated must show a contribution to farm income in excess of fixed costs.

4.4 Remarks

The most important conclusion of the analysis presented here is that explicit consideration of crop prices and irrigation costs does not result in materially lower estimates of per hectare demand for irrigation than those obtained by Arthur. Indeed, the estimated demands per hectare obtained in Section 4.2 are about the same as those reported by Arthur.

This certainly underscores the gravity of the results reported by Strzepek (1981). Consideration of economic factors bearing on individual farmer's irrigation demand decisions does not alter the conclusion that potential water demand and water supply could become seriously out of balance.

The only possible reprieve is that the crop market effects of irrigation could moderate demand. This possibility is considered in Section 5.

5 EFFECT OF IRRIGATION ON CROP MARKETS

Sections 1 and 2 indicated that increased yields due to irrigation could affect the markets for crops. Since the demand estimates presented in Section 4 are based on the assumption of a given set of market conditions, the possibility must be considered that irrigation-induced changes in crop market conditions may result in changes in irrigation demands. Estimating the effect of increased yields in a conceptually correct manner requires a fairly detailed model of the agricultural sector. To support a complete analysis of the effects of irrigation on crop markets and of the feedback effects on irrigation demand, this model would have to explain both factor demands per hectare and the determination of land areas planted in different crops. The model developed in Section 3 explains only irrigation demands per hectare. Without a model that explains crop areas as well, the most that can be done is to estimate roughly the possible ranges of outcomes and to examine how outcomes in these ranges could affect the demand for irrigation water.

To analyze the possible effects of irrigation on crop markets, let us first consider the basic principles of price and quantity determination in national crop markets. To begin, we assume that the national crop market in question is completely open (i.e., there are no barriers to trade) and perfectly competitive.

Determination of output and price in a competitive open national crop market is shown in Figure 4. Curve D–D represents the domestic demand for the commodity. p_w represents the world price for the commodity, and s_0 represents the domestic supply of the commodity. Ignore the curves p_d and s_1 for the moment.

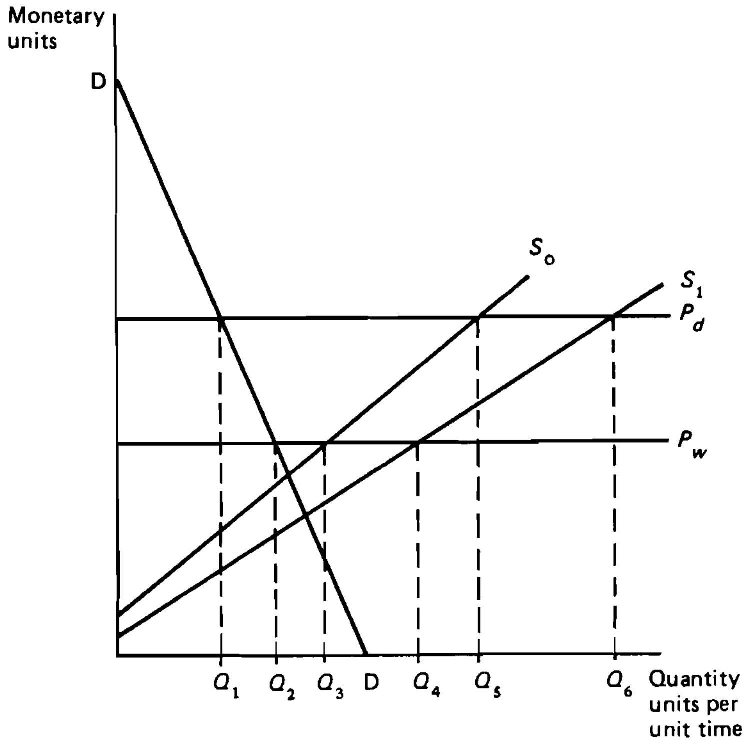


FIGURE 4 Price and output determination in national crop markets.

Under the open competitive market assumption, domestic producers may produce and sell as much output as they wish at the world price p_w. In this situation, domestic producers would produce a quantity corresponding to the point at which the world price curve p_w intersects the domestic supply curve s_0. Output also would be sold domestically at the world price p_w. Hence, domestic producers would produce Q_3 units of output per unit of time, and domestic consumers would purchase Q_2 units of output per unit of time. The difference $Q_3 - Q_2$ would be exported at the world price p_w. If the domestic supply and demand conditions relative to world market price were such that domestic consumers demanded more at the world price than domestic suppliers produced, then the balance would be imported at the world price.

However, the analysis would be somewhat different if domestic prices were supported at p_d, with foreign producers discouraged from selling in the domestic market by an import duty marginally larger than the difference between p_d and p_w. In this case domestic consumption would be Q_1 while domestic production would be Q_5. In this case, exports would not automatically close the gap between domestic production and consumption. The difference between Q_5 and Q_1 would have to be taken up through storage or perhaps through subsidized export. If the latter alternative were chosen, the cost of the subsidy required would be equal to the difference between the domestic and the world price, times the difference between the quantity of the commodity produced and the quantity of the commodity demanded in the domestic market.

Figure 4 also illustrates how an increase in supply -- such as might come about with the adoption of irrigation -- could affect the crop market. If the market were open and competitive, the effect of an increase in supply would be to increase domestic production and increase domestic exports (or reduce domestic imports). There would be no tendency for market price to change since the increment in output would be negligible in relation to the world market. This is seen in Figure 4 by examining the effect of a shift in supply from s_0 to s_1. Domestic output increases to Q_4, and exports increase to $Q_4 - Q_2$.

If the domestic market were insulated from the world market by tariffs and price supports, the effect of a supply shift would be to exacerbate the problems of dealing with crop surpluses. In the case depicted in Figure 4, the effect of the supply shift to s_1 is to increase by $Q_6 - Q_5$ the surplus of output over domestic consumption that must be dealt with in some fashion. The cost of dealing with this problem depends both upon the size of the surplus and the spread between domestic and world prices.

Market effects, such as those depicted in Figure 4, can have profound feedback effects on irrigation demand per hectare of crop and on the land area planted in the crop. Inspection of eqn. (4) reveals immediately that per hectare irrigation demand varies directly with crop price. It can also be shown that crop area varies directly with crop price (see Appendix B). The effects of crop price changes on per hectare irrigation demand and crop area reinforce one another. Estimates of the effects of price changes on per hectare demand thus tend to understate the full effect of crop price changes on demand.

It can be further shown that if crop price does not change as a result of irrigation, crop area would tend to increase (see Appendix B). Thus, although per hectare irrigation demand would be unaffected in the constant crop price case, total irrigation demand would increase.

Figure 4 conveys the essentials of the situation in Sweden. Sweden has taken policy measures to insulate its agricultural sector from the forces of world market competition. These measures include tariffs, subsidized export and storage programs, and price supports for Swedish farmers designed to insure that prices are adequate to cover costs and that "farmers . . . like others, . . . attain a reasonable income level" (Statens Lantbruks Information 1976).

Sweden today pursues what is termed a "high price" policy for its farm production. The domestic prices of most (although not all) crops are maintained above world levels. The specific levels at which domestic prices are maintained are set in negotiations between the government and various organizations representing producers of the different agricultural commodities. In 1978, the domestic price for table potatoes was about 90 skr per deciton, or about 20 skr per deciton above the world market price. However, the Swedish domestic price for sugar beets was roughly equal to the world market price in that year.

The analysis in Figure 4 demonstrates that, as long as the Swedish domestic price and the world price of sugar beets are approximately equal and the world market is competitive, the effect of increased irrigation of sugar beets is to increase production and, other things being equal, to increase exports. Hence, the sugar beet market conditions assumed in the calculation of per hectare irrigation demands of individual farmers (see Section 4) is not affected by irrigation. Crop market adjustments thus have no effect on demand for irrigation per hectare of sugar beets. Moreover, since irrigation increases the relative profitability of sugar beet production, land area planted in sugar beets can increase. This would result in an increase in *total* water demand, not per hectare water demand, for irrigation of sugar beets.

The effects of irrigation of table potatoes on the Swedish market are less clear-cut. Because of the spread between the Swedish domestic price and the world price of table potatoes, the nature and magnitude of effects on the potato crop market would depend upon the adjustment policy pursued. In this regard, it is useful to consider two limiting cases.

In the first case, we assume that the domestic price of table potatoes is lowered as far as is needed to absorb increased yields. This case could result in elimination of the divergence between domestic and world prices, and a decrease in irrigation water demand per hectare of potatoes.

In the second case, we assume that the 1978 spread between the domestic and the world price of table potatoes is maintained, and that the additional surplus is disposed of via subsidized exports. Of course, this would have the effect of increasing the cost of the price support policy, and, if land area planted in table potatoes increased, increasing total irrigation water demand. However, per hectare irrigation demand of table potatoes is unaffected.

5.1 Price Adjustment

Let us consider the case in which price adjustment alone accommodates the increased supply of table potatoes. Other things being equal, an increase in the yield of table potatoes would tend to cause the price to fall. This is because consumers are induced to buy the larger quantity of crop available only if the price falls. The extent of the price decrease necessary to bring increased crop supply into balance with demand, neglecting other (factor demand) adjustments, can be calculated approximately by using an estimate of the *elasticity of demand*.†

To see how the elasticity of demand can be used to estimate the effect of an increase in yield on crop price, suppose that the elasticity of demand for table potatoes is η and that – as a result of the increase in yield – the quantity of this crop offered for sale increases by n percent. Then the approximate change in the price of table potatoes that would be required to bring supply and demand back into balance would be

$$\Delta p = \frac{1}{\eta}\frac{\Delta q}{q}p = \frac{1}{\eta}np \tag{9}$$

†The elasticity of demand is defined as the percentage change in quantity demanded that would result from a given percentage change in price. The formula for this measure is

$$\text{elasticity} = \frac{\Delta q}{q} \bigg/ \frac{\Delta p}{p} = \frac{\Delta q}{\Delta p}\left(\frac{p}{q}\right)$$

where q is the quantity demanded and p is the price.

In a world in which agricultural commodities were freely and competitively traded, the demand facing producers in any country would be highly price elastic (i.e., η would be very large), and increments to output, of the size that would result from irrigation in Skåne, would result in imperceptible downward pressure on prices. Equation (9) is thus completely consistent with our analysis of Figure 4.

Rough estimates of the parameters needed to apply the formula in eqn. (9) and implied estimates of table potato price changes are shown in Table 8. These estimates assume

TABLE 8 Estimated effect of increased yields on table potato prices, assuming no producer adjustment.

(1) Approximate demand elasticity	0.1
(2) Approximate yield increase	0.29
(3) Approximate Skåne share	0.18
(4) Initial wholesale price (skr/dt)	90 (70)
(5) Wholesale price at increased yield (skr/dt)	70
(6) Incremental cost (skr/dt)	10
(7) Farm price at increased yield (skr/dt)	60

that there is no producer response (e.g., reduction in irrigation or reduction in land area planted in table potatoes) to price changes. Thus, they tend to represent upper bounds on the size of the price effect on per hectare irrigation demand that might be expected to follow from irrigation.

Line (1) of Table 8 gives estimates of the price elasticities of Swedish demand for table potatoes.† The elasticity is less than one, that is the domestic demand for this crop is inelastic.

Line (2) gives the estimated increase in per hectare yield obtained when the irrigation plan reflected in Table 6 is applied. Estimated fractional increases are obtained by dividing our estimates of the increases in yield by the estimated yields without irrigation reported in Table 3. Thus, our estimates imply that irrigation will increase expected yields of potatoes by about 29 percent (i.e., $78.70 \div 270.97 \times 100 = 29.01$).

Line (3) gives the fraction of total Swedish production of table potatoes accounted for by the part of Skåne -- Malmöhus County -- which is of interest here. This is the fraction of production which we assume to be affected by irrigation.†† Our estimate of the fraction by which total Swedish production of table potatoes is increased, is thus equal to the product of the fractions in lines (2) and (3) of Table 8.

Line (4) gives the assumed initial (i.e., without irrigation) Swedish wholesale price for table potatoes, and an approximate world market price in parentheses.†††

†Estimate provided by F. Desmond McCarthy, Food and Agriculture Program, IIASA.

††The Malmöhus County Board of Agriculture estimates that 100 percent of the crop area in potatoes in Malmöhus Country potentially will be irrigated.

†††World price was calculated by dividing the total dollar receipts for exports of potatoes by European nations, by the reported export quantities, and converting to skr. Data for these calculations are taken from Food and Agricultural Organization (1978).

As shown by comparison of these figures, the Swedish wholesale price for potatoes is about 20 skr above the world market price.

Line (5) gives the calculated Swedish wholesale price of table potatoes after the fractional production increases implied by the products of lines (2) and (3). This price is determined by calculating the price decrease implied by eqn. (15), and subtracting this calculated decrease from the initial price shown in line (4).

Line (6) gives the cost per deciton of processing table potatoes for wholesale. This is the difference between wholesale price and net farm price. Subtraction of line (6) from line (5) thus provides an estimate of the net farm price of table potatoes at the increased production level implied by the earlier lines of the table. This estimate is reported in line (7). On *a priori* grounds, the price effects calculated in Table 8 could have significant repercussions on the demand for irrigation water. In particular, significant decreases in the net farm price of potatoes of the magnitude reflected in our calculations in Table 8 could result in substantial reductions in demand for supplementary irrigation.

Juxtaposition of the results presented in Tables 7a and 8 provides some indication of the extent to which the price effects shown in Table 8 could moderate demand. The column of Table 7a which most nearly corresponds to our estimated farm price at increased yields is the column which assumes a net farm price of 55 skr/dt for table potatoes. According to our estimates, irrigation of potatoes would still be profitable by a wide margin (i.e., the contribution to farm income exceeds the fixed costs of irrigation) if irrigation costs remained constant, and the quantity of irrigation water demanded per hectare would be only slightly lower than it would be at 1978 prices.

5.2 Surplus Adjustment

Failing any adjustment of price to reflect the effects of supplementary irrigation on yield and production, additional surpluses of table potatoes would be produced. While the total production of sugar beets would also increase, as noted earlier this additional production could be exported without subsidy under the price and cost conditions assumed in Table 8.

Under our simplifying assumption that land areas planted do not adjust, if the domestic price of table potatoes were held constant, the increase in crop surplus that would have to be disposed of would be

$$\Delta s = A \, \Delta Y \tag{10}$$

where A is crop acreage irrigated, ΔY is the increase in expected yield associated with irrigation, and Δs is the increase in surplus. The cost of disposing of this extra surplus, assuming that it was exported at the world price, would be

$$\text{Export subsidy} = (p_d - p_w)\, \Delta s \tag{11}$$

where p_d and p_w are respectively the domestic and world prices.

When the calculation given in eqn. (10) is carried out we obtain an estimate that the incremental surplus production of table potatoes, assuming that all land planted in table potatoes in Malmöhus County in 1978 were irrigated, would be 316,250 decitons. At a price spread between domestic and world prices of 20 skr per deciton, the cost of disposing of this additional surplus would be 6,325,000 skr.†

5.3 Remarks

Two rather strong assumptions are implicit in the calculations of price adjustments using eqn. (9), and surplus adjustments using eqns. (10) and (11). As shown in Appendix B, these assumptions tend to bias downward our estimate of price that would result from irrigation, bias upward our estimate of the reduction in irrigation demand per hectare that could result from crop price effects, and bias downward our estimate of the increment to surplus that would result if crop prices were maintained.

The first assumption (in reality it is a simplification of our calculations) we have made is that we have not considered the effect of reduced irrigation on prices. That is, as prices fall, irrigation per hectare tends to fall, reducing yields, and relieving some of the downward pressure on prices. We could allow for this moderating effect by reformulating the model presented in Section 2 to take price effects into account. We have not done so because, in view of the other assumptions made, the gains from this refinement would not be worth the extra computation.

The second assumption we have made is that no other adjustments take place. In fact, as is demonstrated in Appendix B, one would expect to see adjustments in areas planted in crops in response to changes in the relative profitability of producing different crops. This would tend to moderate the price adjustments calculated in Section 5.1.

In the surplus adjustment case considered in Section 5.2, one would expect to see the acreage planted in irrigated crops increase since the profitability of these crops relative to other crops is increased, *ceteris paribus* (see Appendix B). The estimate of surplus we present in Section 5.2 thus tends to understate the additional surplus that would have to be disposed of in the pure surplus adjustment case.

These qualifications notwithstanding, it is important to note that one very important conclusion can be drawn from our approximate calculations: crop market effects alone probably would not result in a sufficiently large decrease in the demand for irrigation water in Skåne to rebalance water supply and potential demand in the region. Even the relatively large price decrease for potatoes reflected in Table 8 (which, for reasons discussed above, is larger than we would actually expect to observe) would not substantially reduce per hectare water demand by itself. Only if crop price had a substantial effect on crop acreage would there be any material effect on the total quantity of irrigation water demanded.

---

† As noted earlier, Swedish agricultural experts estimate that 100 percent of the potato crop in Malmöhus County will be irrigated. The increase in yield used in the above calculations is taken from Table 6, and the 1978 figure of 4,022 hectares of table potatoes is used as the estimate of irrigated crop area.

6 IRRIGATION DEMAND AT HIGHER IRRIGATION COSTS

The evidence examined suggests that water demands at roughly current prices and costs could well be as great as Arthur (1980) and Strzepek (1981) have estimated them to be, and that there is little prospect that irrigation-induced crop market effects would reduce irrigation water demands by very much.

In the calculations presented in Section 4, the cost of irrigation was estimated as the investment in equipment and outlays for operations to withdraw and apply water from surface water and groundwater sources. No cost was attached to the water *per se*. If adequate water supplies were available to meet all demands, the cost of irrigation estimated in this fashion would fairly represent the economic costs of irrigation. If, however, water were scarce relative to demands – as is potentially the case in Skåne if irrigation demands attain the levels projected in Section 4 – the true social cost of irrigation would be somewhat higher than simply the investment and operating costs needed to withdraw and apply it. In order to accommodate irrigation usage, water would have to be reallocated from other uses, or additional capacity would have to be added, and both of these options involve further costs.

Thus, the sole remaining options for rebalancing supply and demand are some reallocation of existing water supplies or the addition of supply capacity. Either option, as is indicated in Section 1, is likely to result in a substantial increase in the opportunity cost of irrigation.

Reallocation of existing water supplies would mean that some current users of water in the region would have to curtail their usage. As noted in Section 1, there may be no increase in water supply system costs occasioned by a reallocation of water from current users. However, there would be an opportunity cost associated with reallocated water. Current users pay on average about 5.5 skr/m^3 for municipal water in Skåne. This means that each cubic meter of water used may be viewed as being worth *at least* 5.5 skr to the user.

If this figure is taken as a benchmark of the opportunity cost of reallocating water from current users, then the implied variable cost of water per hectare irrigated is 55 skr/mm/ha, and the variable cost of irrigation is about 59 skr/mm/ha (i.e., 55 skr/mm/ha for water and 4 skr/mm/ha for operation of irrigation equipment). This is a substantial increase in irrigation costs over current levels (note that we estimate that the variable cost of irrigation is about 4 skr/mm/ha). Moreover, 55 skr/mm/ha may tend to underestimate the true opportunity cost of the reallocated water since it assumes that only the lowest valued uses (i.e., those valued at just 5.5 skr/m^3) are displaced and that no higher valued uses are displaced.

Inspection of the results presented in Tables 7a and 7b shows the dramatic effect that an increase in irrigation costs to 59 skr/mm/ha would have on per hectare irrigation demand. It is unlikely that irrigation of sugar beets, considered in isolation from other crops, would be profitable. Irrigation of table potatoes *might* continue to be profitable, depending upon what is assumed about crop prices. However, even if irrigation of table potatoes remained profitable, it is possible that per hectare irrigation quantities would be reduced between 30 and 50 percent from the levels we would expect at 1978 crop prices and irrigation costs (see Table 7a).

If charges were levied on use of water for irrigation at levels reflecting the opportunity cost of displaced uses, this suggests that the result could – depending upon what happens to crop prices – be a substantial decrease in quantities of water used per hectare for irrigation. The decrease in usage could be sufficient to rebalance demand and supply without additions to existing capacity.

Another option for rebalancing supply and demand is addition to current water supply system capacity. Indeed, system capacity is currently being expanded through construction of the Bolmen project. Taking estimated average cost of water from Bolmen as a measure of the cost of additional capacity, we may approximate the cost of water from expanded capacity between 4 and 5 skr/m^3, or just about the same amount as the cost of reallocating existing supplies. The conclusions stated earlier (i.e., that demands would be significantly reduced at higher cost levels) thus remain valid if the option considered for rebalancing supply and demand is capacity expansion. Note, however, that it is doubtful that any addition to system capacity would be required if agricultural users were charged a price for water that reflected either its opportunity cost or the cost of additional capacity, and if crop prices were established solely by market forces. It is, therefore, doubtful that demands for supplementary irrigation would provide an economic justification for capacity expansion at costs approximating those of the Bolmen project.

These conclusions are reinforced if the effect of increases in irrigation cost on crop areas is considered. In Appendix B it is demonstrated that the effect of an increase in irrigation cost is to reduce crop area. The crop area effect (which we are unable to estimate with the data at our disposal), taken in conjunction with the effects on per hectare demands estimated above, imply a very substantial reduction in total irrigation demand.

We remarked above that Swedish institutions governing the withdrawal and use of water for irrigation provide no mechanism for pricing of the water used. Thus, if prices were to be used in an effort to balance water supply and demand, substantial changes in the current institutional framework for water management might be required.

7 CONCLUSIONS

The following conclusions can be drawn, based upon our analysis.

First, an economic analysis of demand for irrigation water suggests that the per hectare demands for irrigation water for table potatoes and sugar beets, at 1978 crop prices and variable costs of irrigation, are about the same amount as Arthur (1980) estimated. Based on the analysis of Strzepek (1981), we thus conclude that water supply and demand in Skåne could become seriously out of balance.

Second, the feedback effects of increased yields and production on irrigation demand, even under relatively extreme assumptions probably would not reduce the demand for irrigation water enough (if at all) to rebalance demand and supply. This conclusion is obvious if crop prices do not fall – either because they are equal to world price levels or because they are prevented from doing so by crop price supports. Our calculations show that even if prices did fall, and even if market adjustment mechanisms that tend to moderate price decreases were ignored, the fall in prices probably would not reduce demand by more than a modest amount.

Third, if the variable costs of irrigation were increased to levels reflecting the opportunity cost of the water used (e.g., through levying a charge on irrigation water usage) the reduction in demand would be large enough to rebalance demand and existing supply, depending upon what happens to crop prices. If crop prices fell in response to increased yields, quantities of water used can fall between 30 to 50 percent or more from levels that are used under 1978 price and cost conditions. If prices were supported at well-above world price levels, the effect of increased costs on demand would be smaller.

The analysis on which these conclusions is based contains a number of simplifications and approximations. Even so, in our opinion the results of the analysis can hardly be classified as "simple" or "clear-cut." The actual extent to which supplementary irrigation will be practiced in Skåne clearly will depend upon a number of factors about which we can only speculate. We have seen, for example, that conclusions concerning demand quantities depend upon what one assumes about agricultural commodity price policy. If prices are allowed to adjust to reflect improved yields due to irrigation, water demand may be moderated slightly and the balancing of water supply and demand may be facilitated. If price supports are maintained and additional production is accumulated as surplus. water demand may be amplified as land is reallocated to relatively more profitable irrigated crops, and the potential water supply–demand imbalance may be exacerbated. If institutions are so structured as to confront the farmer with the full social opportunity cost of the water resources he uses, then cost considerations will moderate demands and help balance water supply and demand. If institutions are not so structured, then little help in balancing supply and demand can be expected from this quarter.

What is absolutely clear, from this work and that of Strzepek and Arthur, is that agricultural water demand in Skåne is an important – perhaps even *the critical* – factor in future planning and management of the water supply system in this region. A more detailed investigation which examines and evaluates the assumptions and approximations we have made and the course of Swedish policy with respect to agricultural commodities and water use is thus certainly required. This undoubtedly is our most important conclusion.

ACKNOWLEDGMENTS

During the preparation of this report I have been fortunate to have had the encouragement and constructive criticism of a great many people. It is a pleasure to thank Åke Andersson, Susan Arthur, Jesse Ausubel, Thomas Crocker, Ronald Cummings, Donald Erlenkotter, Janusz Kindler, Lennart de Maré, Desmond McCarthy, Edwin Mills, Kirit Parikh, Folke Snickars, Kenneth Strzepek, and Henry Vaux, Jr., for their helpful comments and suggestions on earlier drafts. However, they cannot be held responsible for any errors, ambiguities, or other faults that may remain.

REFERENCES

Anderson, R, Jr. (1980) The probability distribution of water inputs and the economic benefits of supplementary irrigation. WP-80-165. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Arthur, S. (1980) Irrigation in Skåne – estimated water needs and effect on water available to crops. WP-80-112. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Burt, O.R., and M.S. Stauber (1971) Economic analysis of irrigation in subhumid climate. American Journal of Agricultural Economics 53(1): 33–46.

Córdova, J.R., and R.L. Bras (1979) Stochastic control of irrigation systems. Cambridge, Massachusetts: Department of Civil Engineering, Massachusetts Institute of Technology.

Flinn, J.C., and W.F. Musgreave (1967) Development and analysis of input–output relations for irrigation water. Australian Journal of Agricultural Economics 11: 1–19.

Food and Agricultural Organization (1978) FAO Trade Yearbook. Volume 32.

Hagan, R.M., Y. Vaadia, and M.W. Russell (1959) Interpreting plant responses to soil moisture regimes, in Advances in Agronomy, Volume 11, edited by A.G. Norman. New York: Academic Press.

Hall, W.A., and W.S. Butcher (1968) Optimal timing of irrigation. Journal of the Irrigation and Drainage Division, American Society of Civil Engineers 94: 267–275.

Hallgren, G. (1947) Studies on the influence of precipitation on crop yields in Sweden with special reference to field irrigation. The Annals of the Royal Agricultural College of Sweden 14: 175–279.

Hanks, R.J. (1974) A model of predicting plant yield as influenced by water use. Agronomy Journal 66(5): 660.

Hexem, R.W., and E.O. Heady (1978) Water Production Functions for Irrigated Agriculture. Ames, Iowa: Iowa State University Press.

Hiler, E.A., and R.W. Clark (1971) Stress day index to characterize effects of water stress on crop yields. Transactions of the American Society of Agricultural Engineers 14: 757–761.

Jensen, M.E. (1968) Water Consumption by Agricultural Plants in Water Deficits and Plant Growth, Volume II, edited by T.T. Koglowski. New York: Academic Press.

Johansson, W., and H. Linner (1977) Bevattning. LTs förlag.

Minhas, B., K. Parikh, and T.N. Srinivasan (1974) Toward the structure of a production function for wheat yields with dated inputs of irrigation water. Water Resources Research 10(3): 383.

Moore, C.V. (1961) A general analytical framework for estimating the production function for crops using irrigation water. Journal of Farm Economics 43(4): 876–888.

Organization for Economic Cooperation and Development (1974) Agricultural Policy in Sweden.

Statens Lantbruks Information (1977) Lönar sig bevattning?

Statens Lantbruks Information (1976) Swedish Agriculture. p. 19.

Strzepek, K.M. (1981) MITSIM-2: A simulation model for planning and operational analysis of river basin systems. WP-81-000. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Swedish University of Agricultural Sciences (1966–79) (annual) Results of field experiments on drainage, land improvement, and irrigation.

Veihmeyer, F.J., and A.H. Hendrickson (1955) Does transpiration decrease as the soil moisture decreases? Transactions American Geophysical Union 36: 425–448.

Yaron, D. (1971) Estimation and use of water production functions in crops. Journal of the Irrigation and Planning Division of the American Society of Civil Engineers 97: 291–303.

APPENDIX A: ESTIMATED IRRIGATION COSTS

The following estimated irrigation costs were prepared by Lennart de Maré, IIASA. The calculations were based on 1978 costs, 30 hectares were irrigated. All costs are given in skr.

Investments:

	Surface Water (skr)	Groundwater (skr)
Well		30,000
Pump & electric installations	25,000	30,000
Main pipe		
1200 m	30,000	
700 m		17,500
Movable pipe	7,500	7,500
Hydrants	2,500	2,500
Irrigation machine	40,000	40,000
	105,000	127,500

Amortization of investment:

	Well & main pipe (%)	Pump & electric installation, movable pipe (%)	Irrigation machine (%)
Depreciation	5	11	14
Interest	4	4	4
Maintenance	1	5	6
	10	20	24

Other fixed costs:

	skr/yr
Storage, insurance, etc., at irrigation machine	1,500
Electricity (basic fee)	1,200
	2,700

Variable costs and cost factors:

	Factor	Units
Electricity use charge	0.13	Skr/kwhr
Labor cost	35	skr/hr
Tractor cost	17	skr/hr
Electricity use per hour	20.25	kwh/hr
Irrigation factor		
Potatoes	25	mm/ha/irrigation
	2	ha/18 hour-day
Grain sugar beets	35	mm/ha/irrigation
	1.5	ha/18 hour-day
Labor use	1	hr/ha/irrigation
	1.8	hr/irrigation-day
Tractor use	0.3	hr/ha/irrigation
	1	hr/irrigation-day

Irrigation assumptions:

15 ha sugar beets, 2 irrigations, 35 mm/irrigation
15 ha potatoes, 5 irrigations, 25 mm/irrigation

Calculations:

Fixed costs per year:

Surface water

Investment: 0.1 X 32,500 + 0.2 X 32,500 + 0.24 X 40,000 =	19,350
Other fixed:	2,700
	22,050

Groundwater

Investment: 0.1 X 50,000 + 0.2 X 37,500 + 0.24 X 40,000 =	22,100
Other fixed:	2,700
	24,800

Fixed costs per year per hectare:

Surface water: 735
Groundwater: 827

Variable costs per year:

Potato irrigation days:	15/2 X 5 = 37.5	
Sugar beet irrigation days:	15/1.5 X 2 = 20.0	
Total irrigation days/yr	57.5	
	X 18	
Total irrigation hr/yr	1,035	
Electric cost @ 1,035 X 20.25 X 0.13		2,724.64
Labor cost		
Potatoes. 5 X 15 X 35 X 1.0		2,625.00
1.8 X 37.5 X 35		2,362.50
Sugar beets: 2 X 15 X 35 X 1.0		1,050.00
1.8 X 20 X 35		1,260.00
Tractor cost		
Potatoes: 5 X 15 X 17 X 0.3		382.50
1.0 X 37.5 X 17		637.50
Sugar beets: 2 X 15 X 17 X 0.3		153.00
1.0 X 20 X 17		340.00
		11,535.14

Variable costs per hectare-millimeter/yr:

Hectare-millimeters:

Potatoes:	5 X 15 X 25 = 1,875
Sugar beets:	2 X 15 X 35 = 1,050
	2,925

Variable cost per hectare-millimeter: 11,535.14 ÷ 2,925 = 3.94

APPENDIX B: ANALYSIS OF THE SENSITIVITY OF IRRIGATION DEMAND PER HECTARE AND CROP AREA TO SELECTED PARAMETERS

In Sections 5 and 6 of this report, approximate computations are made of effects of changes in irrigation costs and crop market conditions on the demand for irrigation. Our formal model of irrigation demand (see Section 3) characterizes only the determination of irrigation demand per hectare of crop. Total irrigation demand is the product of demand per unit crop area and total crop area. This appendix qualitatively analyzes the sensitivity of total irrigation demand for a crop to key parameters (irrigation cost and crop price) and shows the effects of changes in these parameters on total demand and per hectare demand.

B.1 Notations and Assumptions

For this purpose, let us denote the production function for a crop by

$$Q = AF(A,w)$$

where Q is total crop output, A is crop area, and F is a function which relates crop yield per unit crop area to total crop area and irrigation water usage per unit crop area, denoted by w. In addition, we assume that the yield function F possesses the following properties:

$$\frac{\partial Q}{\partial A} = AF_1 + F > 0 \qquad \frac{\partial Q}{\partial W} = AF_2 > 0$$

$$\frac{\partial F}{\partial A} = F_1 < 0 \qquad \frac{\partial^2 Q}{\partial W^2} = AF_{22} < 0$$

$$F_{11} > 0 \qquad \frac{\partial^2 Q}{\partial A \partial W} = AF_{12} + F_2 = F_2 = \frac{\partial F}{\partial w} > 0$$

$$\frac{\partial^2 Q}{\partial A^2} = AF_{11} + 2F_1 < 0$$

We have assumed that output is a concave function of inputs, that yield per unit land area declines with increasing land area (as less and less suitable land is brought into production), and that the yield function F is separable in land area and water per unit land area.

We assume that farmers select input levels to maximize farm income. The necessary conditions for this to be achieved are

$$\begin{aligned} &p(AF_1 + F) - r - cw = 0 \\ &pAF_2 - cA = 0 \end{aligned} \tag{B1}$$

where r is the rent per unit area of land, and c is the irrigation cost per unit of water.

B.2 Sensitivity of Factor Demands to Irrigation Cost

Section 6 examines what would happen to irrigation demand per hectare w, if the cost of irrigation increased markedly. Using the extended model described in this appendix, we can also examine the effect of changes in irrigation costs on crop area A and on total water demand, $W = wA$.

To do this, let us first take the total differential of the first-order conditions (B1) we obtain

$$\begin{aligned} &\mathrm{d}p(AF_1 + F) + p(AF_{11} + 2F_1)\mathrm{d}A + (pF_2 - c)\mathrm{d}w - \mathrm{d}r - w\mathrm{d}c = 0 \\ &\mathrm{d}pAF_2 + (pF_2 - c)\mathrm{d}A + pAF_{22}\mathrm{d}w - A\mathrm{d}c = 0 \end{aligned} \tag{B2}$$

Setting $dp = dr = 0$, and solving for dw/dc and dA/dc, we obtain

$$\frac{dw}{dc} = \frac{A}{pAF_{22}} < 0$$

$$\frac{dA}{dc} = \frac{w}{p(AF_{11} + 2F_1)} < 0$$

We conclude that the effect of a change in the cost of irrigation per unit of water applied c is to reduce the quantity of water used per unit of land area w and to reduce the crop area A.

The change in total water usage is given by

$$\frac{d}{dc}(wA) = w\frac{dA}{dc} + A\frac{dw}{dc}$$

and the elasticity of total water demand e with respect to c is given by

$$e_c^T = \frac{c}{A}\frac{dA}{dc} + \frac{c}{w}\frac{dw}{dc} = e_c^A + e_c^w$$

We conclude that if irrigation cost c increases, total demand for irrigation falls on two accounts. First, the quantity of water demanded per unit of crop area w falls. Second, crop area A falls. Estimates of the effect of irrigation costs on demand that encompass only effects on irrigation demand per unit land area (as is the case with the estimates presented in Section 6) thus tend to understate the effect of cost on demand.

B.3 Sensitivity of Crop Area to Introduction of Irrigation

We have noted several times that the adoption of irrigation raises the profitability of irrigated crops. Other things being equal, this should result in an increase in crop area. To investigate, let us compare the solutions to the first-order conditions (B1) when w is constrained to be 0 (i.e., the no-irrigation case) and when it takes on the optimal value given by (B1).

Subtracting the first equation of (B1) evaluated at $w = 0$ from the first equation with w set at its optimal value, we obtain

$$\Delta = pAF_1(A,w) - pA^1F_1(A^1,0) + p[F(A,w) - F(A^1,0)] - cw = 0$$

By the concavity of F in w

$$F(A^1,0) < F(A^1,w) - F_2(A^1,w)w$$

and by separability

$$F_2(A^1,w) = F_2(A,w)$$

Substituting into the expression for Δ and making use of the second equation of (B1) to eliminate a term

$$\Delta \geqslant pAF_1(A,w) - pA^1F_1(A^1,0) + p[F(A,w) - F(A^1,w)] \leqslant 0$$

By the convexity of F in A

$$F(A,w) > F(A^1,w) + F_1(A^1,w)(A - A^1)$$

which implies

$$\Delta \geqslant pAF_1(A,w) - pA^1F_1(A^1,0) - pF_1(A^1,w)(A - A^1) = p[F_1(A,w) - F_1(A^1,0)]A \leqslant 0$$

which recalling the separability of F implies

$$A \geqslant A^1$$

Thus, we conclude that the effect of introduction of irrigation is to increase crop area.

This is an important finding. It could mean, for example, that the crop areas assumed in Strzepek's simulations are smaller than might be expected with full adoption of irrigation.

B.4 Sensitivity of Irrigation Demand to Crop Price

Section 5 examines the effect of crop price changes on irrigation demand per unit of crop area. Consideration of the total differential equations (B2) above shows what happens to both demand per unit land area and to crop area. In particular, (B2) implies that

$$\frac{dw}{dp} = AF_2/(pAF_{22}) > 0$$

$$\frac{dA}{dp} = AF_1 + F/(pAF_{11} + 2F_1) > 0$$

and

$$\frac{dW}{dp} = A\frac{dw}{dp} + w\frac{dA}{dp}$$

and

$$e_p^T = \frac{p}{w}\frac{dw}{dp} + \frac{p}{A}\frac{dA}{dp} = e_p^w + e_p^A$$

That is, irrigation demand per unit crop area, and crop area vary directly with crop price.

MODEL MIGRATION SCHEDULES

Andrei Rogers and Luis J. Castro
International Institute for Applied Systems Analysis, Laxenburg, Austria

SUMMARY

This report draws on the fundamental regularity exhibited by age profiles of migration all over the world to develop a system of hypothetical model schedules that can be used in multiregional population analyses carried out in countries that lack adequate migration data.

---

1 INTRODUCTION

Most human populations experience rates of age-specific fertility and mortality that exhibit remarkably persistent regularities. Consequently, demographers have found it possible to summarize and codify such regularities by means of mathematical expressions called model schedules. Although the development of model fertility and mortality schedules has received considerable attention in demographic studies, the construction of model migration schedules has not, even though the techniques that have been successfully applied to treat the former can be readily extended to deal with the latter.

We begin this report with an examination of regularities in age profile exhibited by empirical schedules of migration rates and go on to adopt the notion of model migration schedules to express these regularities in mathematical form. We then use model schedules to examine patterns of variation present in a large data bank of such schedules. Drawing on this comparative analysis of "observed" model schedules, we develop several "families" of schedules and conclude by indicating how they might be used to generate hypothetical "estimated" schedules for use in Third World migration studies -- settings where the available migration data are often inadequate or inaccurate.

2 AGE PATTERNS OF MIGRATION

Migration measurement can usefully apply concepts borrowed from both mortality and fertility analysis, modifying them where necessary to take into account aspects that are peculiar to spatial mobility. From mortality analysis, migration studies can borrow the notion of the life table, extending it to include increments as well as decrements, in order to reflect the mutual interaction of several regional cohorts (Rogers 1973a, b, 1975, Rogers and Ledent 1976). From fertility analysis, migration studies can borrow well-developed techniques for graduating age-specific schedules (Rogers et al. 1978). Fundamental to both "borrowings" is a workable definition of the migration rate.

2.1 Migration Rates and Migration Schedules

The simplest and most common measure of migration is the crude migration rate, defined as the ratio of the *number of migrants*, leaving a particular population located in space and time, to the average *number of persons* (more exactly, the number of person-years) exposed to the risk of becoming migrants. Data on nonsurviving migrants are often unavailable, therefore the numerator in this ratio generally excludes them.

Because migration is highly age selective, with a large fraction of migrants being young, our understanding of migration patterns and dynamics is aided by computing migration rates for each single year of age. Summing these rates over all ages of life gives the *gross migraproduction rate* (*GMR*), the migration analog of fertility's gross reproduction rate. This rate reflects the level at which migration occurs out of a given region.

The age-specific migration schedules of multiregional populations exhibit remarkably persistent regularities. For example, when comparing the age-specific annual rates of residential migration among whites and blacks in the United States during 1966–1971, one finds a common profile (Figure 1). Migration rates among infants and young children mirrored the relatively high rates of their parents, young adults in their late twenties. The mobility of adolescents was lower but exceeded that of young teens, with the latter showing a local low point around age 15. Thereafter migration rates increased, attaining a high peak at about age 22 and then declining monotonically with age to the ages of retirement. The migration *levels* of both whites and blacks were roughly similar, with whites showing a *GMR* of about 14 migrations and blacks one of approximately 15 over a lifetime undisturbed by mortality before the end of the mobile ages.

Although it has frequently been asserted that migration is strongly sex selective, with males being more mobile than females, recent research indicates that sex selectivity is much less pronounced than age selectivity and is less uniform across time and space. Nevertheless, because most models and studies of population dynamics distinguish between the sexes, most migration measures do also.

Figure 2 illustrates the age profiles of male and female migration schedules in four different countries at about the same point in time between roughly comparable areal units: communes in the Netherlands and Sweden, voivodships in Poland, and counties in the United States. The migration levels for all but Poland are similar, varying between 3.5 and 5.3 migrations per lifetime; and the levels for males and females are roughly the same. The age profiles, however, show a distinct, and consistent, difference. The high peak of the female schedule precedes that of the male schedule by an amount that appears to approximate the difference between the average ages at marriage of the two sexes.

Under normal statistical conditions, point-to-point movements are aggregated into streams between one civil division and another; consequently, the level of interregional migration depends on the size of the areal unit selected. Thus if the areal unit chosen is a

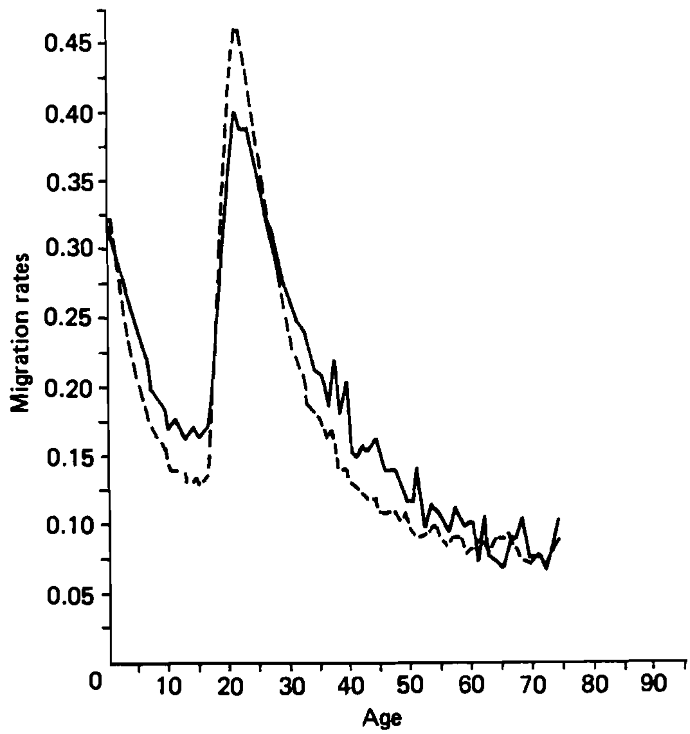


FIGURE 1 Observed annual migration rates by color (– – – white, —— black) and single years of age: the United States, 1966–1971.

minor civil division such as a county or a commune, a greater proportion of residential location will be included as migration than if the areal unit chosen is a major civil division such as a state or a province.

Figure 3 presents the age profiles of female migration schedules as measured by different sizes of areal units: (1) all migrations from one residence to another, (2) changes of residence within county boundaries, (3) migration between counties, and (4) migration between states. The respective four *GMR*s are 14.3, 9.3, 5.0, and 2.5. The four age profiles appear to be remarkably similar, indicating that the regularity in age pattern persists across areal delineations of different size.

Finally, migration occurs over time as well as across space; therefore, studies of its patterns must trace its occurrence with respect to a time interval, as well as over a system of geographical areas. In general, the longer the time interval, the larger the number of return movers and nonsurviving migrants and, hence, the more the count of *migrants* will understate the number of interarea *movers* (and, of course, also of moves). Philip Rees, for example, after examining the ratios of one-year to five-year migrants between the Standard Regions of Great Britain, found that

> . . . the number of migrants recorded over five years in an interregional flow varies from four times to two times the number of migrants recorded over one year. (Rees 1977, p. 247)

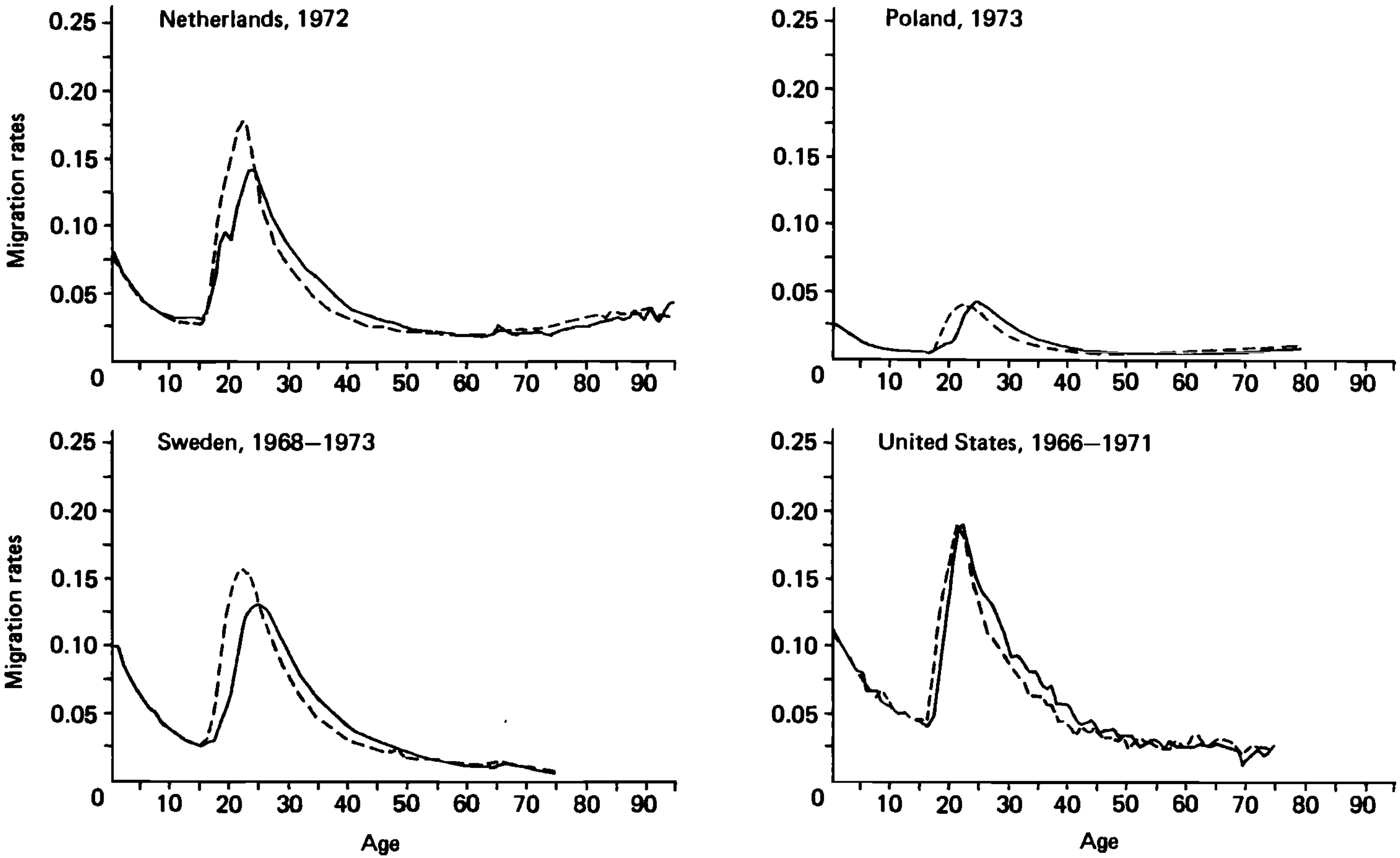


FIGURE 2 Observed annual migration rates by sex (– – – females, —— males) and single years of age: the Netherlands (intercommunal), Poland (intervoivodship), Sweden (intercommunal), and the United States (intercounty); around 1970.

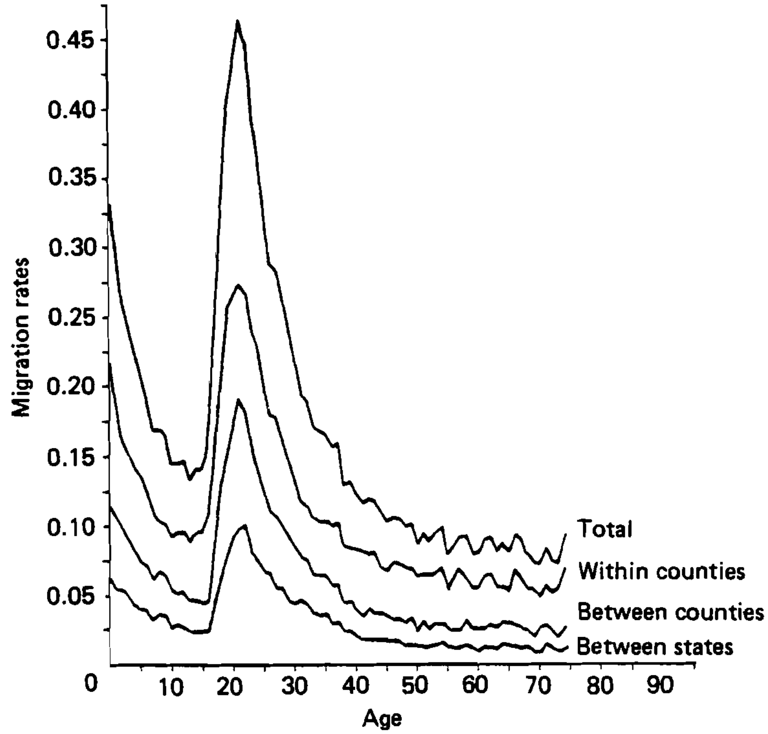


FIGURE 3 Observed average annual migration rates of females by levels of areal aggregation and single years of age: the United States, 1966–1971.

2.2 Model Migration Schedules

From the preceding section it appears that the most prominent regularity found in empirical schedules of age-specific migration rates is the selectivity of migration with respect to age. Young adults in their early twenties generally show the highest migration rates and young teenagers the lowest. The migration rates of children mirror those of their parents; hence the migration rates of infants exceed those of adolescents. Finally, migration streams directed toward regions with warmer climates and into or out of large cities with relatively high levels of social services and cultural amenities often exhibit a "retirement peak" at ages in the mid-sixties or beyond.

Figure 4 illustrates a typical *observed* age-specific migration schedule (the jagged outline) and its graduation by a *model* schedule (the superimposed smooth outline) defined as the sum of four components:

1. A single negative exponential curve of the *pre-labor force* ages, with its rate of descent α_1
2. A left-skewed unimodal curve of the *labor force* ages positioned at mean age μ_2 on the age axis and exhibiting rates of ascent λ_2 and descent α_2

α_1 = rate of descent of pre-labor force component
λ_2 = rate of ascent of labor force component
α_2 = rate of descent of labor force component
λ_3 = rate of ascent of post-labor force component
α_3 = rate of descent of post-labor force component
c = constant

x_l = low point
x_h = high peak
x_r = retirement peak
X = labor force shift
A = parental shift
B = jump

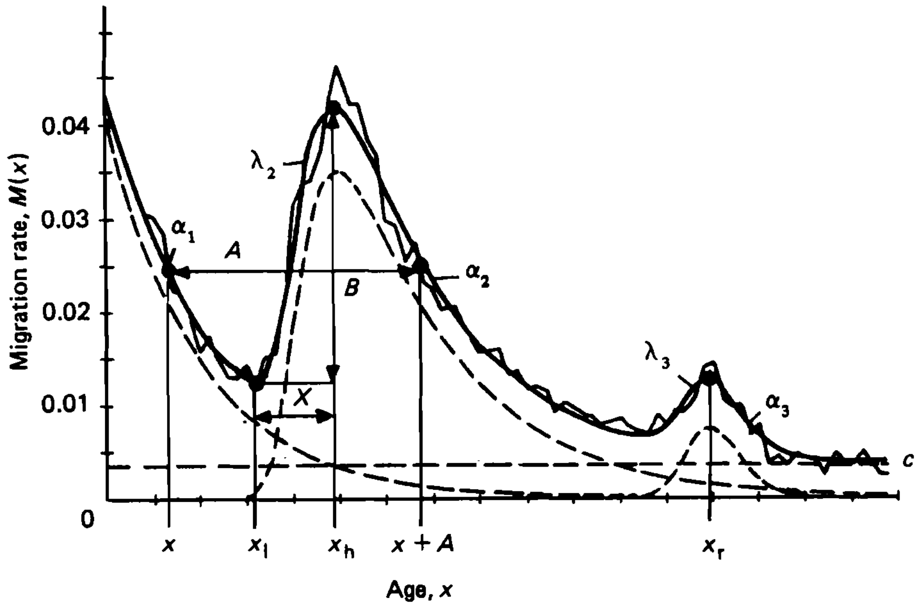


FIGURE 4 The model migration schedule.

3. An almost bell-shaped curve of the *post-labor force* ages positioned at μ_3 on the age axis and exhibiting rates of ascent λ_3 and descent α_3
4. A constant curve c, the inclusion of which improves the fit of the mathematical expression to the observed schedule

The decomposition described above suggests the following simple sum of four curves (Rogers et al. 1978):

$$\left.\begin{aligned} M(x) = {} & a_1 \exp(-\alpha_1 x) \\ & + a_2 \exp\{-\alpha_2(x-\mu_2) - \exp[-\lambda_2(x-\mu_2)]\} \\ & + a_3 \exp\{-\alpha_3(x-\mu_3) - \exp[-\lambda_3(x-\mu_3)]\} \\ & + c \end{aligned}\right\} \quad x = 0,1,2,\ldots,z \qquad (1)$$

The labor force and the post-labor force components in eq. (1) adopt the "double exponential" curve formulated by Coale and McNeil (1972) for their studies of nuptiality and fertility.

The "full" model schedule in eq. (1) has 11 parameters: $a_1, \alpha_1, a_2, \mu_2, \alpha_2, \lambda_2, a_3, \mu_3, \alpha_3, \lambda_3$, and c. The *profile* of the full model schedule is defined by 7 of the 11 parameters: $\alpha_1, \mu_2, \alpha_2, \lambda_2, \mu_3, \alpha_3$, and λ_3. Its *level* is determined by the remaining 4 parameters: a_1, a_2, a_3, and c. A change in the value of the *GMR* of a particular model schedule alters proportionally the values of the latter but does not affect the former. As we shall see in the next section, however, certain aspects of the profile also depend on the allocation of the schedule's level among the pre-labor, labor, and post-labor force age components and on the share of the total level accounted for by the constant term c. Finally, migration schedules without a retirement peak may be represented by a "reduced" model with seven parameters, because in such instances the third component of eq. (1) is omitted.

Table 1 sets out illustrative values of the basic and derived measures presented in Figure 4. The 1974 data refer to migration schedules for an eight-region disaggregation of Sweden (Andersson and Holmberg 1980). The method chosen for fitting the model schedule to the data is a functional-minimization procedure known as the modified Levenberg–Marquardt algorithm (see Appendix A, Brown and Dennis 1972, Levenberg 1944, Marquardt 1963). Minimum chi-square estimators are used to give more weight to age groups with smaller rates of migration.

To assess the goodness-of-fit that the model schedule provides when it is applied to observed data, we calculate E, the mean of the absolute differences between estimated and observed values expressed as a percentage of the observed mean:

$$E = \frac{(1/n) \sum_x |\hat{M}(x) - M(x)|}{(1/n) \sum_x M(x)} 100 \tag{2}$$

This measure indicates that the fit of the model to the Swedish data is reasonably good, the eight regional indices of goodness-of-fit E being 6.87, 6.41, 12.15, 11.01, 9.31, 10.77, 11.74, and 14.82 for males and 7.30, 7.23, 10.71, 8.78, 9.31, 11.61, 11.38, and 13.28 for females. Figure 5 illustrates graphically this goodness-of-fit of the model schedule to the observed regional migration data for Swedish females.

Model migration schedules of the form specified in eq. (1) may be classified into *families* according to the ranges of values taken on by their principal parameters. For example, we may order schedules according to their migration levels as defined by the values of the four level parameters in eq. (1), i.e., a_1, a_2, a_3, and c (or by their associated *GMR*s). Alternatively, we may distinguish schedules with a retirement peak from those without one, or we may refer to schedules with relatively low or high values for the rate of ascent of the labor force curve λ_2 or the mean age $\bar{n}$. In many applications, it is also meaningful to characterize migration schedules in terms of several of the fundamental measures illustrated in Figure 4, such as the low point x_l, the high peak x_h, and the retirement peak x_r. Associated with the first pair of points is the labor force shift X, which is defined to be the difference in years between the ages of the high peak and the low point, i.e., $X = x_h - x_l$. The increase in the migration rate of individuals aged x_h over those aged x_l will be called the jump B.

TABLE 1 Parameters and variables defining observed model migration schedules: outmigration from the 8

Parameters and variables[a]	Region: 1. Stockholm		2. East Middle		3. South Middle		4. South	
	Male	Female	Male	Female	Male	Female	Male	Female
GMR[b]	1.45	1.43	1.44	1.48	1.33	1.41	0.87	0.84
a_1	0.033	0.041	0.035	0.039	0.032	0.033	0.025	0.021
α_1	0.097	0.091	0.088	0.108	0.096	0.106	0.117	0.104
a_2	0.059	0.067	0.079	0.096	0.091	0.112	0.066	0.067
μ_2	20.80	19.32	20.27	18.52	19.92	18.49	21.17	19.88
α_2	0.077	0.094	0.090	0.109	0.104	0.127	0.115	0.129
λ_2	0.374	0.369	0.406	0.491	0.404	0.560	0.269	0.442
a_3	0.000	0.000						
μ_3	76.55	85.01						
α_3	0.776	0.369						
λ_3	0.145	0.072						
c	0.003	0.003	0.003	0.004	0.003	0.004	0.002	0.002
$\bar{n}$	31.02	29.54	29.17	28.38	28.29	27.96	28.26	28.14
%(0–14)	25.61	25.95	22.81	22.59	21.40	20.67	22.76	21.93
%(15–64)	64.49	65.10	70.38	69.48	72.47	71.73	70.73	70.76
%(65+)	9.90	8.94	6.81	7.94	6.13	7.60	6.51	7.31
δ_{1c}	13.56	13.06	12.14	9.79	12.26	8.90	13.27	9.93
δ_{12}	0.716	0.604	0.446	0.403	0.350	0.293	0.377	0.312
δ_{32}	0.003	0.003						
β_{12}	1.26	0.977	0.981	0.993	0.921	0.883	1.02	0.809
σ_2	4.86	3.94	4.52	4.49	3.88	4.40	2.34	3.43
σ_3	0.187	0.196						
x_l	16.39	14.81	15.92	14.80	15.41	15.07	14.52	15.61
x_h	24.68	22.70	23.78	21.46	23.12	21.06	24.16	22.58
x_r	64.80	61.47						
X	8.29	7.89	7.86	6.66	7.71	5.99	9.64	6.97
A	27.87	25.49	29.99	27.32	29.93	27.27	29.90	27.87
B	0.029	0.030	0.040	0.022	0.044	0.059	0.026	0.032

[a] All parameters and variables are briefly defined in Appendix B and discussed more comprehensively in the
[b] The *GMR*, its percentage distribution across the three major age categories (i.e., 0–14, 15–64, 65+), and

The close correspondence between the migration rates of children and those of their parents suggests another important shift in observed migration schedules. If, for each point x on the post-high-peak part of the migration curve, we obtain by interpolation the age (where it exists), $x - A_x$ say, with the identical rate of migration on the pre-low-point part of the migration curve, then the average of the values of A_x, calculated incrementally for the number of years between zero and the low point x_l, will be defined as the observed parental shift A.

An observed (or a graduated) age-specific migration schedule may be described in a number of useful ways. For example, references may be made to the heights at particular ages, to locations of important peaks or troughs, to slopes along the schedule's age profile, to ratios between particular heights or slopes, to areas under parts of the curve, and to both horizontal and vertical distances between important heights and locations. The various descriptive measures characterizing an age-specific model migration schedule may be conveniently grouped into the following categories and subcategories:

Swedish regions, 1974 observed data by single years of age.

5. West		6. North Middle		7. Lower North		8. Upper North	
Male	Female	Male	Female	Male	Female	Male	Female
0.80	0.82	1.22	1.33	1.33	1.46	1.03	1.24
0.021	0.022	0.031	0.027	0.034	0.031	0.024	0.023
0.090	0.106	0.104	0.102	0.123	0.119	0.135	0.128
0.046	0.055	0.084	0.116	0.109	0.141	0.079	0.116
20.36	19.36	19.75	18.18	19.62	17.93	19.47	17.62
0.091	0.114	0.103	0.139	0.118	0.148	0.114	0.143
0.416	0.442	0.437	0.561	0.427	0.701	0.449	0.711
0.001	0.002	0.002	0.004	0.003	0.004	0.003	0.004
28.49	28.39	28.09	28.17	28.24	27.93	29.91	28.99
23.54	23.18	21.52	19.40	19.84	18.26	18.29	16.40
70.34	69.03	72.51	72.45	73.61	73.65	73.46	74.56
6.12	7.79	5.97	8.15	6.55	8.09	8.25	9.04
14.42	10.11	13.34	7.27	11.38	7.41	8.29	5.84
0.457	0.395	0.369	0.237	0.310	0.219	0.305	0.198
0.979	0.926	1.00	0.730	1.04	0.801	1.19	0.890
4.55	3.87	4.23	4.03	3.63	4.74	3.95	4.95
16.11	15.23	15.56	14.71	15.19	15.07	15.21	14.77
23.80	22.30	22.93	20.60	22.56	20.12	22.47	19.85
7.69	7.07	7.37	5.89	7.37	5.05	7.26	5.08
29.57	27.42	29.92	27.01	30.15	26.94	31.61	28.30
0.023	0.027	0.042	0.059	0.053	0.077	0.040	0.063

following text.
the mean age $\bar{n}$ are all calculated with a model schedule spanning an age range of 95 years.

1. Basic measures (the 11 fundamental parameters and their ratios)
   heights: a_1, a_2, a_3, c
   locations: μ_2, μ_3
   slopes: $\alpha_1, \alpha_2, \lambda_2, \alpha_3, \lambda_3$
   ratios: $\delta_{1c} = a_1/c$, $\delta_{12} = a_1/a_2$, $\delta_{32} = a_3/a_2$, $\beta_{12} = \alpha_1/\alpha_2$, $\sigma_2 = \lambda_2/\alpha_2$, $\sigma_3 = \lambda_3/\alpha_3$

2. Derived measures (properties of the model schedule)
   areas: GMR, %(0–14), %(15–64), %(65+)
   locations: $\bar{n}, x_l, x_h, x_r$
   distances: X, A, B

A convenient approach for characterizing an observed model migration schedule (i.e., an empirical schedule graduated by eq. (1)) is to begin with the central labor force curve

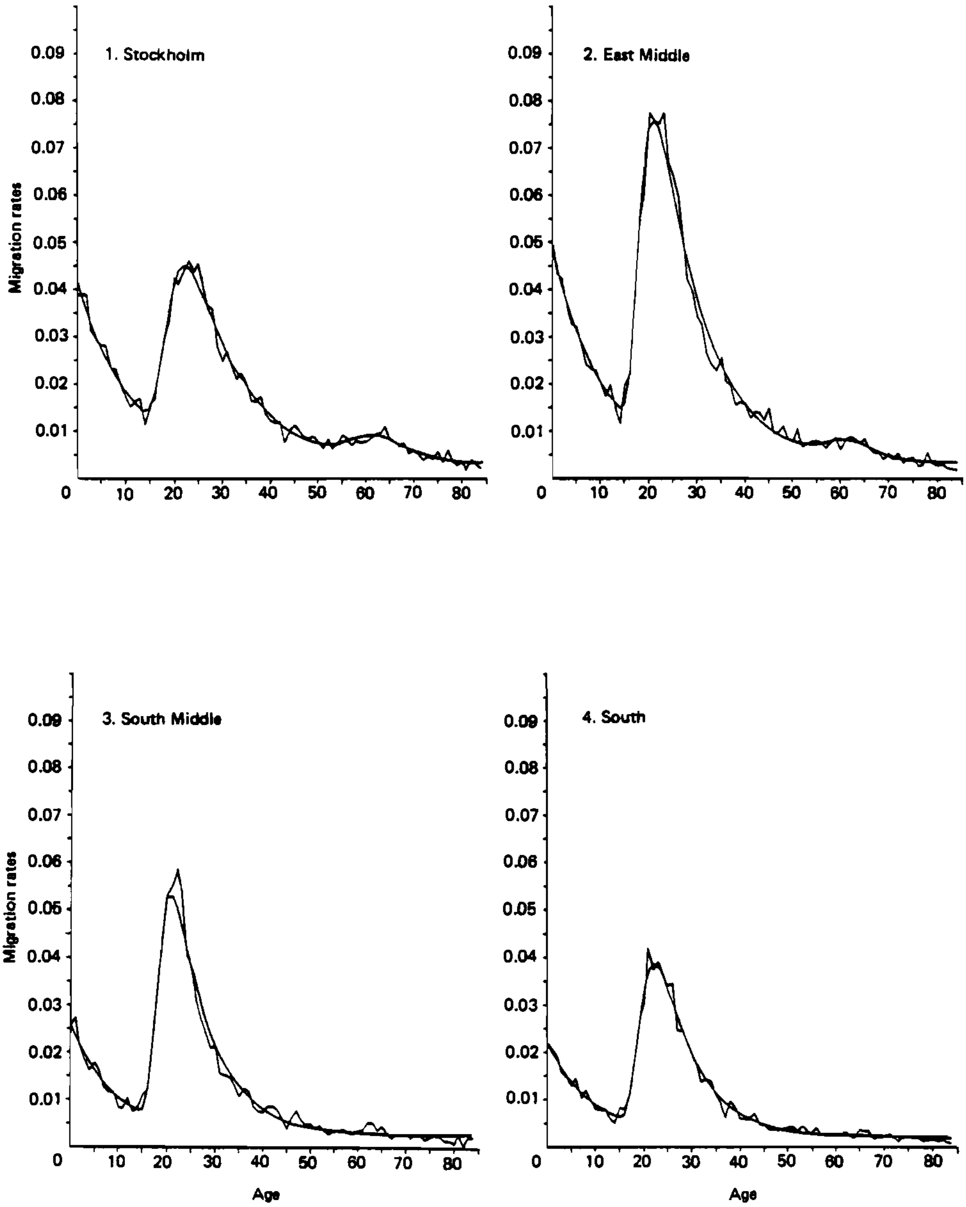


FIGURE 5 *continued on facing page.*

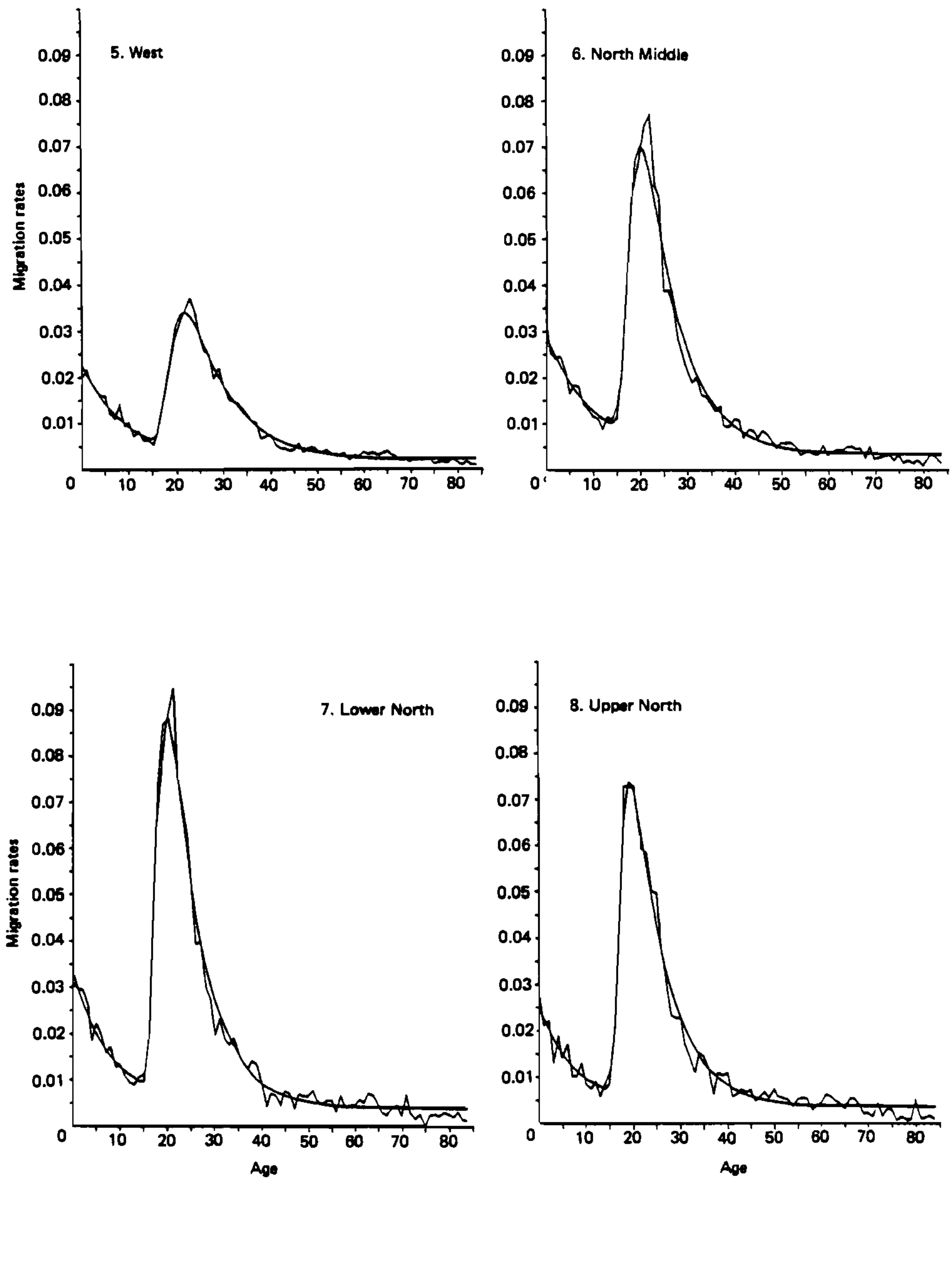


FIGURE 5 Observed (jagged line) and model (smooth line) migration schedules: females, Swedish regions, 1974.

and then to "add on" the pre-labor force, post-labor force, and constant components. This approach is represented graphically in Figure 6.

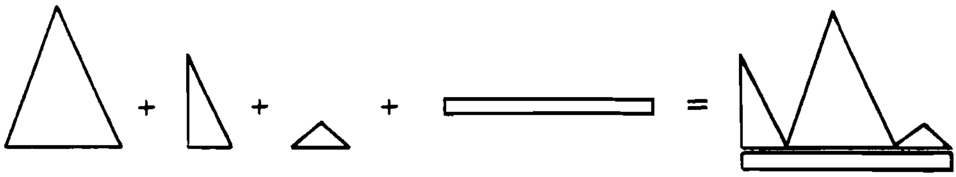


FIGURE 6 A schematic diagram of the fundamental components of the full model migration schedule.

One can imagine describing a decomposition of the model migration schedule along the vertical and horizontal dimensions; e.g., allocating a fraction of its level to the constant component and then dividing the remainder among the other three (or two) components. The ratio $\delta_{1c} = a_1/c$ measures the former allocation, and $\delta_{12} = a_1/a_2$ and $\delta_{32} = a_3/a_2$ reflect the latter division.

The heights of the labor force and pre-labor force components are reflected in the parameters a_2 and a_1, respectively, therefore the ratio a_2/a_1 indicates the degree of "labor dominance", and its reciprocal, $\delta_{12} = a_1/a_2$, the index of child dependency, measures the pace at which children migrate with their parents. Thus the lower the value of δ_{12}, the lower the degree of child dependency exhibited by a migration schedule and, correspondingly, the greater its labor dominance. This suggests a dichotomous classification of migration schedules into *child dependent* and *labor dominant* categories.

An analogous argument applies to the post-labor force curve, and $\delta_{32} = a_3/a_2$ suggests itself as the appropriate index. It will be sufficient for our purposes, however, to rely simply on the value taken on by the parameter α_3, with positive values pointing out the presence of a retirement peak and a zero value indicating its absence.

Labor dominance reflects the relative migration levels of those in the working ages relative to those of children and pensioners. *Labor asymmetry* refers to the shape of the left-skewed unimodal curve describing the age profile of labor force migration. Imagine that a perpendicular line, connecting the high peak with the base of the bell-shaped curve (i.e., the jump B), divides the base into two segments g and h as in Figure 7. Clearly, the ratio h/g is an indicator of the degree of asymmetry of the curve. A more convenient index, using only two parameters of the model schedule is the ratio $\sigma_2 = \lambda_2/\alpha_2$, the index of labor asymmetry. Its movement is highly correlated with that of h/g, because of the approximate relation

$$\sigma_2 = \lambda_2/\alpha_2 \propto \frac{B/g}{B/h} = h/g$$

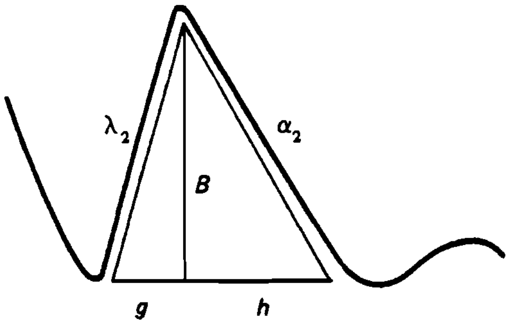


FIGURE 7 A schematic diagram of the curve describing the age profile of labor force migration.

where $\propto$ denotes proportionality. Thus σ_2 may be used to classify migration schedules according to their degree of labor asymmetry.

Again, an analogous argument applies to the post-labor force curve, and $\sigma_3 = \lambda_3/\alpha_3$ may be defined as the index of retirement asymmetry.

When "adding on" a pre-labor force curve of a given *level* to the labor force component, it is also important to indicate something of its *shape*. For example, if the migration rates of children mirror those of their parents, then α_1 should be approximately equal to α_2, and $\beta_{12} = \alpha_1/\alpha_2$, the index of parental-shift regularity, should be close to unity.

The Swedish regional migration patterns described in Figure 5 and in Table 1 may be characterized in terms of the various basic and derived measures defined above. We begin with the observation that the outmigration levels in all of the regions are similar, with *GMR*s ranging from a low of 0.80 for males in Region 5 to a high of 1.48 for females in Region 2. This similarity permits a reasonably accurate visual assessment and characterization of the profiles in Figure 5.

Large differences in *GMR*s, however, give rise to slopes and vertical relationships among schedules that are noncomparable when examined visually. Recourse then must be made to a standardization of the areas under the migration curves, for example, a general rescaling to a *GMR* of unity. Note that this difficulty does not arise in the numerical data in Table 1, because, as we pointed out earlier, the principal slope and location parameters and ratios used to characterize the schedules are not affected by changes in levels. Only heights, areas, and vertical distances, such as the jump, are level-dependent measures.

Among the eight regions examined, only the first two exhibit a definite retirement peak, the male peak being the more dominant one in each case. The index of child dependency δ_{12} is highest in Region 1 and lowest in Region 8, distinguishing the latter region's labor dominant profile from Stockholm's child dependent outmigration pattern. The index of labor asymmetry σ_2 varies from a low of 2.34, in the case of males in Region 4 to a high of 4.95 for the female outmigration profile of Region 8. Finally, with the possible exception of males in Region 1 and females in Region 6, the migration rates of children in Sweden do indeed seem to mirror those of their parents. The index of parental-shift regularity β_{12} is 1.26 in the former case and 0.730 in the latter; for most of the other schedules it is close to unity.

3 A COMPARATIVE ANALYSIS OF OBSERVED MODEL MIGRATION SCHEDULES

Section 2 demonstrated that age-specific rates of migration exhibit a fundamental age profile, which can be expressed in mathematical form as a model migration schedule defined by a total of 11 parameters. In this section we seek to establish the ranges of values typically assumed by each of these parameters and their associated derived variables. This exercise is made possible by the availability of a relatively large data base collected by the Comparative Migration and Settlement Study, recently concluded at IIASA (Rogers 1976a, 1976b, 1978, Rogers and Willekens 1978, Willekens and Rogers 1978). The migration data for each of the 17 countries included in this study are set out in individual case studies, which are listed at the end of this report.

3.1 Data Preparation, Parameter Estimation, and Summary Statistics

The age-specific migration rates that were used to demonstrate the fits of the model migration schedule in the last section were single-year rates. Such data are scarce at the regional level and, in our comparative analysis, are available only for Sweden. All other region-specific migration data are reported for five-year age groups only and, therefore, must be interpolated to provide the necessary input data by single years of age. In all such instances the region-specific migration schedules were first scaled to a *GMR* of unity (*GMR* = 1) before being subjected to a cubic-spline interpolation (McNeil et al. 1977).

Starting with a migration schedule with a *GMR* of unity and rates by single years of age, the nonlinear parameter estimation algorithm ultimately yields a set of estimates for the model schedule's parameters (see Appendix A for details). Table 1 in section 2 presented the results that were obtained using the data for Sweden. Since these data were available for single years of age, the influence of the interpolation procedure could be

TABLE 2 Parameters defining observed model migration schedules and parameters obtained after a cubic-

	Region and width of age group							
	1. Stockholm		2. East Middle		3. South Middle		4. South	
Parameters	1 yr	5 yr	1 yr	5 yr	1 yr	5 yr	1 yr	5 yr
a_1	0.029	0.028	0.026	0.026	0.023	0.023	0.025	0.025
α_1	0.091	0.089	0.108	0.106	0.106	0.105	0.104	0.106
a_2	0.047	0.049	0.065	0.070	0.080	0.087	0.080	0.085
μ_2	19.32	19.69	18.52	18.99	18.49	18.93	19.88	20.23
α_2	0.094	0.098	0.109	0.117	0.127	0.136	0.129	0.135
λ_2	0.369	0.313	0.491	0.351	0.560	0.375	0.442	0.367
c	0.002	0.002	0.003	0.003	0.003	0.003	0.003	0.003
a_3	0.000	0.000						
μ_3	85.01	81.20						
α_3	0.369	0.364						
λ_3	0.072	0.080						

[a]Observed data are for single years of age (1 yr); the cubic-spline-interpolated inputs are obtained from observed

assessed. Table 2 contrasts the estimates for female schedules in Table 1 with those obtained when the same data are first aggregated to five-year age groups and then disaggregated to single years of age by a cubic-spline interpolation. A comparison of the parameter estimates indicates that the interpolation procedure gives generally satisfactory results.

Table 2 refers to results for rates of migration from each of eight regions to the rest of Sweden. If these rates are disaggregated by region of destination, then $8^2 = 64$ inter-regional schedules need to be examined for each sex, which will complicate comparisons with other nations. To resolve this difficulty we shall associate a "typical" schedule with each collection of national rates by calculating the mean of each parameter and derived variable. Table 3 illustrates the results for the Swedish data.

To avoid the influence of unrepresentative "outlier" observations in the computation of averages defining a typical national schedule, it was decided to delete approximately 10 percent of the "extreme" schedules. Specifically, the parameters and derived variables were ordered from low value to high value; the lowest 5 percent and the highest 5 percent were defined to be extreme values. Schedules with the largest number of low and high extreme values were discarded, in sequence, until only about 90 percent of the original number of schedules remained. This reduced set then served as the population of schedules for the calculation of various summary statistics. Table 4 illustrates the average parameter values obtained with the Swedish data. Since the median, mode, standard deviation-to-mean ratio, and lower and upper bounds are also of interest, they are included as part of the more detailed computer outputs reproduced in Appendix B.

The comparison, in Table 2, of estimates obtained using one-year and five-year age intervals for the same Swedish data indicated that the interpolation procedure gave satisfactory results. It also suggested, however, that the parameter λ_2 was consistently underestimated with five-year data. To confirm this, the results of Table 4 were replicated with the Swedish data base, using an aggregation with five-year age intervals. The results, set out in Table 5, show once again that λ_2 is always underestimated by the interpolation procedure. This tendency should be noted and kept in mind.

spline interpolation: Sweden, 8 regions, females, 1974.[a]

5. West		6. North Middle		7. Lower North		8. Upper North	
1 yr	5 yr	1 yr	5 yr	1 yr	5 yr	1 yr	5 yr
0.027	0.025	0.021	0.022	0.021	0.021	0.019	0.021
0.106	0.095	0.102	0.115	0.119	0.130	0.128	0.160
0.067	0.069	0.087	0.097	0.096	0.118	0.094	0.106
19.36	19.72	18.18	18.57	17.93	19.11	17.62	18.00
0.114	0.121	0.139	0.145	0.148	0.172	0.143	0.150
0.442	0.395	0.561	0.345	0.701	0.305	0.711	0.330
0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003

data by five-year age groups (5 yr).

TABLE 3 Mean values of parameters defining the full set of observed model migration schedules: Sweden, 8 regions, 1974 observed data by single years of age until 84 years and over.

Parameters	Males		Females	
	Without retirement peak (52 schedules)	With retirement peak (11 schedules)	Without retirement peak (58 schedules)	With retirement peak (5 schedules)
a_1	0.029	0.025	0.027	0.023
α_1	0.126	0.080	0.114	0.087
a_2	0.066	0.050	0.078	0.051
μ_2	21.09	21.52	19.13	19.20
α_2	0.113	0.096	0.133	0.101
λ_2	0.459	0.439	0.525	0.377
c	0.003	0.002	0.003	0.003
a_3		0.0012		0.0017
μ_3		75.45		72.07
α_3		0.797		0.688
λ_3		0.294		0.192

[a]Region 1 (Stockholm) is a single-commune region; hence there exists no intraregional schedule for it, leaving $8^2 - 1 = 63$ schedules.

TABLE 4 Mean values of parameters defining the reduced set of observed model migration schedules: Sweden, 8 regions, 1974 observed data by single years of age until 84 years and over.[a]

Parameters	Males		Females	
	Without retirement peak (48 schedules)	With retirement peak (9 schedules)	Without retirement peak (54 schedules)	With retirement peak (3 schedules)
a_1	0.029	0.026	0.026	0.024
α_1	0.124	0.085	0.108	0.093
a_2	0.067	0.051	0.076	0.055
μ_2	20.50	21.25	19.09	18.87
α_2	0.104	0.093	0.127	0.106
λ_2	0.448	0.416	0.537	0.424
c	0.003	0.002	0.003	0.003
a_3		0.0006		0.0001
μ_3		76.71		74.78
α_3		0.847		0.938
λ_3		0.158		0.170

[a]Region 1 (Stockholm) is a single-commune region; hence there exists no intraregional schedule for it, leaving $8^2 - 1 = 63$ schedules, of which 6 were deleted.

It is also important to note the erratic behavior of the retirement peak, apparently due to its extreme sensitivity to the loss of information arising out of the aggregation. Thus, although we shall continue to present results relating to the post-labor force ages, they will not be a part of our search for families of schedules.

TABLE 5 Mean values of parameters defining the reduced set of observed model migration schedules: Sweden, 8 regions, 1974 observed data by five years of age until 80 years and over.[a]

	Males		Females	
Parameters	Without retirement peak (49 schedules)	With retirement peak (8 schedules)	Without retirement peak (54 schedules)	With retirement peak (3 schedules)
a_1	0.028	0.026	0.026	0.026
α_1	0.115	0.088	0.108	0.077
a_2	0.068	0.052	0.080	0.044
μ_2	20.61	20.26	19.52	19.18
α_2	0.105	0.084	0.133	0.089
λ_2	0.396	0.390	0.374	0.341
c	0.002	0.001	0.002	0.002
a_3		0.0017		0.0036
μ_3		77.47		77.72
α_3		0.603		0.375
λ_3		0.148		0.134

[a]Region 1 (Stockholm) is a single-commune region; hence there exists no intraregional schedule for it, leaving $8^2 - 1 = 63$ schedules, of which 6 were deleted.

3.2 National Contrasts

Tables 4 and 5 of the preceding subsection summarized average parameter values for 57 male and 57 female Swedish model migration schedules. In this subsection we shall expand our analysis to include a much larger data base, adding to the 114 Swedish model schedules another 164 schedules from the United Kingdom (Table 6), 114 from Japan, 20 from the Netherlands (Table 7), 58 from the Soviet Union, 8 from the United States, and 32 from Hungary (Table 8). Summary statistics for these 510 schedules are set out in

TABLE 6 Mean values of parameters defining the reduced set of observed model migration schedules: the United Kingdom, 10 regions, 1970.[a]

	Males		Females	
Parameters	Without retirement peak (59 schedules)	With retirement peak (23 schedules)	Without retirement peak (61 schedules)	With retirement peak (21 schedules)
a_1	0.021	0.016	0.021	0.018
α_1	0.099	0.080	0.097	0.089
a_2	0.059	0.053	0.063	0.048
μ_2	22.00	20.42	21.35	21.56
α_2	0.127	0.120	0.151	0.153
λ_2	0.259	0.301	0.327	0.333
c	0.003	0.004	0.003	0.004
a_3		0.007		0.002
μ_3		71.11		71.84
α_3		0.692		0.583
λ_3		0.309		0.403

[a]No intraregional migration data were included in the United Kingdom data; hence $10^2 - 10 = 90$ schedules were analyzed, of which 8 were deleted.

TABLE 7 Mean values of parameters defining the reduced set of observed model migration schedules: Japan, 8 regions, 1970; the Netherlands, 12 regions, 1974.[a]

Parameters	Japan Males Without retirement peak (57 schedules)	Japan Females Without retirement peak (57 schedules)	Netherlands Males With retirement slope (10 schedules)	Netherlands Females With retirement slope (10 schedules)
a_1	0.014	0.021	0.013	0.012
α_1	0.095	0.117	0.080	0.098
a_2	0.075	0.085	0.063	0.084
μ_2	17.63	21.32	20.86	20.10
α_2	0.102	0.152	0.130	0.174
λ_2	0.480	0.350	0.287	0.307
c	0.002	0.004	0.003	0.004
a_3			0.00001	0.00004
α_3			0.077	0.071

[a]Region 1 in Japan (Hokkaido) is a single-prefecture region; hence there exists no intraregional schedule for it, leaving $8^2 - 1 = 63$ schedules, of which 6 were deleted. The only migration schedules available for the Netherlands were the migration rates out of each region without regard to destination; hence only 12 schedules were used, of which 2 were deleted.

TABLE 8 Mean values of parameters defining the reduced set of observed total (males plus females) model migration schedules: the Soviet Union, 8 regions, 1974; the United States, 4 regions, 1970–1971; Hungary, 6 regions, 1974.[a]

Parameters	Soviet Union Without retirement peak (58 schedules)	United States With retirement peak (8 schedules)	Hungary Without retirement slope (7 schedules)	Hungary With retirement slope (25 schedules)
a_1	0.005	0.021	0.010	0.015
α_1	0.302	0.075	0.245	0.193
a_2	0.126	0.060	0.090	0.099
μ_2	19.14	20.14	17.22	18.74
α_2	0.176	0.118	0.130	0.159
λ_2	0.310	0.569	0.415	0.274
c	0.004	0.002	0.004	0.003
a_3		0.002		0.00032
μ_3		81.80		
α_3		0.430		0.033
λ_3		0.119		

[a]Intraregional migration was included in the Soviet Union and Hungarian data but not in the United States data; hence there were $8^2 = 64$ schedules for the Soviet Union, of which 6 were deleted, $6^2 = 36$ schedules for Hungary, of which 4 were deleted, and $4^2 - 4 = 12$ schedules for the United States, of which 2 were deleted because they lacked a retirement peak and another 2 were deleted because of their extreme values.

Appendix B; 206 are male schedules, 206 are female schedules, and 98 are for the combination of both sexes (males plus females).*

*This total does not include the 56 schedules excluded as "extreme" schedules. During the process of fitting the model schedule to these more than 500 interregional migration schedules, a frequently encountered problem was the occurrence of a negative value for the constant c. In all such instances

A significant number of schedules exhibited a pattern of migration in the post-labor force ages that differed from that of the 11-parameter model migration schedule defined in eq. (1). Instead of a retirement peak, the age profile took on the form of an "upward slope". In such instances the following 9-parameter modification of the basic model migration was introduced

$$\left.\begin{aligned} M(x) &= a_1 \exp(-\alpha_1 x) \\ &+ a_2 \exp\{-\alpha_2(x-\mu_2) - \exp[-\lambda_2(x-\mu_2)]\} \\ &+ a_3 \exp(\alpha_3 x) \\ &+ c \end{aligned}\right\} \quad x = 0,1,2,\ldots,z \qquad (3)$$

The right-hand side of Table 7, for example, sets out the mean parameter estimates of this modified form of the model migration schedule for the Netherlands.

Tables 4 through 8 present a wealth of information about national patterns of migration by age. The parameters, given in columns, define a wide range of model migration schedules. Four refer only to migration level: a_1, a_2, a_3, and c. Their values are for a *GMR* of unity; to obtain corresponding values for other levels of migration, these four numbers need to be multiplied by the desired level of *GMR*. For example, the observed *GMR* for female migration out of the Stockholm region in 1974 was 1.43. Multiplying $a_1 = 0.029$ by 1.43 gives 0.041, the appropriate value of a_1 with which to generate the migration schedule having a *GMR* of 1.43.

The remaining model schedule parameters refer to migration age profile: $\alpha_1, \mu_2, \alpha_2, \lambda_2, \mu_3, \alpha_3$, and λ_3. Their values remain constant for all levels of the *GMR*. Taken together, they define the age profile of migration from one region to another. Schedules without a retirement peak yield only the four profile parameters: $\alpha_1, \mu_2, \alpha_2$, and λ_2, and schedules with a retirement slope have an additional profile parameter α_3.

A detailed analysis of the parameters defining the various classes of schedules is beyond the scope of this report. Nevertheless a few basic contrasts among national average age profiles may be usefully highlighted.

Let us begin with an examination of the labor force component defined by the four parameters a_2, μ_2, α_2, and λ_2. The national average values for these parameters generally lie within the following ranges:

$$0.05 < a_2 < 0.10$$

$$17 < \mu_2 < 22$$

$$0.10 < \alpha_2 < 0.20$$

$$0.25 < \lambda_2 < 0.60$$

---

the initial value of c was set equal to the lowest observed migration rate and the nonlinear estimation procedure was started once again.

In all but two instances, the female values for a_2, α_2, and λ_2 are larger than those for males. The reverse is the case for μ_2, with two exceptions, the most important of which is exhibited by Japan's females, who consistently show a high peak that is older than that of males. This apparently is a consequence of the tradition in Japan that girls leave the family home at a later age than boys.

The two parameters defining the pre-labor force component, a_1 and α_1, generally lie within the ranges of 0.01 to 0.03 and 0.08 to 0.12, respectively. The exceptions are the Soviet Union and Hungary, which exhibit unusually high values for α_1. Unlike the case of the labor force component, consistent sex differentials are difficult to identify.

Average national migration age profiles, like most aggregations, hide more than they reveal. Some insight into the ranges of variations that are averaged out may be found by consulting the lower and upper bounds and standard-deviation-to-mean ratios listed in Appendix B for each set of national schedules. Additional details are set out in Appendix C. Finally, Table 9 illustrates how parameters vary in several *unaveraged* national schedules, by way of example. The model schedules presented there describe migration flows out of and into the capital regions of each of six countries: Helsinki, Finland; Budapest, Hungary; Tokyo, Japan; Amsterdam, the Netherlands; Stockholm, Sweden; and London, the United Kingdom. All are illustrated in Figure 8.

The most apparent difference between the age profiles of the outflow and inflow migration schedules of the six national capitals is the dominance of young labor force migrants in the inflow, that is, proportionately more migrants in the young labor force ages appear in the inflow schedules. The larger values of the product $a_2\lambda_2$ in the inflow schedules and of the ratio $\delta_{12} = a_1/a_2$ in the outflow schedules indicate this labor dominance.

A second profile attribute is the degree of asymmetry in the labor force component of the migration schedule, i.e., the ratio of the rate of ascent λ_2 to the rate of descent α_2 defined as σ_2 in section 2. In all but the Japanese case, the labor force curves of the capital-region outmigration profiles are more asymmetric than those of the corresponding inmigration profiles. We refer to this characteristic as labor asymmetry.

Examining the observed rates of descent of the labor and pre-labor force curves, α_2 and α_1, respectively, we find, for example, that they are close to being equal in the outflow

TABLE 9 Parameters defining observed total (males plus females) model migration schedules for flows 1974; the United Kingdom, 1970.

	Finland		Hungary		Japan	
Parameters	From Helsinki	To Helsinki	From Budapest	To Budapest	From Tokyo	To Tokyo
a_1	0.037	0.024	0.015	0.008	0.019	0.008
α_1	0.127	0.170	0.239	0.262	0.157	0.149
a_2	0.081	0.130	0.082	0.094	0.064	0.096
μ_2	21.42	22.13	17.10	17.69	20.70	15.74
α_2	0.124	0.198	0.130	0.152	0.111	0.134
λ_2	0.231	0.231	0.355	0.305	0.204	0.577
c	0.000	0.003	0.003	0.003	0.003	0.002
a_3	0.00027		0.00001	0.00005	0.00002	0.00131
μ_3	99.32					
α_3	0.204		0.072	0.059	0.061	0.000
λ_3	0.042					

schedules of Helsinki and Stockholm and are highly unequal in the cases of Budapest, Tokyo, and Amsterdam. In four of the six capital-region inflow profiles $\alpha_2 > \alpha_1$. Profiles with significantly different values for α_2 and α_1 are said to be irregular.

In conclusion, the empirical migration data of six industrialized nations suggest the following hypothesis. *The age profile of a typical capital-region inmigration schedule is, in general, more labor dominant and more labor symmetric than the age profile of the corresponding capital-region outmigration schedule.* No comparable hypothesis can be made regarding its anticipated degree of irregularity.

3.3 Families of Schedules

Three sets of model migration schedules have been defined in this report: the 11-parameter schedule with a retirement peak, the alternative 9-parameter schedule with a retirement slope, and the simple 7-parameter schedule with neither a peak nor a slope. Thus we have at least three broad families of schedules.

Additional dimensions for classifying schedules into families are suggested by the above comparative analysis of national migration age profiles and the basic measures and derived variables defined in section 2. These dimensions reflect different locations on the horizontal and vertical axes of the schedule, as well as different ratios of slopes and heights.

Of the 524 model migration schedules studied in this section, 412 are sex-specific and, of these, only 336 exhibit neither a retirement peak nor a retirement slope. Because the parameter estimates describing the age profile of post-labor force migration behave erratically, we shall restrict our search for families of schedules to these 164 male and 172 female model schedules, summary statistics for which are set out in Tables 10 and 11.

An examination of the parametric values exhibited by the 336 migration schedules summarized in Tables 10 and 11 suggests that a large fraction of the variation shown by these schedules is a consequence of changes in the values of the following four parameters and derived variables: μ_2, δ_{12}, σ_2, and β_{12}.

from and to capital cities: Finland, 1974; Hungary, 1974; Japan, 1970; the Netherlands, 1974; Sweden,

Netherlands		Sweden		United Kingdom	
From Amsterdam	To Amsterdam	From Stockholm	To Stockholm	From London	To London
0.015	0.012	0.028	0.018	0.015	0.014
0.085	0.108	0.098	0.102	0.090	0.072
0.050	0.093	0.046	0.093	0.048	0.067
21.62	19.66	20.48	19.20	19.65	18.81
0.141	0.150	0.095	0.134	0.111	0.123
0.284	0.288	0.322	0.323	0.327	0.320
0.002	0.003	0.003	0.002	0.005	0.004
0.00229	0.00002	0.00004	0.00003	0.00003	
		80.32	73.19	81.13	
0.012	0.066	0.616	1.359	0.676	
		0.105	0.255	0.112	

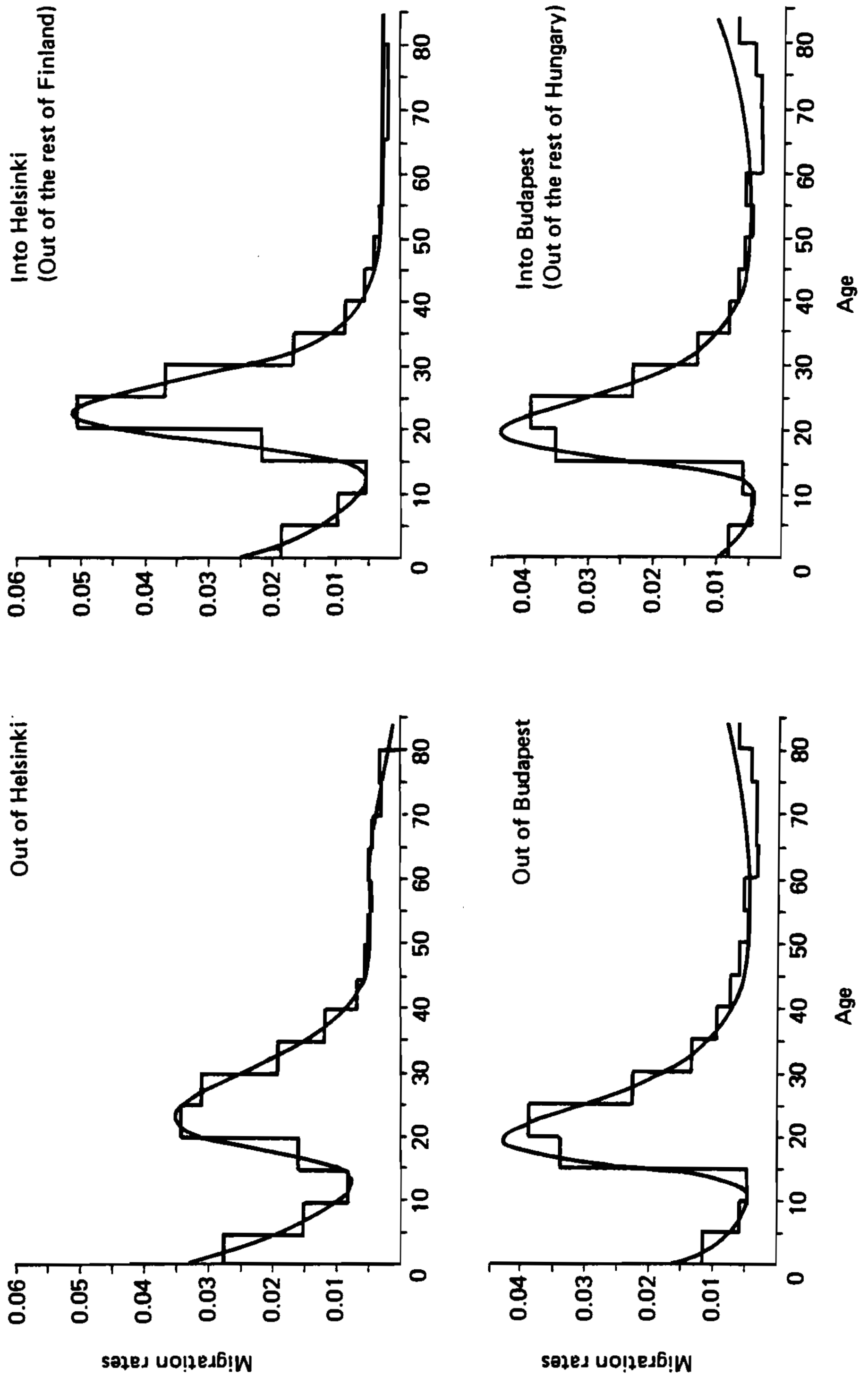


FIGURE 8 *continued on facing page.*

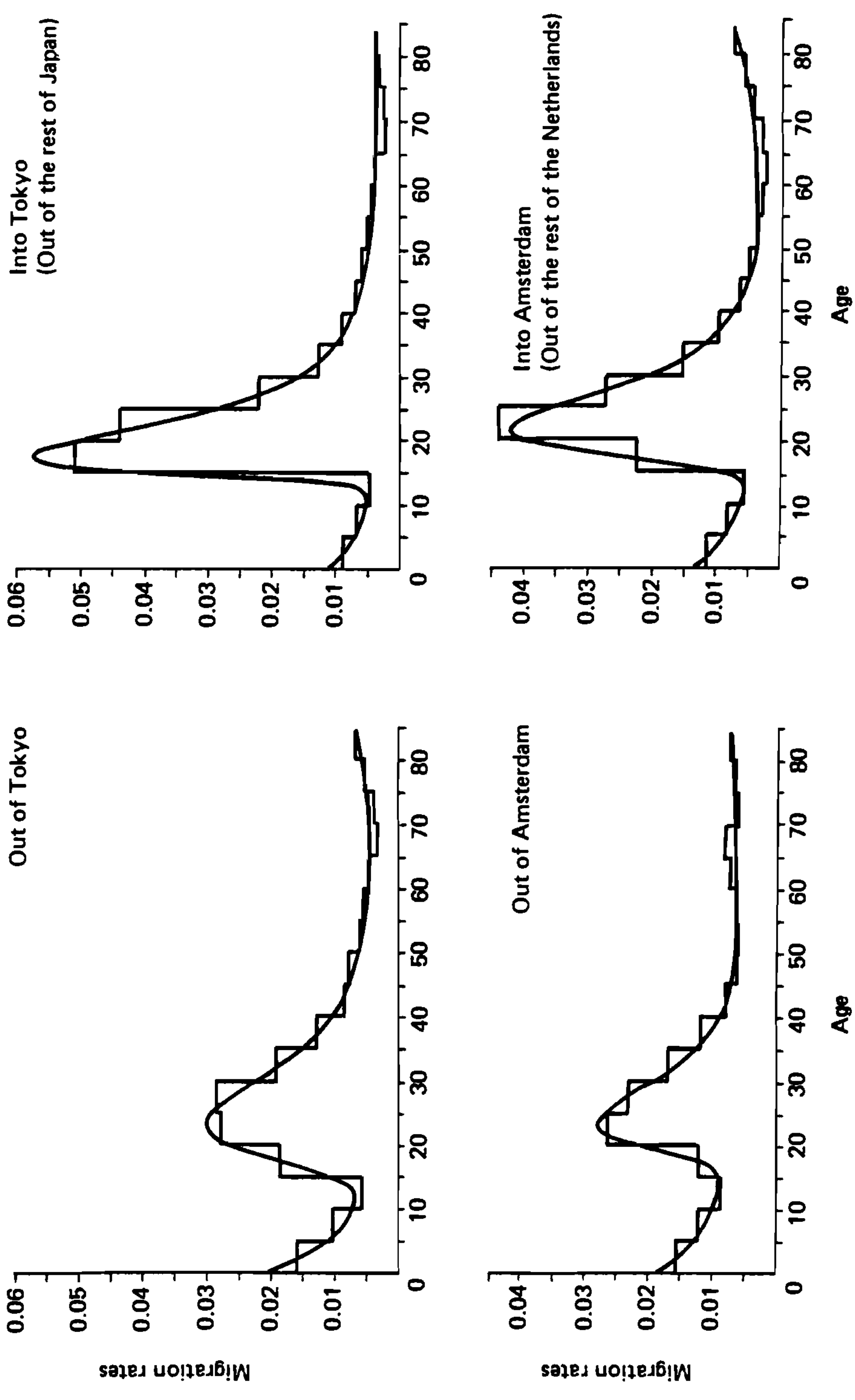


FIGURE 8 *continued overleaf.*

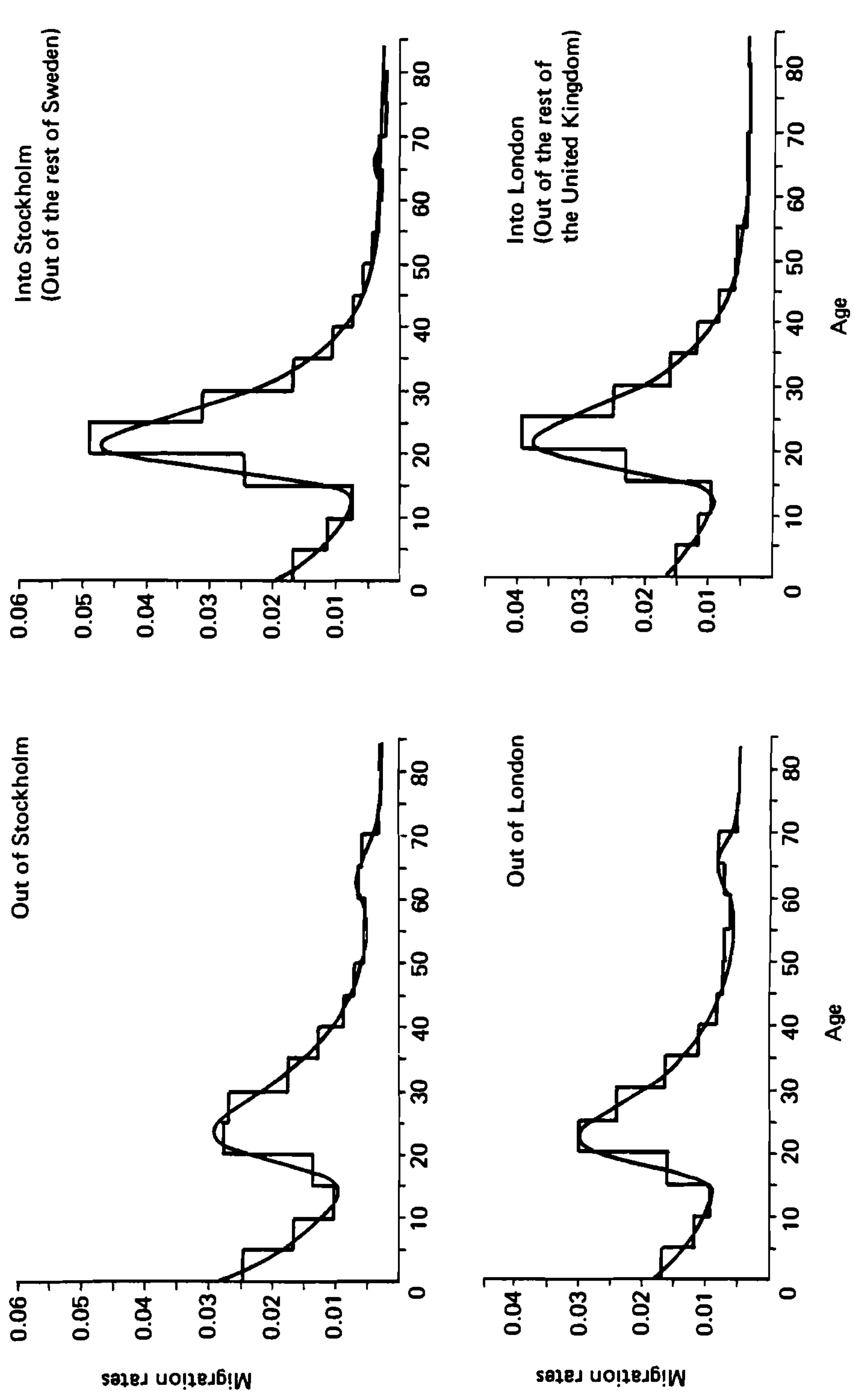


FIGURE 8 Migration age profiles of outflows from and inflows to capital cities: Helsinki, Budapest, Tokyo, Amsterdam, Stockholm, and London.

TABLE 10 Estimated summary statistics of parameters and variables associated with reduced sets of observed model migration schedules for Sweden, the United Kingdom, and Japan: males, 164 schedules.[a]

Parameters and variables	Summary statistics: Lowest value	Highest value	Mean value	Median	Mode	Standard deviation	Standard deviation/ mean
GMR (observed)	0.00539	1.81309	0.22642	0.13176	0.09578	0.27380	1.20928
GMR (model)	1.00000	1.00000	1.00000	1.00000	1.00000	0.00000	0.00000
E	4.75751	62.98674	16.22228	13.10527	13.49189	9.95789	0.61384
a_1	0.00173	0.04891	0.02084	0.01992	0.01824	0.00879	0.42204
α_1	0.00009	0.40526	0.10491	0.10390	0.10138	0.05358	0.51077
a_2	0.01559	0.22707	0.06716	0.06471	0.06846	0.02578	0.38391
μ_2	14.68744	43.96579	20.04227	19.67385	19.07919	3.95015	0.19709
α_2	0.03471	0.29735	0.11164	0.10618	0.10037	0.04389	0.39316
λ_2	0.06951	1.76712	0.39110	0.37244	0.31650	0.21146	0.54068
c	0.00003	0.00704	0.00266	0.00263	0.00248	0.00130	0.48947
$\bar{n}$	24.71596	40.53283	30.71751	30.41339	30.25187	2.72144	0.08860
%(0–14)	4.92484	29.69068	18.93871	19.02262	18.54605	4.91304	0.25942
%(15–64)	60.27293	86.29065	72.08085	71.29800	66.77736	5.10213	0.07078
%(65+)	1.35294	17.31658	8.98045	8.71650	8.53658	3.49047	0.38867
δ_{1c}	0.37762	712.88135	14.36314	6.79034	36.00280	56.75620	3.95152
δ_{12}	0.02274	1.53679	0.35774	0.33571	0.24985	0.20221	0.56523
β_{12}	0.00092	7.47530	1.11318	1.02442	1.12208	0.81866	0.73542
σ_2	0.30349	24.23831	4.27564	3.42123	3.89371	3.26113	0.76272
x_l	6.91004	18.26030	13.72508	13.34019	12.01766	2.14485	0.15627
x_h	17.11028	28.14053	22.50278	22.95041	23.17692	2.14731	0.09542
X	2.90007	16.93039	8.77770	8.38019	7.81068	2.28557	0.26038
A	22.33532	102.41312	32.97422	31.54365	34.34699	7.58660	0.23008
B	0.01107	0.07343	0.02994	0.02775	0.02666	0.01036	0.34609

[a] A list of definitions for the parameters and variables appears in Appendix B.

TABLE 11 Estimated summary statistics of parameters and variables associated with reduced sets of observed model migration schedules for Sweden, the United Kingdom, and Japan: females, 172 schedules.[a]

Parameters and variables	Summary statistics: Lowest value	Highest value	Mean value	Median	Mode	Standard deviation	Standard deviation/ mean
GMR (observed)	0.00388	1.59564	0.19909	0.11590	0.08347	0.24085	1.20973
GMR (model)	1.00000	1.00000	1.00000	1.00000	1.00000	0.00000	0.00000
E	4.17964	60.83579	15.42092	12.26192	7.01245	9.85544	0.63910
a_1	0.00526	0.04496	0.02259	0.02209	0.01916	0.00851	0.37664
α_1	0.01585	0.41038	0.10698	0.10883	0.11448	0.05091	0.47587
a_2	0.02207	0.18944	0.07426	0.06935	0.06391	0.02693	0.36263
μ_2	15.06610	37.76019	20.63237	19.88280	18.47021	3.50346	0.16980
α_2	0.05467	0.33556	0.14355	0.13434	0.12489	0.04993	0.34784
λ_2	0.08367	1.49869	0.40032	0.37870	0.29592	0.19248	0.48081
c	0.00012	0.00685	0.00347	0.00350	0.00315	0.00139	0.39940
$\bar{n}$	24.51402	37.86541	30.65265	30.53835	29.18701	2.69720	0.08799
%(0–14)	9.37675	31.87480	20.93872	20.68939	19.50087	4.26504	0.20369
%(15–64)	60.55278	81.17286	68.65491	68.07751	67.76981	4.34828	0.06334
%(65+)	1.46164	19.56255	10.40638	10.32867	9.60705	3.40400	0.32711
δ_{1c}	0.89359	192.60318	9.39987	5.95881	10.47907	16.22411	1.72602
δ_{12}	0.02828	0.90435	0.34847	0.32367	0.33490	0.17420	0.49989
β_{12}	0.09121	2.48385	0.81472	0.84944	0.92863	0.37720	0.46298
σ_2	0.38917	12.23371	3.26434	2.89784	2.16585	2.12718	0.65164
x_l	10.32012	21.79038	14.51330	14.75022	14.33471	1.95309	0.13457
x_h	17.03028	30.92059	22.49959	22.46040	21.89189	2.14262	0.09523
X	2.89007	15.09035	7.98629	7.61017	7.16017	2.11207	0.26446
A	23.73040	37.24700	28.50972	28.17807	27.10955	2.47098	0.08667
B	0.00831	0.09111	0.03118	0.02970	0.02901	0.01149	0.36845

[a] A list of definitions for the parameters and variables appears in Appendix B.

Migration schedules may be early or late peaking, depending on the location of μ_2 on the horizontal (age) axis. Although this parameter generally takes on a value close to 20, roughly three out of four observations fall within the range 17–25. We shall call those below age 19 early peaking schedules and those above 22 late peaking schedules.

The ratio of the two basic vertical parameters, a_1 and a_2, is a measure of the relative importance of the migration of children in a model migration schedule. The index of child dependency, $\delta_{12} = a_1/a_2$, tends to exhibit a mean value of about one-third with 80 percent of the values falling between one-fifth and four-fifths. Schedules with an index of one-fifth or less will be said to be labor dominant; those above two-fifths will be called child dependent.

Migration schedules with labor force components that take the form of a relatively symmetrical bell shape will be said to be *labor symmetrical.* These schedules will tend to exhibit an index of labor asymmetry ($\sigma_2 = \lambda_2/\alpha_2$) that is less than 2. Labor asymmetric schedules, on the other hand, will usually assume values for σ_2 of 5 or more. The average migration schedule will tend to show a σ_2 value of about 4, with approximately five out of six schedules exhibiting a σ_2 within the range 1–8.

Finally, the index of parental-shift regularity in many schedules is close to unity, with approximately 70 percent of the values lying between one-third and four-thirds. Values of $\beta_{12} = \alpha_1/\alpha_2$ that are lower than four-fifths or higher than six-fifths will be called irregular.

We may imagine a 3 X 4 cross-classification of migration schedules that defines a dozen "average families" (Table 12). Introducing a low and a high value for each parameter gives rise to 16 additional families for each of the three classes of schedules. Thus we may conceive of a minimum set of 60 families, equally divided among schedules with a retirement peak, schedules with a retirement slope, and schedules with neither a retirement peak nor a retirement slope (a reduced form).

TABLE 12 A cross-classification of migration schedules.

	Measures (average values)			
Schedule	Peaking ($\mu_2 = 20$)	Dominance ($\delta_{12} = 1/3$)	Asymmetry ($\sigma_2 = 4$)	Regularity ($\beta_{12} = 1$)
Retirement peak	+	+	+	+
Retirement slope	+	+	+	+
Reduced form	+	+	+	+

To complement the above discussion with a few visual illustrations, in Figure 9(a) we present six labor dominant profiles, with δ_{1c} fixed at 22. The tallest three exhibit a steep rate of descent $\alpha_2 = 0.3$; the shortest three show a much more moderate slope of $\alpha_2 = 0.06$. Within each family of three curves, one finds variations in μ_2 and in the rate of ascent λ_2. Increasing μ_2 shifts the curve to the right along the horizontal axis; increasing λ_2 raises the relative height of the high peak.

The six schedules in Figure 9(b) depict the corresponding two families of child dependent profiles. The results are generally similar to those in Figure 9(a), with the

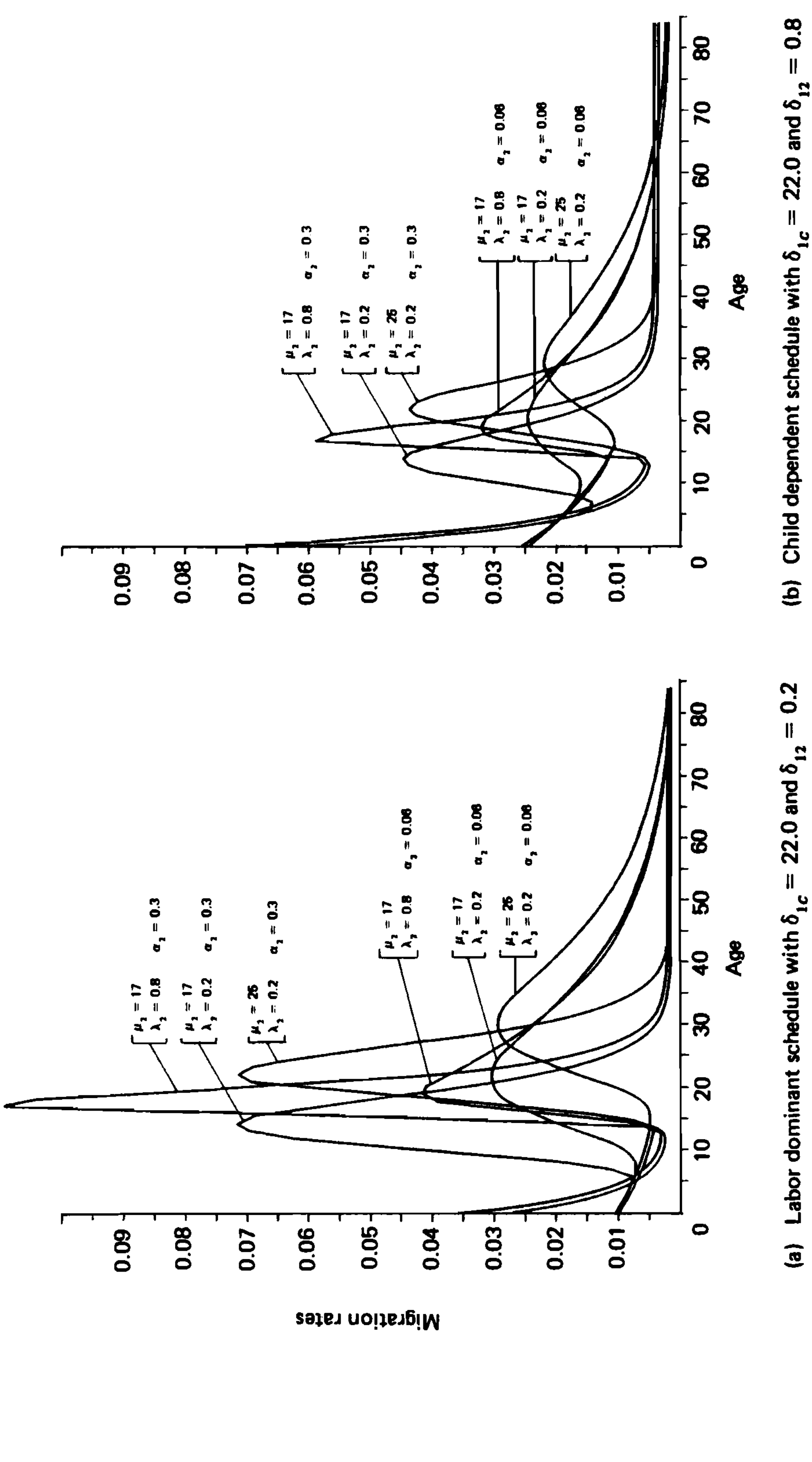
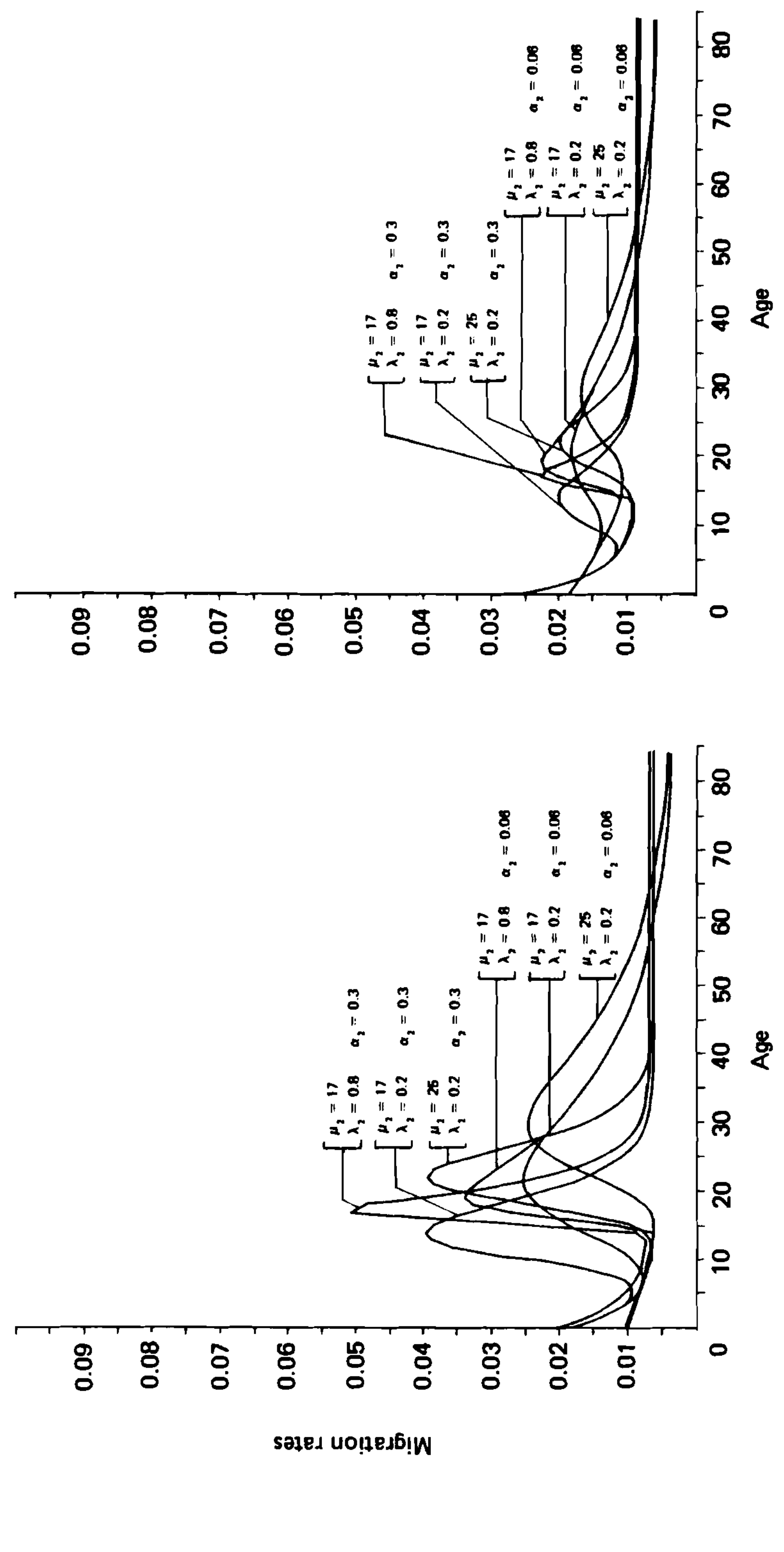


FIGURE 9 *continued on facing page.*



(c) Labor dominant schedule with $\delta_{1c} = 2.6$ and $\delta_{12} = 0.2$

(d) Child dependent schedule with $\delta_{1c} = 2.6$ and $\delta_{12} = 0.8$

FIGURE 9 Hypothetical model migration schedules with unit *GMR*s, $\beta_{12} = 1$, and different parameter combinations.

exception that the relative importance of migration in the pre-labor force age groups is increased considerably. The principal effects of the change in δ_{12} are: (1) a raising of the intercept $a_1 + c$ along the vertical axis, and (2) a simultaneous reduction in the height of the labor force component in order to maintain a constant area of unity under each curve.

Finally, the dozen schedules in Figures 9(c) and 9(d) describe similar families of migration curves, but in these profiles the relative contribution of the constant component to the unit *GMR* has been increased significantly (i.e., $\delta_{1c} = 2.6$). It is important to note that such "pure" measures of profiles as x_1, x_h, X, and A remain unaffected by this change, whereas "impure" profile measures, such as the mean age of migration $\bar{n}$, now take on a different set of values.

3.4 Sensitivity Analysis

The preceding subsections have focused on a comparison of the fundamental parameters defining the model migration age profiles of a number of nations. The comparison yielded ranges of values within which each parameter may be expected to fall and suggested a classification of schedules into families. We now turn to an analytic examination of how changes in several of the more important parameters become manifested in the age profile of the model schedule. For analytical convenience we begin by focusing on the properties of the double exponential curve that describes the labor force component:

$$f_2(x) = a_2 \exp\{-\alpha_2(x - \mu_2) - \exp[-\lambda_2(x - \mu_2)]\} \tag{4}$$

We begin by observing that if α_2 is set equal to λ_2 in the above expression, then the labor force component assumes the shape of a well-known extreme value distribution used in the study of flood flows (Gumbel 1941, Kimball 1946). In such a case $x_h = \mu_2$ and the function $f_2(x)$ achieves its maximum y_h at that point. To analyze the more general case where $\alpha_2 \neq \lambda_2$, we may derive analytical expressions for both of these variables by differentiating eq. (4) with respect to x, setting the result equal to zero, and then solving to find

$$x_h = \mu_2 - (1/\lambda_2) \ln(\alpha_2/\lambda_2) \tag{5}$$

an expression that does not involve a_2, and

$$y_h = a_2(\alpha_2/\lambda_2)^{\alpha_2/\lambda_2} \exp(-\alpha_2/\lambda_2) \tag{6}$$

an expression that does not involve μ_2.

Note that if $\lambda_2 > \alpha_2$, which is almost always the case, then $x_h > \mu_2$. And observe that if $\alpha_2 = \lambda_2$, then the above two equations simplify to

$$x_h = \mu_2$$

and

$$y_h = a_2/e$$

Since μ_2 affects x_h only as a displacement, we may focus on the variation of x_h as a function of α_2 and λ_2. A plot of x_h against α_2, for a fixed λ_2, shows that increases in α_2 lead to decreases in x_h. Analogously, increases in λ_2, for a fixed α_2, produce increases in x_h but at a rate that decreases rapidly as the latter variable approaches its asymptote.

The behavior of y_h is independent of μ_2 and varies proportionately with a_2. Hence its variation also depends fundamentally only on the two variables α_2 and λ_2. A plot of y_h against α_2, for a fixed λ_2, gives rise to a U-shaped curve that reaches its minimum at $\alpha_2 = \lambda_2$. Increasing λ_2 widens the shape of the U.

The influence of α_2 and λ_2 on the labor force component may be assessed by examining the proportional rate of change of the function $f_2(x)$:

$$\frac{f_2'(x)}{f_2(x)} = -\alpha_2 + \lambda_2 \exp[-\lambda_2(x - \mu_2)] \tag{7}$$

Equation (7) defines this rate of change as the sum of two components: $-\alpha_2$ and the exponential $\lambda_2 \exp[-\lambda_2(x - \mu_2)]$. To demonstrate how the actual rates of ascent and descent are related to λ_2 and α_2 we may take, for example, a typical set of parameter values such as $\alpha_2 = 0.1$, $\lambda_2 = 0.4$, and $\mu_2 = 20$ and then proceed to calculate the quantities presented in Table 13. The calculations indicate that, at ages above 30, the actual rate of descent is almost identical to $-\alpha_2$. The actual rates of ascent are very different from the λ_2 value, except for ages close to $x = \mu_2$.*

TABLE 13 Impacts of λ_2 and α_2 on the actual rates of ascent and descent of the labor force component: $\lambda_2 = 0.4, \alpha_2 = 0.1$, and $\mu_2 = 20$.

Range of age	Age (x)	Actual rates of ascent and descent: $g(x) = \lambda_2 \exp[-\lambda_2(x - \mu_2)]$	$-\alpha_2 + g(x)$
In this range the impact of α_2 can be ignored	0	1192	1192
	5	161	161
	10	22	22
	15	3	3
	16	1.98	1.88
	17	1.33	1.23
	18	0.89	0.79
	19	0.60	0.50
$x = \mu_2$ →	20	0.40	0.30
	21	0.27	0.17
	22	0.18	0.08
	23	0.12	0.02
x_{max} →	24	0.08	−0.02
	25	0.05	−0.05
In this range the impact of λ_2 can be ignored	30	0.007	−0.093
	35	0.001	−0.100

*We are grateful to Kao-Lee Liaw for suggesting the examination of eq. (7) and for pointing out that the parameters λ_2 and α_2 are not truly rates of ascent and descent, respectively.

The introduction of the pre-labor force component into the profile generally moves x_h to a slightly younger age and raises y_h by about $a_1 \exp(-\alpha_1 x_h)$, usually a negligible quantity. The addition of the constant term c, of course, affects only y_h, raising it by the amount of the constant. Thus the migration rate at age x_h may be expressed as

$$M(x_h) \approx a_1 \exp(-\alpha_1 x_h) + y_h + c$$

A variable that interrelates the pre-labor force and labor force components is the parental shift A. To simplify our analysis of its dependence on the fundamental parameters, it is convenient to assume that α_1 and α_2 are approximately equal. In such instances, for ages immediately following the high peak x_h, the labor force component of the model migration schedule is closely approximated by the function $a_2 \exp[-\alpha_2(x_2 - \mu_2)]$. Recalling that the pre-labor force curve is given by $a_1 \exp(-\alpha_2 x_1)$ when $\alpha_1 = \alpha_2$, we may equate the two functions to solve for the difference in ages that we have called the parental shift:

$$A = x_2 - x_1 = \mu_2 + (1/\alpha_2) \ln (1/\delta_{12}) \qquad (8)$$

This equation shows that the parental shift will increase with increasing values of μ_2 and will decrease with increasing values of α_2 and δ_{12}. Table 14 compares the values of this analytically defined "theoretical" parental shift with the corresponding observed parental shifts presented earlier in Table 1 for Swedish males and females. The two definitions appear to produce similar numerical values, but the analytical definition has the advantage of being simpler to calculate and analyze.

Consider the rural-to-urban migration age profile defined by the parameters in Table 15. In this profile the values of α_2 and λ_2 are almost equal, making it a suitable illustration of several points raised in the above discussion.

First, calculating x_h with eq. (5) gives

$$x_h = 21.10 - (1/0.270) \ln (0.237/0.270) = 21.58$$

as against $x_h = 21.59$ set out in Table 15. Deriving y_h using eq. (6) gives

$$y_h = 0.187(0.878)^{0.878} \exp(-0.878) = 0.069$$

where $\alpha_2/\lambda_2 = 0.237/0.270 = 0.878$. Thus $M(21.59)$ is approximately equal to $y_h + c = 0.069 + 0.004 = 0.073$. The value given by the model migration schedule equation is also 0.073.

Since $\alpha_1 \neq \alpha_2$, we cannot adequately test the accuracy of eq. (8) as an estimator of A. Nevertheless, it can be used to help account for the unusually large value of the parental shift. Substituting the values for μ_2, α_2, and δ_{12} into eq. (8), we find

$$A = 21.10 + (1/0.237) \ln (1/0.011)$$

$$= 21.10 + 4.51/0.237 = 40.13$$

And although this is an underestimate of 45.13, it does suggest that the principal cause for the unusually high value of A is the unusually low value of δ_{12}. If this latter parameter

TABLE 14 Observed and theoretical values of the parental shift: Sweden, 8 regions, 1974.

Parental shift	Regions of Sweden							
	1. Stockholm	2. East Middle	3. South Middle	4. South	5. West	6. North Middle	7. Lower North	8. Upper North
Observed,[a] males	27.87	29.99	29.93	29.90	29.57	29.92	30.15	31.61
Theoretical, males	25.14	29.24	30.01	29.65	28.97	29.43	26.61	29.89
Observed,[a] females	25.49	27.32	27.27	27.87	27.42	27.01	26.94	28.30
Theoretical, females	24.68	26.85	28.16	28.91	27.51	28.54	28.19	28.95

[a]Source: Table 1.

TABLE 15 Parameters and variables defining observed total (males plus females) model migration schedules for urban-to-rural and rural-to-urban flows: the Soviet Union, 1974.

Parameters and variables[a]	Urban-to-rural	Rural-to-urban
GMR	0.74	3.41
a_1	0.005	0.002
α_1	0.313	0.431
a_2	0.127	0.187
μ_2	19.26	21.10
α_2	0.177	0.237
λ_2	0.286	0.270
c	0.005	0.004
$\bar{n}$	33.66	31.24
%(0–14)	8.63	5.59
%(15–64)	78.30	84.60
%(65+)	13.07	9.81
δ_{1c}	0.977	0.548
δ_{12}	0.038	0.011
β_{12}	1.77	1.82
σ_2	1.61	1.14
x_l	11.09	11.38
x_h	20.94	21.59
X	9.85	10.21
A	42.30	45.13
B	0.045	0.063

[a]A list of definitions for the parameters and variables appears in Appendix B.

had the value found for Stockholm's males, for example, the parental shift would exhibit the much lower value of 22.52.

4 ESTIMATED MODEL MIGRATION SCHEDULES

An estimated model schedule is a collection of age-specific rates derived from patterns observed in various populations other than the one being studied plus some incomplete data on the population under examination. The justification for such an approach is that age profiles of fertility, mortality, and geographical mobility vary within predetermined limits for most human populations. Birth, death, and migration rates for one age group are highly correlated with the corresponding rates for other age groups, and expressions of such interrelationships form the basis of model schedule construction. The use of these regularities to develop hypothetical schedules that are deemed to be close approximations of the unobserved schedules of populations lacking accurate vital and mobility registration statistics has been a rapidly growing area of contemporary demographic research.

4.1 Introduction: Alternative Perspectives

The earliest efforts in the development of model schedules were based on only one parameter and hence had very little flexibility (United Nations 1955). Demographers soon discovered that variations in the mortality and fertility regimes of different populations required more complex formulations. In mortality studies greater flexibility was introduced by providing families of schedules (Coale and Demeny 1966) or by enlarging the number of parameters used to describe the age pattern (Brass 1975). The latter strategy was also adopted in the creation of improved model fertility schedules and was augmented by the use of analytical descriptions of age profiles (Coale and Trussell 1974).

Since the age patterns of migration normally exhibit a greater degree of variability across regions than do mortality and fertility schedules, it is to be expected that the development of an adequate set of model migration schedules will require a greater number both of families and of parameters. Although many alternative methods could be devised to summarize regularities in the form of families of model schedules defined by several parameters, three have received the widest popularity and dissemination:

1. The regression approach of the Coale–Demeny model life tables (Coale and Demeny 1966)
2. The logit system of Brass (Brass 1971)
3. The double exponential graduation of Coale, McNeil, and Trussell (Coale 1977, Coale and McNeil 1972, Coale and Trussell 1974)

The regression approach embodies a *correlational* perspective that associates rates at different ages to an index of level, where the particular associations may differ from one "family" of schedules to another. For example, in the Coale–Demeny model life tables, the index of level is the expectation of remaining life at age 10, and a different set of regression equations is established for each of four "regions" of the world. Each of the four regions (North, South, East, and West) defines a collection of similar mortality schedules that are more uniform in pattern than the totality of observed life tables.

Brass's logit system reflects a *relational* perspective in which rates at different ages are given by a standard schedule whose shape and level may be suitably modified to be appropriate for a particular population.

The Coale–Trussell model fertility schedules are relational in perspective (using a Swedish standard first-marriage schedule), but they also introduce an analytic description of the age profile by adopting a double exponential curve that defines the shape of the age-specific first-marriage function.

In this study we mix the above three approaches to define two alternative perspectives for estimating model migration schedules in situations where only inadequate or defective data on internal (origin–destination) migration flows are available. Both perspectives rely on the analytic (double plus single exponential) graduation defined by the basic model migration schedule set out earlier in this study. Both ultimately depend on the availability of some limited data to obtain the appropriate model schedule, for example, at least two age-specific rates, such as $M(0-4)$ and $M(20-24)$, and informed guesses regarding the values of a few key variables, such as the low and high points of the schedule. They differ only in the method by which a schedule is identified as being appropriate for a particular population.

The first perspective, the regression approach, associates variations in the parameters and derived variables of the model schedule to each other and then to age-specific migration rates. The second, the logit approach, embodies different relationships between the model schedule parameters in several standard schedules and then associates the logits of the migration rates in a standard to those of the population in question.

4.2 The Correlational Perspective: The Regression Migration System

A straightforward way of obtaining an estimated model migration schedule from limited observed data is to associate such data with the basic model schedule's parameters by means of regression equations. For example, given estimates of the migration rates of infants and young adults, $M(0-4)$ and $M(20-24)$ say, we may use equations of the form

$$Q_i = b_0\,[M(0-4)]^{b_1}\,[M(20-24)]^{b_2}$$

to estimate the set of parameters Q_i that define the model schedule. The parameters of the fitted model schedules are not independent of each other, however. Higher than average values of λ_2, for example, tend to be associated with lower than average values of a_1. The incorporation of such dependencies into the regression approach would surely improve the accuracy and consistency of the estimation procedure. An examination of empirical associations among model schedule parameters and variables, therefore, is a necessary first step.

Regularities in the covariations of the model schedule's parameters suggest a strategy of model schedule construction that builds on regression equations embodying these covariations. Given the values for δ_{12}, x_l, and x_h, for example, one can proceed to derive μ_2, λ_2, σ_2, and β_{12}. Since $\sigma_2 = \lambda_2/\alpha_2$ we obtain, at the same time, an estimate for α_2, which we then can use to find a_2. With a_2 established, a_1 may be obtained by drawing on the definitional equation $\delta_{12} = a_1/a_2$, and α_1 may be found with the similar equation $\beta_{12} = \alpha_1/\alpha_2$. An initial value for c is obtained by setting $c = a_1/\delta_{1c}$, where δ_{1c} is estimated by regressing it on δ_{12}, and a_1, a_2, and c are scaled to give a *GMR* of unity.

Conceptually, this approach to model schedule construction begins with the labor force component and then appends to it the pre-labor force part of the curve. The value given for δ_{12} reflects the relative weights of these two components, with low values defining a labor dominant curve and high values pointing to a family dominant curve. (The behavior of the post-labor force curve is assumed here to be treated exogenously.)

We begin the calculations with μ_2 to establish the location of the curve on the age axis; is it an early or late peaking curve? Next, we turn to the determination of its two slope parameters λ_2 and α_2 by resolving whether or not it is a labor symmetric curve. Values of σ_2 between 1 and 2 generally characterize a labor symmetric curve; higher values describe an asymmetric age profile. The regression of a_2 on α_2 produces the fourth parameter needed to define the labor force component. With values for μ_2, λ_2, α_2, and a_2 the construction procedure turns to the estimation of the pre-labor force curve, which is defined by the two parameters α_1 and a_1. Its relative share of the total unit area under the model migration schedule is set by the value given to δ_{12}. The retirement peak and the upward slope are introduced exogenously by setting their parameters equal to those of the "observed" model migration schedule.

The collection of regression equations given in Table 16 exemplifies a regression system that may be defined to represent the "child dependency" set, inasmuch as their central independent variable δ_{12} is the index of child dependency. It is also possible to replace this independent variable with others, such as σ_2 or β_{12} for example, to create a "labor asymmetry" or a "parental-shift regularity" set. The regression coefficients were obtained using the age-specific interregional migration schedules (scaled to unit *GMR*) of Sweden, the United Kingdom, and Japan. Of the three variants, the child dependency set gave the best fits in about half of the female schedules tested, whereas the parental-shift

TABLE 16 A basic set of regression equations.

Dependent variables		Regression coefficients of independent variables				
		Intercept	δ_{12}	x_l	x_h	α_2
μ_2	(males)	−3.26	3.28	−0.67	1.39	
	(females)	−7.69	−2.14	−0.53	1.63	
λ_2	(males)	1.31	0.15	0.08	−0.09	
	(females)	1.19	0.13	0.08	−0.09	
σ_2	(males)	16.43	5.59	0.89	−1.17	
	(females)	10.97	6.05	0.63	−0.85	
β_{12}	(males)	1.90	1.33	−0.03	−0.04	
	(females)	1.82	1.42	−0.04	−0.04	
a_2	(males)	0.03				0.30
	(females)	0.04				0.25
δ_{1c}	(males)	9.41	13.83			
	(females)	0.19	26.43			

regularity set was overwhelmingly the best fitting variant for the male schedules (see Rogers and Castro 1981).

To use the basic regression equations presented in Table 16, one first needs to obtain estimates of δ_{12}, x_l, and x_h. Values for these three variables may be selected to reflect informed guesses, historical data, or empirical regularities between such model schedule variables and observed migration data. For example, suppose that a fertility survey has produced a crude estimate of the ratio of infant to parent migration rates: $M = M(0\text{–}4)/M(20\text{–}24)$, say. A linear association between δ_{12} and this M ratio, with the regression equation forced through the origin, gives

$$_F\hat{\delta}_{12} = 0.6M$$

for females, and

$$_M\hat{\delta}_{12} = 0.7M$$

for males.

Figure 10 illustrates examples of the goodness-of-fit provided by the estimated schedules to the observed model migration data. Two sets of estimated schedules are shown: those obtained with the observed index of child dependency (δ_{12}) and those found with the estimated index ($\hat{\delta}_{12}$), both calculated using the above regressions. In each case x_l and x_h were set equal to the values given by the observed model migration schedules.

4.3 The Relational Perspective: The Logit Migration System

Among the most popular methods for estimating mortality from inadequate or defective data, is the so-called logit system developed by William Brass about twenty years ago

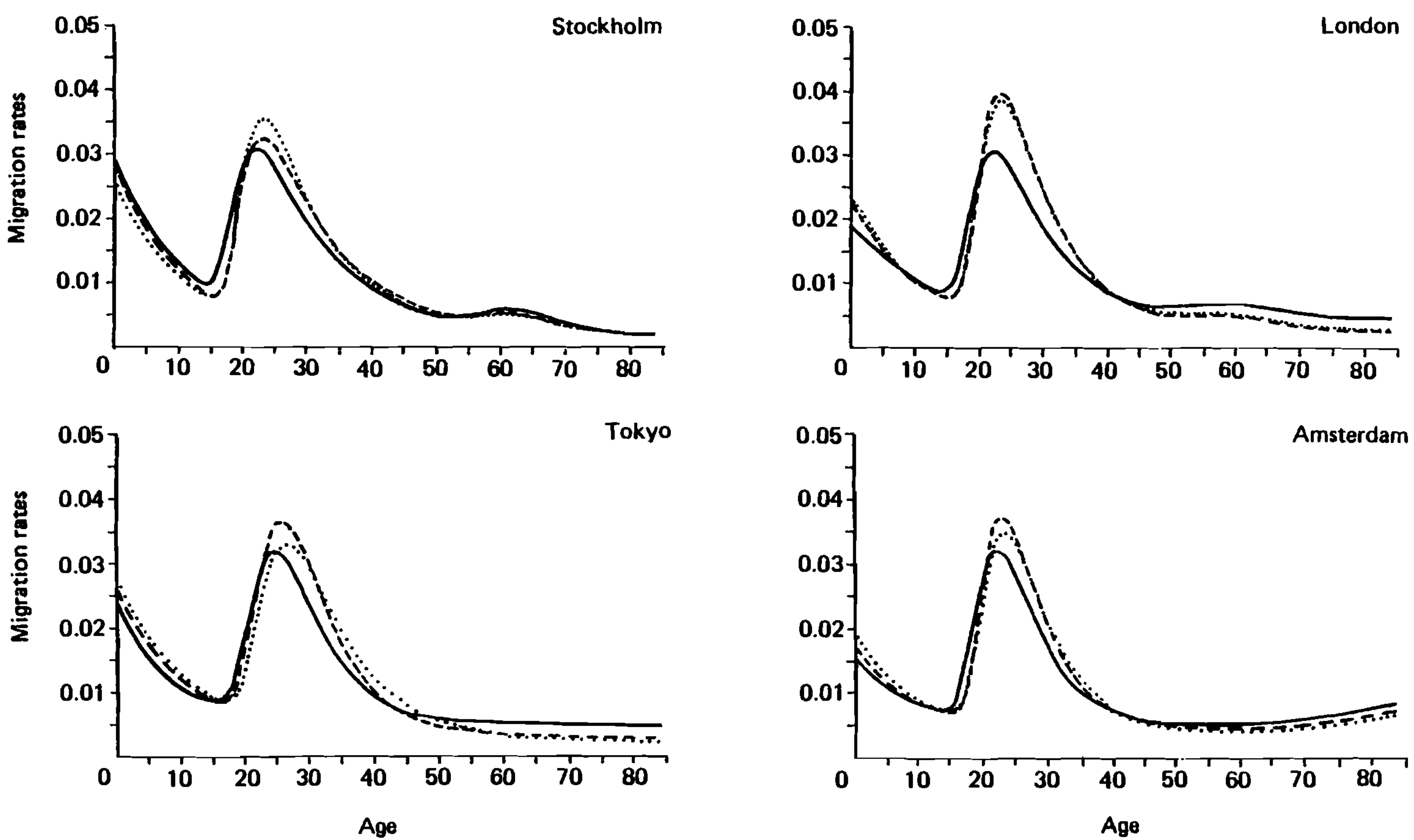


FIGURE 10 The fits of the correlational approach when using δ_{12} from the model migration schedule (– – –) and $\hat{\delta}_{12}$ from the observed M ratio (· · ·) compared with the observed (———) data for the female populations of Stockholm, London, Tokyo, and Amsterdam.

and now widely applied by demographers all over the world (Brass 1971, Brass and Coale 1968, Carrier and Hobcraft 1971, Hill and Trussell 1977, Zaba 1979). The logit approach to model schedules is founded on the assumption that different mortality schedules can be related to each other by a linear transformation of the logits of their respective survivorship probabilities. That is, given an observed series of survivorship probabilities $l(x)$ for ages $x = 1,2,\ldots,\omega$, it is possible to associate these observed series with a "standard" series $l_s(x)$ by means of the linear relationship

$$\text{logit}\ [1 - l(x)] = \gamma + \rho\ \text{logit}\ [1 - l_s(x)]$$

where, say,

$$\text{logit}\ [y(x)] = (1/2) \ln\ [y(x)/(1 - y(x))] = Y(x) \qquad 0 < y(x) < 1$$

or

$$Y(x) = \gamma + \rho\ Y_s(x)$$

The inverse of this function is

$$l(x) = 1/\{1 + \exp[2Y(x)]\}$$

The principal result of this mathematical transformation of the nonlinear $l(x)$ function is a more nearly linear function in x, with a range of minus and plus infinity rather than unity and zero.

Given a standard schedule, such as the set of standard logits, $Y_s(x)$, proposed by Brass, a life table can be created by selecting appropriate values for γ and ρ. In the Brass system γ reflects the level of mortality and ρ defines the relationship between child and adult mortality. The closer γ is to zero and ρ to unity, the more the estimated life table is like the standard.

The logit perspective can be readily applied to migration schedules. Let ${}_uM(x)$ denote the age-specific migration rates of a schedule scaled to a unit *GMR*, and let ${}_uM_s(x)$ denote the corresponding standard schedule. Taking logits of both sets of rates gives the logit migration system

$${}_uY(x) = \gamma + \rho\ {}_uY_s(x)$$

and

$${}_uM(x) = \frac{1}{1 + \exp\{-2[\gamma + \rho\ {}_uY_s(x)]\}}$$

where, for example,

$$\text{logit}\ [{}_uM_s(x)] = {}_uY_s(x) = (1/2) \ln {}_uM_s(x)/[1 - {}_uM_s(x)]$$

The selection of a particular migration schedule as a standard reflects the belief that it is broadly representative of the age pattern of migration in the multiregional population system under consideration. (Our standard schedules will always have a unit *GMR*; hence the left subscript on ${}_{\mathrm{u}}Y_s(x)$ will be dropped.) To illustrate a number of calculations carried out with several sets of multiregional data, we shall adopt the national age profile as the standard in each case and strive to estimate regional outmigration age profiles by relating them to the national one. Specifically, given an $m \times m$ table of interregional migration flows for any age x, we divide each origin–destination-specific flow $O_{ij}(x)$ by the population in the origin region $K_i(x)$ to define the associated age-specific migration rate $M_{ij}(x)$. Summing these over all origins and destinations gives the corresponding national rate $M..(x)$, and scaling all schedules to unit *GMR* gives ${}_{\mathrm{u}}M_{ij}(x)$ and ${}_{\mathrm{u}}M..(x)$, respectively.

Figure 11 presents national male standards for Sweden, the United Kingdom, Japan, and the Netherlands. (We shall deal only with graduated fits inasmuch as all of our non-Swedish data are for five-year age intervals and therefore need to be graduated first in order to provide single-year profiles by means of interpolation.) The differences in age profiles are marked. Only the Swedish and the United Kingdom standards exhibit a retirement peak. Japan's profile is described without such a peak because the age distribution of migrants given by the census data ends with the open interval of 65 years and over. The data for the Netherlands, on the other hand, show a definite upward slope at the post-labor force ages and therefore have been graduated with the 9-parameter model schedule with an upward slope.

Regressing the logits of the age-specific outmigration rates of each region on those of its national standard (the *GMR*s of both first being scaled to unity) gives estimated values for γ and ρ. Reversing the procedure and combining selected values of γ and ρ with a national standard of logit values, identifies the following important regularity: *whenever* $\gamma = 2(\rho - 1)$ *then the GMR of the estimated model schedule is approximately unity* (Rogers and Castro 1981). Linear regressions of the form

$$\gamma = d_0 + d_1 \rho$$

fitted to our data for Sweden, the United Kingdom, Japan, and the Netherlands, consistently produce estimates for d_0 and d_1 that are approximately equal to 2 in magnitude and that differ only in sign, i.e., $\hat{d}_0 = -2$, and $\hat{d}_1 = +2$. Thus

$$\gamma = -2 + 2\rho = 2(\rho - 1)$$

Differences in the national standard schedules illustrated in Figure 11 suggest that a single standard schedule may be a more restrictive assumption in migration analysis than in mortality studies. It therefore may be necessary to follow the Coale–Demeny strategy of developing families of appropriate schedules (Coale and Demeny 1966).

The comparative analysis of national and interregional migration patterns carried out in section 3 identified at least three distinct families of age profiles. First, there was the 11-parameter *basic model migration schedule* with a retirement peak that adequately described a number of interregional flows, for example, the age profiles of outmigrants leaving capital regions such as Stockholm and London. The elimination of the retirement peak gave rise to the 7-parameter *reduced form* of this basic schedule, a form that was used to describe a large number of labor dominant profiles and the age patterns of migration schedules with a single open-ended age interval for the post-labor force population,

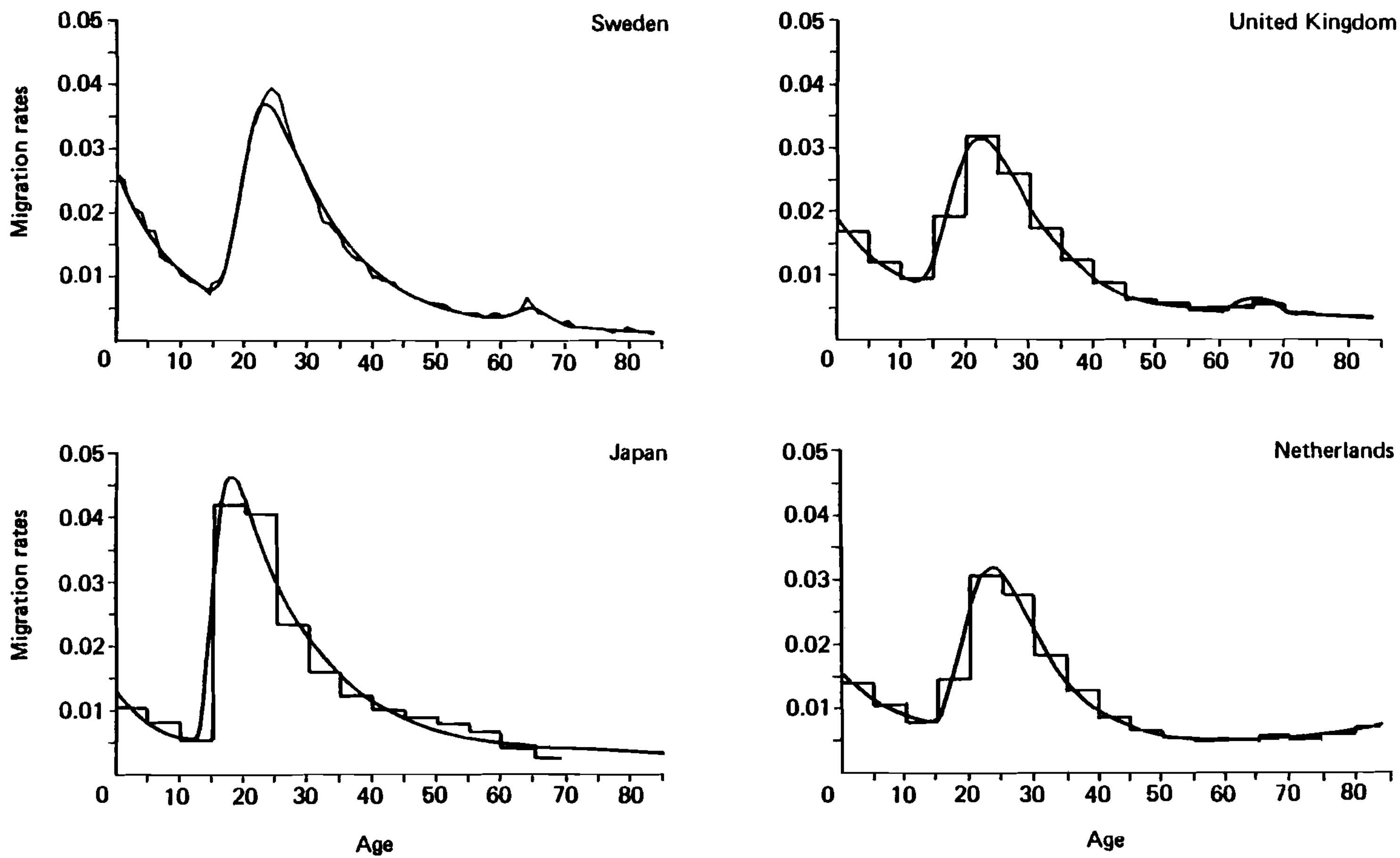


FIGURE 11 Observed (jagged line) and model (smooth line) national male standard migration schedules: Sweden, the United Kingdom, Japan, the Netherlands.

for example, Japan's migration schedules. Finally, the existence of a monotonically rising tail in migration schedules such as those exhibited by the Dutch data led to the definition of a third profile: the 9-parameter *model migration schedule with an upward slope.*

Within each family of schedules, a number of key parameters or variables may be put forward in order to further classify different categories of migration profiles. For example, in section 3 we identified the special importance of the following aspects of shape and location along the age axis:

1. Peaking: early peaking versus late peaking (μ_2)
2. Dominance: child dependence versus labor dominance (δ_{12})
3. Asymmetry: labor symmetry versus labor asymmetry (σ_2)
4. Regularity: parental-shift regularity versus parental-shift irregularity (β_{12})

These fundamental families and four key parameters give rise to a large variety of standard schedules. For example, even if the four key parameters are restricted to only dichotomous values, one already needs $2^4 = 16$ standard schedules. If, in addition, the sexes are to be differentiated, then 32 standard schedules are a minimum. A large number of standard schedules would make the logit approach a less desirable alternative. Hence we shall examine the feasibility of adopting only a single standard for both sexes and assume that the shape of the post-labor force part of the schedule may be determined exogenously. In tests of our logit migration system, therefore, we shall always set the post-labor force retirement peak or upward slope equal to observed model schedule values.

The similarity of the male and female median parameter values set out in Tables 10 and 11 (for Sweden, the United Kingdom, and Japan), suggests that one could use the average of the values for the two sexes to define a unisexual standard. A rough rounding of these averages would simplify matters even more. Table 17 presents the simplified basic standard parameters obtained in this way. The values of a_1, a_2, and c are initial values only and need to be scaled proportionately to ensure a unit *GMR*. Figure 12 illustrates the age profile of this simplified basic standard migration schedule.

TABLE 17 The simplified basic standard migration schedule.

Fundamental parameters	Fundamental ratios
$a_1 = 0.02$	$\delta_{12} = 1/3$
$\alpha_1 = 0.10$	$\sigma_2 = 4$
$a_2 = 0.06$	$\beta_{12} = 1$
$\mu_2 = 20$	$\delta_{1c} = 6$
$\alpha_2 = 0.10$	
$\lambda_2 = 0.40$	
$c = 0.003$	

We have noted before that when $\gamma = 0$ and $\rho = 1$, the estimated model schedule is identical to the standard. Moreover since the *GMR* of the standard is always unity, values of γ and ρ that satisfy the equality $\gamma = 2(\rho - 1)$ guarantee a *GMR* of unity for the estimated schedule. What are the effects of other combinations of values for these two parameters?

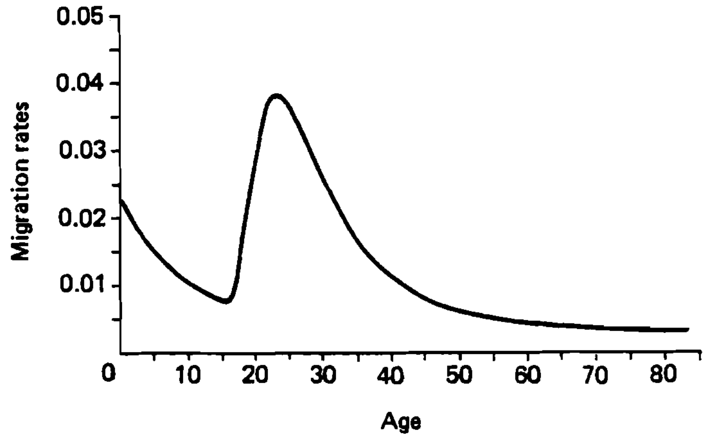


FIGURE 12 Simplified basic standard migration schedule.

Figure 13 illustrates how the simplified basic standard schedule is transformed when γ and ρ are assigned particular pairs of values. Figure 13(a) shows that fixing $\gamma = 0$ and increasing ρ from 0.75 to 1.25 lowers the schedule, giving migration rates that are smaller in value than those of the standard. On the other hand, fixing $\rho = 0.75$, and increasing γ from -1 to 0 raises the schedule, according to Figure 13(b). Finally, fixing $GMR = 1$ by selecting values of γ and ρ that satisfy the equality $\gamma = 2(\rho - 1)$ shows that as γ and ρ both increase, so does the degree of labor dominance exhibited by the estimated schedule. For example, moving from an estimated schedule with $\gamma = -0.5$ and $\rho = 0.75$ to one with $\gamma = 0.5$ and $\rho = 1.25$ does not alter the area under the curve ($GMR = 1$), but it does increase its labor dominance (Figure 13(c)).

Given a standard schedule and a few observed rates, such as $M(0\text{–}4)$ and $M(20\text{–}24)$, for example, how can one find estimates for γ and ρ, and with those estimates go on to obtain the entire estimated schedule?

First, taking logits of the two observed migration rates gives $Y(0\text{–}4)$ and $Y(20\text{–}24)$ and associating these two logits with the pair of corresponding logits for the standard gives

$$Y(0\text{–}4) = \gamma + \rho\, Y_s(0\text{–}4)$$

$$Y(20\text{–}24) = \gamma + \rho\, Y_s(20\text{–}24)$$

Solving these two equations in two unknowns gives crude estimates for γ and ρ, and applying them to the standard schedule's full set of logits results in a set of logits for the estimated schedule. From these one can obtain the migration rates, as shown earlier. Tests of such a procedure with the migration data for Sweden, the United Kingdom, Japan, and the Netherlands, however, indicate that the method is very erratic in the goodness-of-fits that it produces and, therefore, more refined procedures are necessary. Such procedures (for the case of mortality) are described in the literature on the Brass logit system (for example, in Brass 1975, Carrier and Goh 1972).

A reasonable first approximation to an improved estimation method for the case of migration is suggested by the regression approach described in subsection 4.2. Imagine a

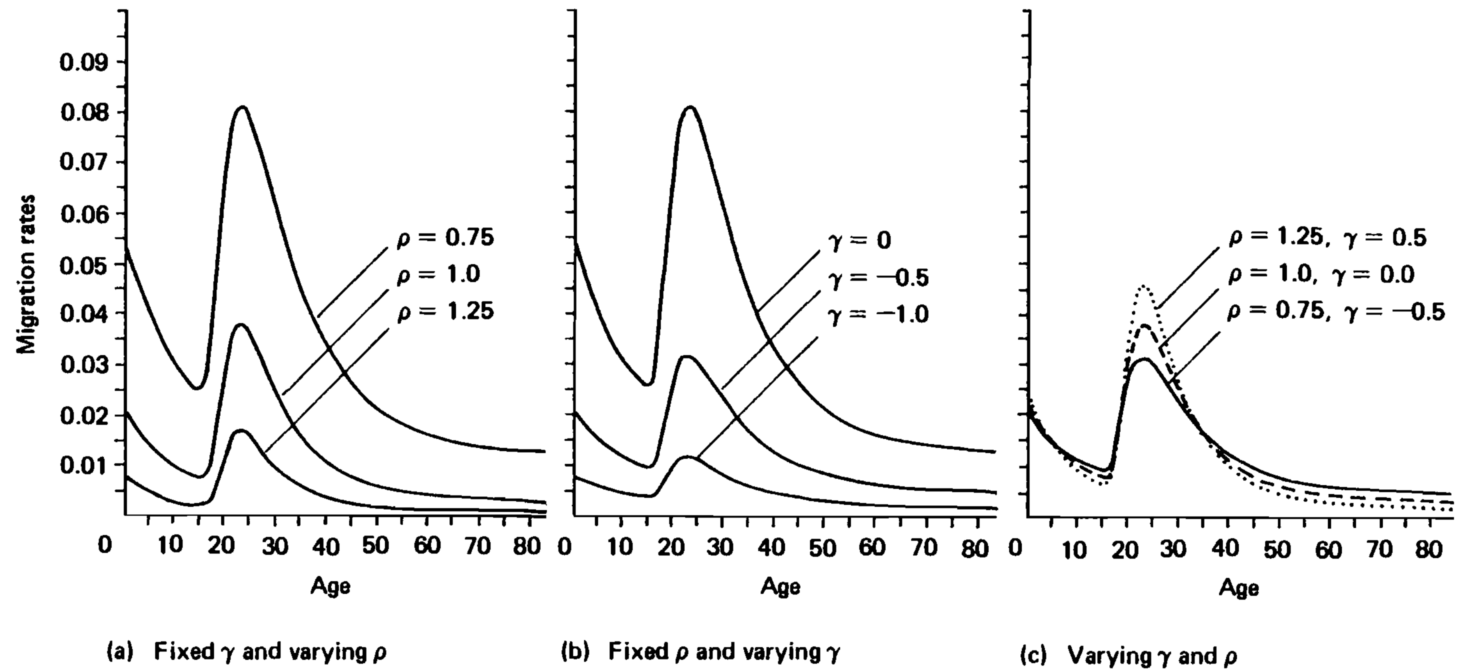


FIGURE 13 Sensitivity of the logit model migration schedule to variations in γ and ρ: simplified basic standard migration schedule.

regression of ρ on the M ratio, $M(0-4)/M(20-24)$. Starting with the simplified basic standard migration schedule and varying ρ within the range of observed values, one may obtain a corresponding set of M ratios. Associating ρ and the M ratio in this way, one may proceed further and use the relational equation to estimate $\hat{\gamma}$ from $\hat{\rho}$:

$$\hat{\gamma} = 2(\hat{\rho} - 1)$$

A further simplification can be made by forcing the regression line to pass through the origin. Since the resulting regression coefficient has a negative sign and the intercept exhibits roughly the same absolute value, but with a positive sign, the regression equations take on the form

$$\hat{\rho} = 2.1(1 - M)$$

where $M = M(0-4)/M(20-24)$.

Given a standard schedule and estimates for γ and ρ, one can proceed to compute the associated estimated model migration schedule. Figure 14 illustrates representative examples of the goodness-of-fit obtained using this procedure. Two estimated schedules are illustrated with each observed model migration schedule: those calculated with the interpolated 85 single-year-of-age observations and the resulting least-squares estimates of γ and ρ, and those computed using the above regression equations of ρ on the M ratio. Although the fits are moderately successful, it is clear that further study of this problem is necessary.

5 CONCLUSION

This report began with the observation that empirical regularities characterize observed migration schedules in ways that are no less important than the corresponding well-established regularities in observed fertility or mortality schedules. Section 2 was devoted to defining mathematically such regularities in observed migration schedules in order to exploit the notational, computational, and analytical advantages that such a formulation provides. Section 3 reported on the results of an examination of over 500 migration schedules that underscored the broad generality of the model migration schedule proposed and helped to identify a number of families of such schedules.

Regularities in age profiles lead naturally to the development of hypothetical model migration schedules that might be suitable for studies of populations with inadequate or defective data. Drawing on techniques used in the corresponding literature in fertility and mortality, section 4 develops procedures for inferring migration patterns in the absence of accurate migration data.

Of what use, then, is the model migration schedule defined in this study? What are some of its concrete practical applications?

The model migration schedule may be used to *graduate* observed data, thereby smoothing out irregularities and ascribing to the data summary measures that can be used for comparative analysis. It may be used to *interpolate* to single years of age, observed migration schedules that are reported for wider age intervals. Assessments of the *reliability*

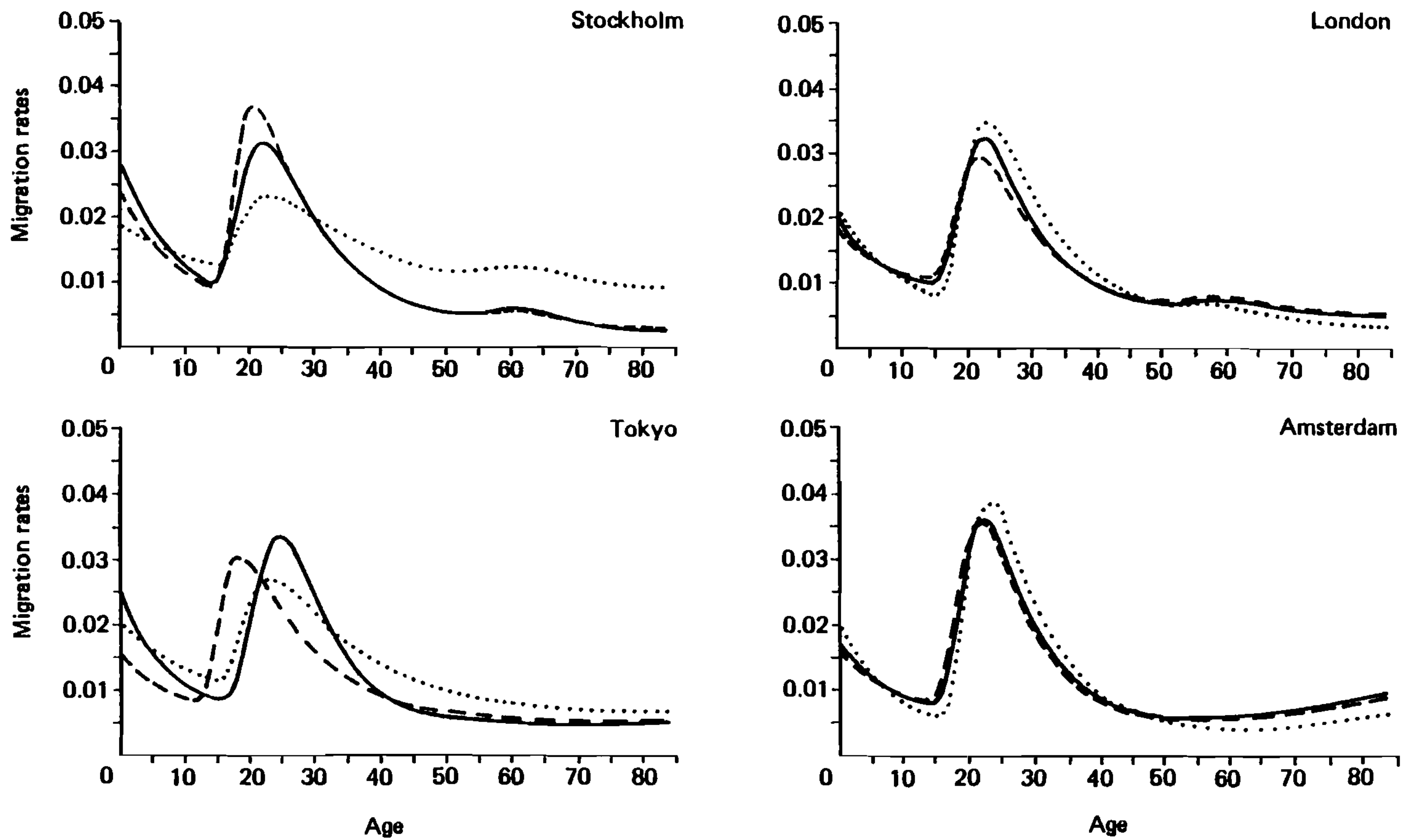


FIGURE 14 The fits of the relational approach when using the estimated parameters from 85 observations (– – –) and the $\hat{\rho}$ parameter from the observed M ratio (· · · ·) compared with the observed (———) data for the female populations of Stockholm, London, Tokyo, and Amsterdam.

of empirical migration data and indications of appropriate strategies for their *correction* are aided by the availability of standard families of migration schedules. Finally, such schedules also may be used to help resolve problems caused by *missing data.*

The analysis of national migration age patterns reported in this study seeks to demonstrate the utility of examining the regularities in age profile exhibited by empirical schedules of interregional migration. Although data limitations have restricted some of the findings to conjectures, a modest start has been made. It is hoped that the results reported here will induce others to devote more attention to this topic.

ACKNOWLEDGMENTS

The authors are grateful to the many national collaborating scholars who have participated in IIASA's Comparative Migration and Settlement Study. This report could not have been written without the data bank produced by their collective efforts. Thanks also go to Richard Raquillet for his contributions to the early phases of this study and to Walter Kogler for his untiring efforts on our behalf in front of a console in IIASA's computer center. Kao-Lee Liaw, Philip Rees, Warren Sanderson, and an anonymous reviewer offered several useful comments on an earlier draft.

REFERENCES

Andersson, A. and I. Holmberg (1980) Migration and Settlement: 3. Sweden. RR-80-5. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Bard, Y. (1974) Nonlinear Parameter Estimation. New York: Academic Press.

Benson, M. (1979) Parameter Fitting in Dynamic Models. Ecological Modelling 6:97–115.

Brass, W. (1971) On the scale of mortality. Pages 69–110 in Biological Aspects of Demography, edited by W. Brass. London: Taylor and Francis, Ltd.

Brass, W. (1975) Methods for Estimating Fertility and Mortality from Limited and Defective Data. Chapel Hill, North Carolina: Laboratories for Population Statistics, University of North Carolina at Chapel Hill.

Brass, W. and A. Coale (1968) Methods of analysis and estimation. Pages 88–150 in The Demography of Tropical Africa, edited by W. Brass et al. Princeton, New Jersey: Princeton University Press.

Brown, K.M. and J.E. Dennis (1972) Derivative free analogues of the Levenberg–Marquardt and Gauss algorithms for nonlinear least squares approximations. Numerische Mathematik 18:289–297.

Carrier, N. and J. Hobcraft (1971) Demographic Estimation for Developing Societies. London: Population Investigation Committee, London School of Economics.

Carrier, N. and T. Goh (1972) The validation of Brass's model life table system. Population Studies 26 (1):29–51.

Coale, A. (1977) The development of new models of nuptiality and fertility. Population, special issue: 131–154.

Coale, A. and P. Demeny (1966) Regional Model Life Tables and Stable Populations. Princeton, New Jersey: Princeton University Press.

Coale, A.J. and D.R. McNeil (1972) The distribution by age of the frequency of first marriage in a female cohort. Journal of the American Statistical Association 67:743–749.

Coale, A. and J. Trussell (1974) Model fertility schedules: variations in the age structure of childbearing in human populations. Population Index 40(2):185–206.

Fiacco, A. and G. McCormick (1968) Nonlinear Programming: Sequential Unconstrained Minimization Techniques. New York: Wiley.

Gumbel, E.J. (1941) The return period of flood flows. Annals of Mathematical Statistics 12:163–190.

Hill, K. and J. Trussell (1977) Further developments in indirect mortality estimation. Population Studies 31(2):313–334.

Kimball, B.F. (1946) Sufficient statistical estimation functions for the parameters of the distribution of maximum values. Annals of Mathematical Statistics 17:299–309.

Levenberg, K. (1944) A method for the solution of certain nonlinear problems in least squares. Quarterly of Applied Mathematics 2:164–168.

Marquardt, D.W. (1963) An algorithm for least-squares estimation of nonlinear parameters. SIAM, Journal of Numerical Analysis 11:431–441.

McNeil, D.R., T.J. Trussell, and J.C. Turner (1977) Spline interpolation of demographic data. Demography 14:245–252.

Rees, P.H. (1977) The measurement of migration, from census data and other sources. Environment and Planning A 1:247–260.

Rees, P.H. (1979) Migration and Settlement: 1. United Kingdom. RR-79-3. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Rogers, A. (1973a) The multiregional life table. The Journal of Mathematical Sociology 3:127–137.

Rogers, A. (1973b) The mathematics of multiregional demographic growth. Environment and Planning 5:3–29.

Rogers, A. (1975) Introduction to Multiregional Mathematical Demography. New York: Wiley.

Rogers, A. (1976a) The Comparative Migration and Settlement Study: A Summary of Workshop Proceedings and Conclusions. RM-76-01. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Rogers, A. (1976b) Two Methodological Notes on Spatial Population Dynamics in the Soviet Union. RM-76-48. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Rogers, A., editor (1978) Migration and settlement: selected essays. Environment and Planning A 10 (5):469–617.

Rogers, A. and L. Castro (1981) Model Schedules in Multistate Demographic Analysis: The Case of Migration. WP-81-22. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Rogers, A. and J. Ledent (1976) Increment–Decrement life tables: a comment. Demography 13:287–290.

Rogers, A., R. Raquillet, and L.J. Castro (1978) Model migration schedules and their applications. Environment and Planning A 10:475–502.

Rogers, A. and F. Willekens (1978) Migration and Settlement: Measurement and Analysis. RR-78-13. Laxenburg, Austria: International Institute for Applied Systems Analysis.

United Nations (1955) Age and Sex Patterns of Mortality: Model Life Tables for Underdeveloped Countries. New York: United Nations.

Willekens, F. and A. Rogers (1978) Spatial Population Analysis: Methods and Computer Programs. RR-78-18. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Zaba, B. (1979) The four-parameter logit life table system. Population Studies 33(1):79–100.

APPENDIX A

NONLINEAR PARAMETER ESTIMATION WITH MODEL MIGRATION SCHEDULES

This appendix briefly illustrates the mathematical programming procedure used to estimate the parameters of the model migration schedule. The nonlinear estimation problem may be defined as the search for the "best" parameter values for the function

$$\begin{aligned} M(x) = {} & a_1 \exp(-\alpha_1 x) \\ & + a_2 \exp\{-\alpha_2(x - \mu_2) - \exp[-\lambda_2(x - \mu_2)]\} \\ & + a_3 \exp\{-\alpha_3(x - \mu_3) - \exp[-\lambda_3(x - \mu_3)]\} \\ & + c \end{aligned} \tag{A1}$$

in the sense that a pre-defined objective function is minimized when the parameters take on these values.

This problem is the classical one of nonlinear parameter estimation in unconstrained optimization. All of the available methods start with a set of given initial conditions, or initial guesses of the parameter values, in the search for better estimates following specific convergence criteria. The iterative sequence ends after a finite number of iterations, and the solution is accepted as giving the best estimates for the parameters.

The problem of selecting an effective method has been usefully summarized by Bard (1974, p. 84) as follows:

> . . . no single method has emerged which is best for the solution of all nonlinear programming problems. One cannot even hope that a "best" method will ever be found, since problems vary so much in size and nature. For parameter estimation problems we must seek methods which are particularly suitable to the special nature of these problems which may be characterized as follows:
>
> 1. A relatively small number of unknowns, rarely exceeding a dozen or so.
> 2. A highly nonlinear (though continuous and differentiable) objective function, whose computation is often very time consuming.
> 3. A relatively small number (sometimes zero) of inequality constraints. Those are usually of a very simple nature, e.g., upper and lower bounds.
> 4. No equality constraints, except in the case of exact structural models (where, incidentally, the number of unknowns is large) . . .

For computational convenience, we have chosen the Marquardt method (Levenberg 1944, Marquardt 1963). This method seeks out a parameter vector P^* that minimizes the following objective function:

$$\phi(P) = f_P \tag{A2}$$

where f_P is the residual vector. For the case of a model schedule with a retirement peak, vector P has the following elements:

$$P^T = [a_1, \alpha_1, a_2, \alpha_2, \mu_2, \lambda_2, a_3, \alpha_3, \mu_3, \lambda_3, c] \tag{A3}$$

where T denotes transposition. The elements of the vector f_P can be computed by either of the following two expressions:

$$f_P(x) = [M(x) - \hat{M}_P(x)]^2 \tag{A4}$$

or

$$f_P(x) = [M(x) - \hat{M}_P(x)]^2/\hat{M}_P(x) \tag{A5}$$

where $M(x)$ is the observed value at age x and $\hat{M}_P(x)$ is the estimated value using eq. (A1) and a given vector P of parameter estimates.

By introducing eq. (A4) in the objective function set out in eq. (A2), the sum of squares is minimized; if, on the other hand, eq. (A5) is introduced instead, the chi-square statistic is minimized.

In matrix notation, the Levenberg–Marquardt method follows the iterative sequence

$$P_{q+1} = P_q - \{\mathbf{J}_q^T \mathbf{J}_q + \lambda_q \mathbf{D}_q\}^{-1} \mathbf{J}_q^T f_{P_q}$$

where λ is a non-negative parameter adjusted to ensure that at each iteration the function (A2) is reduced, $\mathbf{J}_q$ denotes the Jacobian matrix of $\phi(P)$ evaluated at the q iteration, and $\mathbf{D}$ is a diagonal matrix equal to the diagonal of $\mathbf{J}^T\mathbf{J}$.

The principal difficulty in nonlinear parameter estimation is that of convergence, and the method discussed here is no exception. The algorithm starts out by assuming some initial parameters, and then a new vector P is estimated according to the value of λ, which in turn is also modified following some gradient criteria. Once some given stopping values are achieved, vector P^* is assumed to be the optimum. In some cases, however, this P^* reflects local minima that may be improved with better initial conditions and a different set of gradient criteria.

Using the data described in this report, several experiments were carried out to examine the variation in parameter estimates that could result from different initial conditions (assuming Newton's gradient criteria).† Among the cases studied, the most significant differences were found for the vector P with 11 parameters, principally among the parameters of the retirement component. For schedules without the retirement peak, the vector P^* shows no variation in most cases.

†For a complete description of gradient methods, see Fiacco and McCormick 1968, Bard 1974.

The impact of the gradient criteria on the optimal vector P^* was also analyzed, using the Newton and the Steepest Descent methods. The effects of these two alternatives were reflected in the computing times but not in the values of the vector P^*. Nevertheless, Bard (1974) has suggested that both methods can create problems in the estimation, and therefore they should be used with caution in order to avoid unrealistic parameter estimates. It appears that the initial parameter values may be improved by means of an interactive approach suggested by Benson (1979).

Appendix B contains summary statistics of national parameters and variables of the reduced sets of observed migration schedules, and Appendix C contains national parameters and variables of the full sets of observed migration schedules. To save space these data have not been included here. Copies of the complete report may be obtained from:

Office of Communications (Distribution)
International Institute for Applied Systems Analysis,
A-2361 Laxenburg,
Austria.

THE POSSIBLE SHARE OF SOFT/DECENTRALIZED RENEWABLES IN MEETING THE FUTURE ENERGY DEMANDS OF DEVELOPING REGIONS

A.M. Khan
International Institute for Applied Systems Analysis, Laxenburg, Austria
and
Pakistan Atomic Energy Commission, Pakistan

SUMMARY

The consumption of commercial energy in the developing countries is expected to increase by a factor of about 10 over the next 50 years. As most of the infrastructure related to their energy consumption and supply will undergo a major expansion during the next few decades, it should be possible to introduce soft/decentralized technologies based on renewable sources of energy in order to meet a significant fraction of the future energy demand in these countries. This report assesses what could, under favorable conditions, be a maximum share of soft/decentralized renewables in meeting the future commercial energy demand of the three market-economy developing world regions considered in the global energy study of the IIASA Energy Systems Program (Region IV: Latin America (LA); Region V: Africa (except Northern Africa and South Africa) and South and Southeast Asia (Af/SEA); Region VI: Middle East and Northern Africa (ME/NAf)).

A number of soft/decentralized technologies based on renewable sources of energy are looked into and their investment requirements (per unit capacity) and fuel production costs are compared with those of conventional supply schemes. Shortcomings and practical difficulties associated with some of these technologies that render them unsuitable for meeting different categories of demand in rural areas and in small and large urban centers are analyzed. It is concluded that the most promising soft/decentralized renewables are: windmills and small hydropower units for use in irrigation and for supplying electricity to rural households and small towns; charcoal for meeting thermal energy requirements of industry, households, and the service sector; biogas for use in rural households; and solar heat for supplying hot water/steam to industries, households, and services. Maximum feasible shares of soft/decentralized renewables to meet different categories of demand by 2000 and 2030 are stated. The quantities of energy to be supplied by different technologies are estimated by superimposing these shares on the sectoral and subsectoral demands for electricity and useful thermal energy as projected in the High and Low scenarios of the IIASA study.

The shares of soft/decentralized renewables in the commercial final energy demand of the three developing regions in 2030 are estimated to be about 7% for electricity, 17% for nonelectric energy (charcoal 15%, biogas 1%, soft solar 1%), and 15% for total final energy. The soft/decentralized renewables will be required in 2030 to supply 0.7–1.1 TWyr of final energy, which is large in relation to the commercial final energy demand of 0.6 TWyr in 1975 in the three regions.

Soft/decentralized renewables will be of greatest importance in rural areas. With the continuing use of noncommercial fuels in the rural areas of the three regions to satisfy about 60% of their thermal energy requirements, the commercialized soft/decentralized renewables are projected to meet about 35% of the electricity demand and 22% of the nonelectric commercial final energy demand of the rural sector in 2030. For the urban sector, it is estimated that by 2030 soft/decentralized renewables may be invoked to cover about 15% of the electricity demand and about 17% of the nonelectric demand originating from small urban centers.

Among the soft/decentralized renewables charcoal stands out as the most important and, at the same time, the most difficult component of the future supply schemes. In Region IV(LA), a region with large forest resources, the projected High scenario demand for charcoal in 2030 may be met by utilizing about one-third of the annual regenerative capacity of the region's natural forests. However, the same is not true for Regions V and VI (Af/SEA, ME/NAf). If it is assumed that not more than about one-third of a region's natural forest may be harvested for energy purposes, Region V will be required by 2030 to raise energy plantations over an area equivalent to about 10% of its present arable land in order to meet its High scenario demand for charcoal. The situation will be even more critical in Region VI, whose natural forest resources are extremely small. This region will have to undertake extensive energy plantation operations to meet its demand for charcoal in both scenarios. The area required to be put under plantations by 2030 is estimated, in the High scenario, to be about the same as the region's present arable area.

This assessment of the possible role of soft/decentralized renewables is based on optimistic assumptions. The envisaged supply of energy by these renewables can certainly be met if a well organized, large effort is initiated by the developing regions without loss of time, and pursued vigorously for the next 50 years. Any further delay or halfhearted effort would result in a smaller contribution by soft/decentralized renewables than is anticipated here.

---

1 INTRODUCTION

The developing world regions considered in this report are only those comprising market-economy countries and are defined as Regions IV, V, and VI in the IIASA study (Energy Systems Program Group of IIASA 1981). Region IV is Latin America (LA); Region V is Africa (excluding Northern Africa and South Africa) and South and Southeast Asia (Af/SEA); and Region VI is the Middle East and Northern Africa (ME/NAf). The countries in each region are listed in the Appendix.

The consumption of primary commercial energy in these regions in 1975 amounted to 0.80 TWyr yr^{-1}. According to the two IIASA scenarios, corresponding to different projected levels of economic growth, the consumption will increase to about 8–13 times its 1975 value by 2030. As most of the infrastructure related to energy consumption and supply in the developing countries will be established over the next 50 years, it is worth while to explore the extent to which soft/decentralized technologies based on renewable sources of energy may be called upon to meet the future energy demand in these regions. To do this we shall make a brief survey of such technologies that hold promise of wide-scale application in the developing regions; identify the plausible extent of their application to different energy-consuming activities; and then estimate their contribution to meeting the energy demand projected in the two IIASA scenarios for the years 2000 and 2030, assuming that an aggressive policy were to be pursued in favor of soft/decentralized renewables. Finally, we shall consider how such an assessment stands in relation to the resource base of renewables in the three developing regions.

In the context of this paper renewable energy sources are solar energy, wind, hydropower, biomass, geothermal energy, etc. They will frequently be referred to as "renewables." The term "soft" refers to simple technologies such as harvesting of wood from forests and plantations and use of small-scale hydropower, whereas the term "decentralized" implies localized systems, e.g. windmills, small hydropower units, and biogas plants that are not part of centralized supply systems. Large windmills may not be called soft technologies but are still decentralized technologies, whereas wood harvesting may involve a very large organized effort but is still considered here as a soft technology. The technologies included in this assessment are not necessarily "soft" and "decentralized" at the same time, but they belong to at least one of the categories. Centralized supply schemes using relatively advanced technology, such as hydropower plants, interconnected chains of windmills, and solar–thermal–electric conversion, are therefore not considered here. The soft/decentralized technologies based on renewable sources of energy will often be referred to as "S/D renewables."

2 SOME BACKGROUND INFORMATION

In 1975 the developing countries of Africa, Asia, and Latin America had a population of 1874 million, 30% of whom lived in urban areas and 70% in rural areas (United Nations 1976). The main sources of income in rural areas are activities such as agriculture, fishing, and cattle breeding. Most of these activities, in particular farming, are still carried out by way of centuries-old traditional practices requiring intensive use of human labor and draft power. However, the pressure caused by increasing population, limited resources of arable land (0.34 ha per capita in 1975, as against 0.62 ha in the developed regions (Food and Agriculture Organization 1977)), and inadequate supplies of water from precipitation and canals is gradually forcing a change toward more productive, mechanized methods of both farming and irrigation.

At present there is little industrial activity in the rural areas although efforts are being made by various governments to establish handicraft and cottage industries to reduce the population shift to urban areas. Most of the villages have a few hundred inhabitants and are distributed close to the arable land. The facilities of electricity and transportation are generally inadequate. According to a survey by the World Bank in 1975 only about one-sixth of the total rural population of the developing countries (4%

in Africa, 15% in Asia, and 23% in Latin America) had access to electricity. The present trend of rural electrification seems to be dominated by extending grid supplies to rural areas rather than by establishing independent, small generating units.

About 60% of the urban population in the developing regions was concentrated in 1975 in cities of 100,000 or more inhabitants (UN 1976). Although all urban dwellers have potential access to electricity, a large fraction of the poor do not have the financial resources to cover the initial cost of electrification and are still without electricity. A substantial fraction of the electrified households use electricity only for lighting and for operating ceiling fans, radios, and television sets.

The requirements of nonelectric energy in both rural and urban households are dominated by cooking needs. This is because about three-quarters of the population live in areas where hardly any space heating is required in winter and the use of hot water is also minimal, in general. Noncommercial fuels to the extent of 0.5 TWyr were used in 1975 mainly for cooking and heating in households. Understandably the share of noncommercial fuels in meeting the household requirements is much higher in rural than in urban areas. It is generally estimated that the efficiency of noncommercial fuels, as they are presently used, is only 5–10%, compared with 30–60% for fossil fuels. Thus the present level of use of noncommercial fuels serves requirements that could perhaps be met by some 50–100 GWyr of fossil fuels.

The manufacturing activities are practically all confined to urban areas. They are generally based on processes and technologies similar to those that either are presently used in the developed countries or were used by them within the last few decades. The mining and construction operations are, in general, very labor-intensive (except for oil and gas mining), although mechanization is gradually being increased.

Things will change considerably in the next 50 years. In particular there will be much progress in industrial activity coupled with an increased level of urbanization. The rural development programs will help provide electricity to a large fraction of the rural population. Agriculture, construction, and mining activities will also become much more mechanized. In view of the growing scarcity of fossil fuels the renewable forms of energy will certainly have a role in various sectoral activities but this role will vary from sector to sector and will not be the same for all groups of the population. For example, cities with populations of millions will almost exclusively have to rely on centralized electricity grids, whereas it may be economically attractive to supply electricity to scattered and remote villages and to irrigation water pumps in certain areas from windmills and small hydropower units. For a proper assessment of the role of soft/decentralized renewables one therefore needs to look into the distribution of future energy demand among the sectoral activities as well as among various groups of the population.

3 PROMISING SOFT/DECENTRALIZED TECHNOLOGIES

The developing world regions have a large potential of renewable energy sources. For example, the annual increment of wood above ground in the regions' forests is equivalent to about 7 billion tonnes (7×10^9 t) of dry wood with an energy content of about 4 TWyr (Earl 1975, FAO 1977). The energy content of agricultural and animal wastes produced at present in these regions is estimated to be about 0.4 TWyr yr^{-1} (FAO 1977, Parikh 1978, Revelle 1979). This will probably increase by a factor of 3–5 over the next 50 years. The total hydropotential in the developing regions corresponds to about 0.5 TW(e)yr yr^{-1}, of which only about 6% is in use at present, and that mostly through centralized generating schemes (UN 1977, World Energy Conference 1978). There is an abundance of sunshine in most of the regions, with an average solar irradiance of 1500–2000 kWh m^{-2} yr^{-1}. Winds having useful velocities are also found near coastal areas at distances of up to several hundred kilometers from the coastlines. As a rough estimate, the realizable potential of mechanical power from wind available in the developing regions may be taken as 0.5 TWyr yr^{-1}. The realizable potential of wet geothermal energy would also correspond to about 0.5 TWyr yr^{-1}. These potentials of wind power and geothermal energy assumed for the developing regions are half of those estimated for the world as a whole (Energy Systems Program Group of IIASA 1981).

In spite of their large potential the only significant applications, so far, of renewable sources of energy in the developing regions have been centralized hydropower generation and use of noncommercial fuels derived from disorganized cutting of forests and from agricultural and animal wastes. Recently some countries (most notably India) have started promoting the use of biogas plants in rural areas, while Brazil has embarked on a program of production of alcohol from sugarcane for use as fuel. Other applications such as those of windmills, small hydropower units, soft solar devices, and plantation schemes, are lagging further behind and are still in the exploratory stages.

Since the various soft/decentralized technologies have not been commercialized it is not possible to make firm estimates of their investment requirements or fuel production costs. In addition, both the investment requirement per unit of installed capacity and the total cost per unit of energy produced will vary considerably for each technology, depending on the geography, environmental conditions, and indigenous industrial capability. Nevertheless some rough estimates are necessary in order to identify the technologies that hold promise of large-scale utilization in areas where resource conditions are favorable.

Table 1 presents some estimates (in 1975 US dollars) of the capital costs (per unit peak capacity) and the average energy production costs of S/D renewables for electricity generation. The estimates of capital costs are, in general, based on the prices of basic equipment now commercially available in some countries. The cost of electricity production has been calculated by fixing a charge of 10% per annum on capital cost and using the appropriate duty cycle in column 4. The table also allows a comparison of the S/D renewables with conventional centralized systems.

As centralized systems of electricity supply also entail large investments in transmission networks and have associated maintenance and distribution expenses, the actual energy supply costs from centralized systems would be some 50–100% higher than those in Table 1. Thus under favorable conditions the supply from individual windmills and small hydropower units may be more economical. The cost of electricity production from photovoltaic arrays is, however, an order of magnitude too high at present. It is too early to say whether the cost can be reduced sufficiently to make such systems an economically attractive proposition within the next 50 years.

Although windmills and some small hydropower units appear economically competitive with centralized power generation in terms of supply cost per unit of electricity, their energy sources are irregular by nature. This shortcoming makes them unsuitable for supplying regulated power to large cities, major industries, and electrified transportation systems.

TABLE 1 Electricity supply from S/D renewables: estimates of capital costs and electricity production costs, and comparison with centralized systems.

Technology	Capacity	Capital cost[a] (1975\$ kW(e)$^{-1}$)	Assumed duty cycle (h yr^{-1})	Electricity cost[b] (1975\$ kW(e)h^{-1})
Windmills	< 1 kW(e)	3000–6000	2500	0.12–0.24
	5–15 kW(e)	1000–2000	2500	0.04–0.08
	3 MW(e)[c]	450	2500	0.02
Small hydropower units	0.5–10 kW(e)	1000–7000	4000	0.03–0.18
Photovoltaic devices	< 1 kW(e)	15,000–30,000	2000	0.75–1.50
Centralized systems				
Large hydropower units	250 MW(e)	800–1500	4000	0.02–0.04
Coal-fired plants	300 MW(e)	500	4000	0.02
Oil-fired plants	300 MW(e)	400	4000	0.03

[a]These estimates are based essentially on the information given by the National Academy of Sciences (1976), WEC (1978), Cecelski *et al.* (1979), and the Energy Systems Program Group of IIASA (1981).
[b]These costs have been worked out by assuming a fixed charge of 10% per annum on capital investment and neglecting the operating costs. For coal- and oil-fired plants an allowance has also been made for the fuel cost at 25\$ ton^{-1} for coal and 12\$ bbl^{-1} for oil.
[c]Large windmills with capacities in the MW(e) range are in the development stage.

However, this unsteady nature would not pose much of a problem in meeting irrigation water-pumping requirements. Similarly, villages and small towns might tolerate to a considerable extent an irregular electricity supply and could meet part of their requirements from diesel-operated systems or central grids. However, the use of windmills and small hydropower units would call for considerable investments (\$2000–10,000 per kW(e)yr yr^{-1} of supply), which may be difficult for individuals or small groups to afford without government finances.

Table 2 lists the capital costs (per unit peak thermal capacity) and the average energy production costs of some solar devices, biogas plants, and alcohol production plants. The simple solar devices considered for water heating, space heating, and cooking are made from reflecting material that costs about \$100 per m^2 of surface. The biogas plants are the one-family and community units of the Indian design discussed by Parikh (1978). The capital cost for alcohol production was estimated by Goldemberg (1979) and corresponds to the Brazilian situation. The energy costs in column 4 (expressed per unit of useful energy for solar devices and of energy content of fuel for other plants) are, again, based simply on a fixed charge of 10% per annum on capital cost and the assumed duty cycle in column 3. Also listed in Table 2 are the production costs for fuelwood and charcoal estimated by Earl (1975) for two different schemes in East Africa: harvesting natural forests, and raising energy plantations. Unlike the energy costs of solar heating, biogas, and alcohol, which have been worked out by neglecting the operating costs (although they would not be negligible for large biogas and alcohol plants), the costs of fuelwood and charcoal include both the royalty paid on the forest/land area used and the operating costs. For comparison the energy costs of coal and of oil are also shown, but the actual price paid by the user for coal and oil will, in general, be much higher when taxes, profits, and transportation are taken into account.

TABLE 2 Thermal energy supply from S/D renewables: estimates of capital costs and energy production costs.

Technology	Capital cost[a] (1975\$ kW^{-1})	Assumed duty cycle (h yr^{-1})	Energy cost[b] (10^{-3} \$ kWh^{-1})
Solar water heating	300–600	2000	15–30
Solar space heating	400–800	500	80–160
Solar cooking	200–300	500	40–60
Biogas:			
one-family units	500	Continuous	6
community plants	250	Continuous	3
Alcohol production from sugarcane	800	4000	20
Fuelwood production:			
harvesting natural forests	–	–	0.3
energy plantations	–	–	0.4–0.7
Charcoal production:			
harvesting natural forests	–	–	1.2
energy plantations	–	–	1.8–2.5
For comparison			
Coal at 25\$ ton^{-1}	–	–	3
Oil at 12\$ bbl^{-1}	–	–	7

[a]These estimates are based essentially on the information given by NAS (1976), WEC (1978), Parikh (1978), and Goldemberg (1979).

[b]Except for fuelwood and charcoal, the energy costs reflect only the contribution of capital costs at a fixed charge of 10% per annum. The cost estimates for fuelwood and charcoal are based on the data given by Earl (1975).

To the extent that one can rely on the estimates in Table 2, renewable energy supply schemes based on biogas plants, energy plantations, and harvesting of natural forests appear attractive when their energy production costs are compared with the prices of coal and oil in the international market. Alcohol production may become more economical in coming years, as coal and oil prices rise. The energy costs for solar devices are relatively high but may fall as a result of the current R&D effort and the possible introduction of mass production. In any case the relative energy costs of different solar devices will remain roughly in the same proportion as in Table 2; in particular, solar water heating will remain more attractive than solar space heating. Very simple solar cookers, costing as little as \$15 each, were produced in India in the 1960s but failed to be accepted in rural areas even though the government gave financial support to popularize them (Cecelski *et al.* 1979). The experience in other countries has not been very different. This is understandable because solar cookers are very inconvenient for the housewife, and the time when they may be used efficiently does not coincide with the time when most people want to have warm meals. Therefore, we do not expect much success for solar cookers (in spite of lowered production costs) as long as more convenient means of energy supply for cooking remain available at acceptable costs. These alternatives for rural areas, which may be considered the most likely environment for using solar cookers, are fuelwood, agricultural and animal wastes, and biogas.

Biogas generation offers a very efficient and convenient form of fuel but its application is limited to rural areas. In view of the relatively low investment potential of families in villages it is to be expected that large plants for the community, perhaps built with the help of external financing, will be much more successful than the smaller, one-family units. While envisaging the application of biogas plants one should keep in mind the traditional social customs and habits of people in the handling of animal wastes. Taking into account these considerations and the relative availability of forest wood in different developing regions, we feel that the biogas plants will not have much success in the rural areas of Latin America. On the other hand, they may be very successful in the villages of Africa and Asia provided that the necessary investment funds are made available.

In order to consider the long-term prospects of supplying wood/charcoal from natural forests and energy plantations and of producing alcohol from sugarcane or other crops, one should look at the present land utilization patterns in developing regions. Table 3 shows that there are about one billion hectares of forest area in each of Regions IV and V whereas Region VI has a meager 28 million hectares of woodland. Although the indiscriminate cutting of forests in recent years, particularly in Africa and Southeast Asia, has been causing serious deforestation and land erosion, silviculturally sound practices may allow harvesting of large amounts of wood from natural forests without adverse effects, and perhaps even with beneficial effects (Earl 1975).

Both of the other alternatives (energy plantations and alcohol production from sugarcane or other similar crops) are to be seen in competition with the requirements of food production for a growing population. The availability of arable land in Regions IV, V, and VI in 1975 was only 0.45, 0.32, and 0.33 ha per capita, respectively. These figures appear low when compared with 0.62 ha per capita, the average area of arable land available in the developed world regions. There does not appear to be much prospect for expanding arable land in the developing regions, so the per capita availability of good agricultural land in these regions will become even smaller in the next 50 years, over which period the population will increase to about 2.5 times the present number. The production of sugarcane, or of similar crops, requires good agricultural land with an adequate water supply. (The cultivated area of sugarcane required to produce 1 GWyr yr^{-1} of alcohol in Brazil was estimated by Goldemberg (1979) to be 0.4×10^6 ha.) It is unlikely that the production of alcohol from agricultural crops will be able to play any significant role in the long run,

TABLE 3 Distribution, by region, of population and land in 1975. Arable land includes the area under permanent crops. Source: FAO (1977).

Region	Population (10^6)	Arable land (10^6 ha)	Permanent pastures (10^6 ha)	Forests and woodland (10^6 ha)	Other land (10^6 ha)	Total land (10^6 ha)
IV	319	142	527	1071	374	2114
V	1422	456	710	963	1074	3203
Africa	319	184	671	633	840	2328
Asia	1103	272	39	330	234	875
VI	133	45	172	28	802	1047
Africa	57	13	45	3	453	514
Asia	76	32	127	25	349	533

when there will be great pressure to use such land for producing food. On the other hand, energy plantations based on certain species of fast-growing trees can be raised on marginal agricultural land (Earl 1975) and, as such, do not necessarily interfere with food production requirements. Raising such plantations would help to control land erosion in certain areas and may even be a welcome approach in the neighborhood of populated areas far from natural forests.

Although wood can be used directly as fuel it is not as convenient to handle and transport as charcoal and burns with a lower efficiency (for supplying useful energy). It is therefore expected that most of the available wood from forests and plantations will be converted to charcoal for use in industry and urban households. The additional expenditure incurred in conversion of wood to charcoal and the conversion energy losses (equivalent to about 50% of the energy content of wood) will then be more or less counterbalanced by the savings in transportation expenses and the higher burning efficiency of charcoal.

In principle, wood and its associated tree matter (leaves and twigs) can also be used to obtain liquid or gaseous fuel but no proper cost estimates are available for such an operation at a sizable scale. In general one would expect the investment per unit capacity of liquid/gaseous fuel to be much higher for this operation than for production of alcohol from sugarcane. Therefore we do not consider this as a viable alternative to charcoal production.

The use of wood/charcoal for electricity generation at a decentralized level is not considered here as the capital costs (per unit capacity) of small thermal power plants will be too high to make them economically viable. Large thermal power plants (of the order of 100 MW(e)) may be more economical but they will also be more complex systems and will need to be linked to central grids; therefore, they cannot be considered as either a soft or a decentralized technology. In any case it would be preferable to use the available wood/charcoal for direct thermal uses rather than for power generation if the demand for the former use alone is sufficient to put a great pressure on the resources of natural forests. (We shall show that this situation will apply to the three developing regions if the use of S/D renewables is promoted to the extent envisaged in the present assessment.)

The costs of supplying energy from wet geothermal sources are not mentioned in Table 2 as these sources are limited to only a few locations, and hence are unsuitable for wide-scale decentralized use; moreover, the costs will depend very much upon location.

For the reasons given above, we view only four soft/decentralized technologies based on renewables as holding promise for wide application to developing world regions over the next 50 years.

(i) *Windmills, small hydropower units*: for irrigation water pumping and supplying electricity to villages and small towns.

(ii) *Charcoal*: for industry, households, and the service sector, mainly as a source of thermal energy.

(iii) *Biogas*: for rural areas of Africa and Asia, where the handling of animal wastes is traditionally and culturally acceptable.

(iv) *Solar heat*: mainly for supplying hot water/steam to industries and hot water to households and services; of limited use for space heating (in rich households and the service sector where the availability of capital is not a problem).

4 SCENARIO ASSUMPTIONS CONCERNING USE OF SOFT/DECENTRALIZED RENEWABLES

Many factors will determine the extent to which soft/decentralized technologies based on renewable energy sources may be invoked to meet the future energy demand. The most important of these are listed below:

- the cost economics of S/D renewables as compared with those of conventional forms of energy;
- the magnitude of the domestic resources of conventional fuels;
- the production potential of renewables close to demand centers;
- the convenience of use and social preferences;
- the access of different sections of the population to central power grids;
- the investment potential of individuals and small groups for financing independent installations;
- the government loan and investment policies for funding decentralized supply sources in preference to centralized facilities; and
- the problems of institutional changes and management.

All these factors will vary from region to region and much more so from country to country. A detailed analysis that acknowledges so many factors with their inherent uncertainties may prove to be a formidable task. Therefore, we shall make some simplifying assumptions that allow us to estimate the possible share of S/D renewables in meeting the future energy demands of the developing countries.

4.1 General Assumptions

(1) In view of the general scarcity of fossil fuels and the high cost of electricity transmission through centralized grids to small towns and rural areas, vigorous efforts will be made to make increasing use of renewables on a soft/decentralized basis. The institutional, managerial, and financial problems will be overcome through national policies and governmental support.

(2) The electricity requirements of *cities* (large urban agglomerations each with 100,000 or more inhabitants) will be met solely through centralized supply schemes, whereas those of *towns* (urban agglomerations having up to 100,000 inhabitants) and *villages* (covering all rural households) will be met partly by S/D renewable sources. It is assumed that half of the towns and villages will be in areas where renewable sources may be used for electricity generation on a decentralized level. In view of the variable nature of these sources (wind, small-scale hydropower), the problems of energy storage, and the need for diversification of power sources, it is further assumed that of the power requirements in such areas not more than 60% in villages and 30% in towns can actually be supplied by S/D renewables.

(3) The decentralized renewable sources of power will be available in areas covering about half of the agricultural land that will require irrigation by pumping of water. As the demand for irrigation is more flexible than the needs of households, services, and industries, it is assumed that up to 80% of the requirements in favorable areas can be supplied

by decentralized renewables. It is also assumed that the other power needs of the agricultural sector (e.g. for grinding grain) can be met by S/D renewables to the same extent.
(4) The main manufacturing industries will all be in urban areas only and distributed between cities and towns in proportion to their respective populations. The small power requirements of cottage industries in villages will be included in the electricity demand of the rural households.
(5) Rural electrification will increase rapidly in all developing regions and will be complete within the next 50 years. The consumption of electricity will be the same in a village household as in an urban household of the same region. Although this amounts to overemphasizing the share of rural households in electricity consumption, it may be justified by assumption (4).
(6) The energy-consuming activities of the service sector will be confined to urban areas and distributed between cities and towns in proportion to population.
(7) It will not be possible to meet the electricity requirements of the mining, transportation, and construction activities with S/D renewables to any significant extent.
(8) Although, in principle, all the useful thermal energy demand of the household/service sector and most of the industrial demand may be considered potentially suitable for S/D renewables, in practice it will not be possible to make use of these renewables to such an extreme. For example, rich households, sophisticated service establishments, and certain large modern industries will in all likelihood continue to make use of relatively more convenient and clean forms of conventional fuels. The reliance of industries upon fossil fuels for certain uses will also be dictated by specific processes, e.g. those requiring high-temperature furnace heat. Therefore, we shall make the following assumptions.

(i) In the industrial sector, up to 80% of the hot-water/low-temperature steam requirements, 60% of the high-temperature steam requirements as well as coke needs of the steel industry, and 12% of the high-temperature furnace heat requirements may be met by S/D renewables.
(ii) In the household sector, the S/D renewables may be invoked to meet, in combination with noncommercial fuels, as much as 90% of the useful thermal energy requirements (for cooking, water heating, space heating) in villages, 80% in towns, and 60% in cities.
(iii) In the service sector, the share of S/D renewables in meeting the useful energy demand (for water and space heating) may be as large as 60% for towns and 40% for cities.

(9) S/D renewables will not be used to any significant extent to produce liquid fuels. Since practically all the nonelectric demand of transportation, mining, construction, and agricultural activities will be for liquid fuels, the S/D renewables will not be required in any significant amount to meet it. (The gas used in petroleum-mining activities cannot be replaced by renewables either.) Similarly, the use of fossil oil in the production of petrochemical feedstocks will not be replaced by S/D renewables.
(10) The use of S/D renewables will proceed so that by 2030 their share in meeting the energy demand for various sectoral activities will be as high as anticipated in these assumptions.

4.2 Additional Assumptions Concerning Specific Renewables

(A) The use of noncommercial fuels was increasing in the past. It is assumed that, as a result of efforts to organize harvesting of wood from forests and to introduce biogas plants in rural areas, the total quantity of noncommercial fuels used (user-collected firewood, agricultural and animal wastes) will stay about the same as in 1975 in each region. Further, efforts will be made to introduce devices that will help to improve the efficiency of use of noncommercial fuels, by 2030, to 1.6 times its 1975 value in each of Regions V and VI, and to 2.0 times its 1975 value in Region IV. It is also assumed that noncommercial fuels will be used mostly in villages and that the balance, if available, will be used in the households of small towns.
(B) Extensive use will be made of biogas plants in the rural areas of Regions V and VI. By 2030, 90% of the nonelectric energy demand from village households will be allocated to renewables; the part that is not met by noncommercial fuels will be supplied by biogas generation. The efficiency of burning biogas will be the same as that of natural gas.
(C) Soft solar devices for water heating, space heating, and steam generation will find increasing use after the turn of the century. It is assumed that by 2030 such devices, backed by 20% fossil fuel support, will be used to meet:

(i) 30% of the hot-water/low-temperature steam requirements and 10% of the high-temperature steam demand of industries in all regions;
(ii) 30% of the hot-water demand in households of Regions IV and V (15% in Region VI);
(iii) 50% of the space-heating demand of centrally heated, single-family houses in Region IV and 20% of this demand in Region VI; and
(iv) 50% of the heat requirements of low-rise buildings of the service sector in Regions IV and V (20% in Region VI).

(D) All wood supplied commercially will be converted to charcoal. The efficiency of charcoal for different uses will be the same as the average fossil fuel efficiency for corresponding applications.

5 FUTURE DEMANDS FOR SOFT/DECENTRALIZED RENEWABLE ENERGY

The assumptions of Section 4 provide a general framework for estimating the possible overall contribution of S/D renewables to satisfying the future energy demands of the developing regions. It has been assumed that by 2030 S/D renewables in two groups (as sources of electric and nonelectric energy) will penetrate their respective markets covering various sectoral activities to the maximum extent feasible in our judgment. The projected penetrations are listed in Tables 4 and 5. Also listed are appropriate figures for the year 2000, in line with the projected penetrations for 2030. Relatively high penetrations are assumed for nonelectric energy demand in 2000 (Table 5) compared with those for electric energy demand (Table 4). This is because: (i) we feel that it would not be too difficult for industries and the household/service sector to shift from fossil fuels to charcoal if appropriate policy measures were adopted soon enough on a national basis; (ii) the industrial

TABLE 4 Projected penetrations of decentralized renewable sources of energy in the electricity supply schemes of developing regions (expressed as a percentage of electricity demand). Only 60% of the village households in Region IV and 50% in each of Regions V and VI are assumed to have access to electricity in the year 2000.

Demand sector	2000	2030
Households		
Cities	–	–
Towns	2.5	15.0
Villages	5.0	30.0
Service sector		
Cities	–	–
Towns	2.5	15.0
Manufacturing		
Cities	–	–
Towns	2.5	15.0
Agriculture	7.5	40.0
Other sectors		
Transport, mining, construction	–	–

TABLE 5 Projected penetrations of soft renewable sources of energy in the nonelectric energy supply schemes of developing regions (expressed as a percentage of useful energy demand). Penetration of soft renewables in the household sector includes use of noncommercial fuels.

Demand sector	Nature of demand	2000			2030
		Region IV (LA)	Region V (Af/SEA)	Region VI (ME/NAf)	Regions IV, V, VI
Households					
Cities	Cooking, space heating, water heating	40	45	20	60
Towns	Cooking, space heating, water heating	60	70	40	80
Villages	Cooking, space heating, water heating	75	85	60	90
Service sector					
Cities	Space heating, water heating	20	20	20	40
Towns	Space heating, water heating	30	30	30	60
		Regions IV, V, VI			
Manufacturing					
	Low-temperature steam/ hot water		40		80
	High-temperature steam		30		60
	Furnace heat		6		12
	Steel industry (coke replacement)		30		60
	Feedstocks (oil replacement)		–		–
Other sectors					
Transport, agriculture, construction, mining	Mainly liquid fuel demand		–		–

infrastructure is still being established in the developing regions and this favors rapid penetration of renewables in the industrial sector if a number of new industries opt for them; and (iii) renewables, in the form of noncommercial energy, were already supplying in 1975 about 40, 70, and 10% of the useful thermal energy requirements of households (urban and rural) in Regions IV, V, and VI, respectively. The penetrations of S/D renewables in the nonelectric energy demand of households (Table 5) include the use of commercial and noncommercial forms of renewable energy and should, therefore, be considered in relation to the present situation.

There are some additional assumptions for the year 2000 not explicitly covered in Section 4.

(*a*) The fractional uses of soft solar energy for various activities will be a factor of about 2–10 lower than those in 2030.

(*b*) The efficiency in using noncommercial fuels will be only about 15–20% higher than that in 1975.

(*c*) Biogas will be used to supply only 50% of the fraction of the nonelectric energy demand of rural households that is allocated to renewables but not met by noncommercial fuels.

(*d*) Electrification will extend to 60% of the villages in Region IV and 50% of those in Regions V and VI.

To make quantitative estimates of the requirements of soft/decentralized forms of renewable energy in 2000 and 2030, we need the corresponding projections for (i) the population distribution among cities, towns, and villages, and (ii) the sectoral requirements of electricity and useful thermal energy for various activities. The projections for population distribution are given in Table 6 together with the historical data for 1950 and 1975. For the year 2000, the estimates of rural/urban distribution are based on UN (1976) projections, whereas those for the distribution of urban population between cities and towns are extrapolations of the UN projections for 1985 (UN 1976). All the estimates for 2030 are our own, made by extrapolating the historical data and the available UN projections.

TABLE 6 Distribution of population (%) in cities, towns, and villages: historical data and projections. Source: UN(1976).

Population grouping	1950	1975	2000	2030
Region IV				
Cities	19	36	53	65
Towns	22	24	22	20
Villages	59	40	25	15
Region V				
Cities	6	13	23	44
Towns	8	9	11	11
Villages	86	78	66	45
Region VI				
Cities	14	29	45	65
Towns	12	16	17	17
Villages	74	55	38	18

For the projections of useful and final energy demand for different sectoral activities in 2000 and 2030 we shall use the results of a detailed energy demand analysis (Khan and Hölzl 1981) carried out with the help of a model called MEDEE-2 (Lapillonne 1978, Hölzl 1981) in connection with the medium- to long-term global energy study recently completed by the Energy Systems Program Group of IIASA (1981). The various assumptions and projections concerning growth of population, growth and evolution of economy, lifestyle changes, technological improvements, conservation measures, and growth of useful and final energy consumption until 2030, in two different scenarios called Low and High, are described in detail by Chant (1981), Khan and Hölzl (1981), and the Energy Systems Program Group of IIASA (1981) and will not be discussed here. Table 7 summarizes the relevant information on population, gross domestic product, commercial final energy consumption, the share of electricity in this final energy consumption, and noncommercial energy consumption for each of Regions IV, V, and VI for the base year 1975, together with the corresponding projections for 2000 and 2030 in the two scenarios. Additional information about the demands for electricity, useful thermal energy, and liquid fuels (for specific uses), together with their sectoral distributions in 1975, 2000, and 2030, is given in Tables 8, 9, and 10 for Regions IV, V, and VI. The three tables also list those parts of the demands for electricity and useful thermal energy (including coke requirements) that

TABLE 7 Projections of population, GDP (in constant 1975 US dollars), final energy (commercial), and noncommercial energy in the IIASA High and Low scenarios.

Parameter	1975	2000		2030	
		Low	High	Low	High
Region IV					
Population (10^6)	319	575		797	
GDP (10^9 $)	340	918	1272	2229	3569
Final energy (commercial) (GWyr)	255	733	1004	1656	2640
Share of electricity (%)	10	12	12	16	15
Noncommercial energy (GWyr)	109	109		109	
Region V					
Population (10^6)	1422	2528		3550	
GDP (10^9 $)	340	924	1207	1995	3488
Final energy (commercial) (GWyr)	253	802	1063	1876	3173
Share of electricity (%)	9	12	13	15	16
Noncommercial energy (GWyr)	344	344		344	
Region VI					
Population (10^6)	133	247		353	
GDP (10^9 $)	190	643	900	1310	2918
Final energy (commercial) (GWyr)	106	434	578	868	1638
Share of electricity (%)	4	12	12	15	17
Noncommercial energy (GWyr)	10	10		10	
Regions IV + V + VI					
Population (10^6)	1874	3350		4700	
GDP (10^9 $)	870	2485	3379	5534	9975
Final energy (commercial) (GWyr)	614	1969	2645	4400	7451
Share of electricity (%)	8	12	12	15	16
Noncommercial energy (GWyr)	463	463		463	

TABLE 8 Projections of demands for electricity, useful thermal energy, and liquid fuels and of possible shares of S/D renewables for Region IV. Electricity and liquid fuel demands are expressed as final energy. Useful thermal energy is expressed in terms of equivalent requirements of electricity.

	1975	2000		2030	
		Low	High	Low	High
Demand for electricity (GWyr)	24	85	119	256	402
Shares of sectors (%)					
Households	21	22	22	23	23
Services	8	10	9	13	8
Manufacturing	62	57	60	52	58
Agriculture	1.2	4.0	3.2	4.0	2.9
Others	8	7	6	8	8
Demand to be met by S/D renewables (GWyr)	0	0.9	1.3	12.9	18.6
Demand for useful thermal energy (GWyr)	68	182	247	390	616
Shares of sectors (%)					
Households	30	28	21	25	18
Services	1.2	1.4	1.0	2.2	1.1
Manufacturing[a]	69	71	78	73	81
Demand to be met by S/D renewables (GWyr)	8	52	65	179	269
Shares (GWyr)					
Soft solar	0	2.9	4.1	22	33
Charcoal, biogas	0	39	51	141	220
Noncommercials	8	10	10	16	16
Specific demand for liquid fuels[b] (GWyr)	132	389	533	939	1507
Shares of sectors (%)					
Agriculture	0.6	2.2	1.8	2.7	1.9
Transportation	79	78	76	76	75
Others	20	20	22	21	23

[a]Coke requirements of the steel industry are included on an equivalent calorific basis.
[b]Liquids required as feedstocks for petrochemical industries are included.

can be met by S/D renewables on the basis of the assumptions in Section 4 and those made earlier in this section, and that are consistent with the population distribution projections of Table 6.

Using the data of Tables 6–10 as a basis, and the efficiency improvement projections for different fuels and processes as embodied in the MEDEE-2 analysis of Khan and Hölzl (1981), we present in Tables 11 and 12, respectively, the demands in 2000 and 2030 for electricity, nonelectric commercial final energy, and noncommercial energy in villages, towns, and cities of the developing regions and provide details of the contributions from different S/D renewables (wind/hydropower, charcoal, biogas, soft solar) in the three types of demand center.

Assuming that the share of villages in the regional consumption of noncommercial fuels in 1975 was about 60% for Region IV, 85% for Region V, and 75% for Region VI, we estimate that only about 6% of that year's demand, both for total commercial final energy and for electricity alone, in the developing regions (IV + V + VI) originated from

TABLE 9 Projections of demand for electricity, useful thermal energy, and liquid fuels and of possible shares of S/D renewables for Region V. Electricity and liquid fuel demands are expressed as final energy. Useful thermal energy is expressed in terms of equivalent requirements of electricity.

	1975	2000		2030	
		Low	High	Low	High
Demand for electricity (GWyr)	22	95	133	274	509
Shares of sectors (%)					
Households	7	8	7	10	10
Services	10	11	10	10	9
Manufacturing	75	61	66	57	66
Agriculture	6.4	18	15	19	13
Others	1.8	1.8	1.6	3.4	2.6
Demand to be met by S/D renewables (GWyr)	0	2.1	2.6	31	45
Demand for useful thermal energy (GWyr)	100	249	331	508	883
Shares of sectors (%)					
Households	38	31	24	30	18
Services	0.1	0.2	0.2	0.3	0.3
Manufacturing[a]	62	69	76	70	82
Demand to be met by S/D renewables (GWyr)	26	88	104	239	376
Shares (GWyr)					
Soft solar	0	1.0	1.4	15	30
Charcoal, biogas	0	58	73	183	305
Noncommercials	26	29	29	41	41
Specific demand for liquid fuels[b] (GWyr)	80	309	391	896	1425
Shares of sectors (%)					
Agriculture	3.4	10.7	9.5	11.3	8.7
Transportation	82	70	68	67	63
Others[a]	15	19	23	22	28

[a]Coke requirements of the steel industry are included on an equivalent calorific basis.
[b]Liquids required as feedstocks for petrochemical industries are included.

the requirements* of villages. The share of villages in total final energy, for all three regions taken together, does not change significantly over the next 50 years (Table 13) despite increased urbanization. On the other hand, the share of villages increases by a factor of about 2 if only electricity demand is considered. These results are consequences of the increasing energy intensiveness of agriculture assumed in the MEDEE-2-based energy demand projections and the assumptions of Section 3 concerning rural electrification and use of noncommercial fuels.

Table 14 shows the shares of S/D renewables in the electricity, nonelectric commercial final energy, and total commercial final energy demands of villages, towns, and cities

*The commercial final energy requirements of villages are assumed to consist of the demand of the agricultural sector for irrigation and tractor operations, etc. and the commercial energy requirements of rural households for those needs that are not satisfied by the available supplies of noncommercial fuels. The rural population would also have some share in the transportation energy but this is generally very small and has accordingly been neglected here.

TABLE 10 Projection of demands for electricity, useful thermal energy, and liquid fuels and of possible shares of S/D renewables for Region VI. Electricity and liquid fuel demands are expressed as final energy. Useful thermal energy is expressed in terms of equivalent requirements of electricity.

	1975	2000		2030	
		Low	High	Low	High
Demand for electricity (GWyr)	4.7	53	69	133	270
Shares of sectors (%)					
Households	13	9	10	12	15
Services	8	10	9	16	13
Manufacturing	77	77	77	64	67
Agriculture	2.1	3.6	3.2	5.8	3.7
Others	0.4	0.5	0.5	2.4	2.0
Demand to be met by S/D renewables (GWyr)	0	0.5	0.7	7.8	13.9
Demand for useful thermal energy (GWyr)	18.4	113	142	209	395
Shares of sectors (%)					
Households	41	20	18	23	14
Services	2.3	2.6	2.3	5.2	3.8
Manufacturing[a]	57	77	80	72	82
Demand to be met by S/D renewables (GWyr)	0.7	25	30	90	157
Shares (GWyr)					
Soft solar	0	0.5	0.7	8.4	16.1
Charcoal, biogas	0	23	29	80	139
Noncommercials	0.7	0.8	0.8	1.2	1.2
Specific demand for liquid fuels[b] (GWyr)	69	214	300	466	870
Shares of sectors (%)					
Agriculture	0.6	2.3	1.9	2.6	1.8
Transportation	60	67	66	67	70
Others[b]	39	31	32	30	28

[a]Coke requirements of the steel industry are included on an equivalent calorific basis.
[b]Liquids required as feedstocks for petrochemical industries, as well as gas used in petroleum mining, are included.

in Regions IV, V, and VI, based on the IIASA High scenario energy demand projections. The results for the Low scenario are not very different and have, therefore, been left out. According to Table 14, by 2030 S/D renewables should be able to meet about 35% of the electricity requirements of villages and 15% of those of towns in each region. The commercial renewables will be meeting about 22% of the nonelectric commercial final energy demand of villages, and 17% of that in urban areas (towns, cities) of the three regions taken together. The share of commercial renewables in meeting the energy requirements of the developing regions in 2030 could be, according to the present assessment, as high as 7% of electricity and 17% of nonelectric commercial energy, which amounts to about 15% of the total commercial final energy. In quantitative terms S/D renewables would be required to supply in 2030 about 78 GW(e)yr yr^{-1} of electricity and 1055 GWyr yr^{-1} of nonelectric final energy (charcoal, biogas, soft solar), according to the High scenario. The corresponding figures in the Low scenario would be 52 GW(e)yr yr^{-1} and 657 GWyr yr^{-1} (Table 12).

6 SOME SUPPLY CONSIDERATIONS

The amounts of electricity generated by wind/small-scale hydropower, and of non-electric energy in the forms of charcoal, biogas, and soft solar that will be required in 2000 and 2030 in Regions IV, V, and VI, on the basis of our assumptions in conjunction with the IIASA energy demand projections, have been detailed in Tables 11 and 12. The greatest pressure on sources of S/D renewables, within the time horizon of the present assessment, will be in 2030 in each scenario; moreover, this will be greater in the High than in the Low scenario. We shall therefore discuss the supply of S/D renewables mostly with respect to the High scenario demand for the year 2030.

The per capita demand for electricity in the villages of Regions IV, V, and VI in 2030 is expected to be in the range of 0.06–0.27 kW(e)yr yr^{-1} for the High scenario (Table 15). The corresponding demand in the towns of these regions would be 0.2–0.9 kW(e)yr yr^{-1}. For a typical village of 500 inhabitants and a typical town of 20,000, the power requirements would be 30–135 kW(e)yr yr^{-1} and 4–18 MW(e)yr yr^{-1}, respectively. Thus windmills/small hydropower units (or groups of units) with peak power capacities between just a few kW(e) and a few hundred kW(e) would be needed to meet the requirements of villages, whereas larger systems, with capacities between several hundred kW(e) and a few tens of MW(e), would be necessary to meet even the low-priority requirements of towns. Still these larger systems are considered here as S/D renewables since they may well consist of several separate units that may or may not be connected to each other or to a conventional power plant.

A much larger fraction of human population has settled close to rivers and streams than near the coasts (where wind is strong), so it is assumed that about two-thirds of the power generation by S/D renewables in each region would be based on small-scale hydropower and the remaining one-third would be derived from wind energy. Such an assumption does not call for utilizing more than about one-tenth of the hydropower potential in each of Regions IV and V, but in Region VI it would imply utilizing about 40% of the hydropower potential via decentralized power generation (Table 16). Regions IV, V, and VI used only about 7, 4, and 6% of their respective hydropower potentials in 1975 and that output practically all originated from centralized power generation.

Of the noncommercial fuels used in the developing regions in 1975 about 25 GWyr in Region IV, 115 GWyr in Region V, and 8 GWyr in Region VI are estimated to have been produced from agricultural and animal wastes. The total amounts of such wastes produced in the regions in 1975 are estimated to have been about 83, 300, and 23 GWyr, respectively (Parikh 1978, Revelle 1979). By 2030, higher agricultural production will probably increase these amounts by a factor of 3.5–4.5. (Even no change in agricultural production per capita would need an increase in total agricultural production by 2.5 times.) The requirements of biogas for rural households in 2030 have been estimated (Table 12) as 35–40 GWyr for Region V and 12–14 GWyr for Region VI. This implies that by 2030 some 60–65 GWyr yr^{-1} of agricultural and animal wastes would be used for biogas production in Region V and about 20–25 GWyr yr^{-1} in Region VI, at a biogas conversion efficiency of about 60% (Makhijani and Poole 1975). Thus the production of biogas to the extent envisaged in our assessment would not put any excessive pressure on production of agricultural and animal wastes. Most of these wastes will still remain available in each region for use as noncommercial fuels, for returning to the fields, and for other applications.

TABLE 11 Projections of demands for electricity, nonelectric commercial final energy, and noncommercial energy by villages, towns, and cities in the year 2000, and shares of S/D renewables (GWyr). The energy demands of villages are assumed to comprise requirements of rural households and agriculture. The energy demands of towns and cities are obtained by distributing the urban demand in proportion to population.

Demand sector	Low			High		
	IV	V	VI	IV	V	VI
Villages						
Electricity	6.6	20.8	3.1	8.2	24.0	3.8
Contribution from wind, small-scale hydropower	0.4	1.5	0.2	0.5	1.7	0.2
Nonelectric commercial energy	14.1	76.7	19.3	15.2	82.1	21.7
Contribution from						
Charcoal	–	14.3	4.2	–	14.8	4.6
Biogas	–	14.3	4.2	–	14.8	4.6
Noncommercial energy	105.0	344.0	10.0	109.0	344.0	10.0
Towns						
Electricity	22.9	23.8	13.5	32.2	34.9	17.6
Contribution from wind, small-scale hydropower	0.5	0.6	0.3	0.8	0.9	0.4
Nonelectric commercial energy	183.6	201.7	97.5	252.3	271.1	131.5
Contribution from						
Charcoal	17.7	24.7	7.7	22.9	32.3	9.7
Soft solar	1.0	0.3	0.1	1.2	0.5	0.2
Noncommercial energy	4.0	–	–	–	–	–
Cities						
Electricity	56.0	50.5	36.5	78.9	74.3	47.5
Contribution from wind, small-scale hydropower	–	–	–	–	–	–
Nonelectric commercial energy	449.6	428.7	263.7	617.7	576.1	355.5
Contribution from						
Charcoal	43.3	52.5	20.9	56.1	68.7	26.2
Soft solar	2.5	0.7	0.4	2.9	1.0	0.5
Total						
Electricity	85.5	95.1	53.1	119.3	133.2	68.9
Contribution from wind, small-scale hydropower	0.9	2.1	0.5	1.3	2.6	0.7
Nonelectric commercial energy	647.3	707.1	380.5	885.2	929.3	508.7
Contribution from						
Charcoal	61.0	91.5	32.8	79.0	115.8	40.5
Soft solar	3.6	1.0	0.5	4.1	1.4	0.7
Biogas	–	14.3	4.2	–	14.8	4.6
Noncommercial energy	109.0	344.0	10.0	109.0	344.0	10.0

TABLE 12 Projections of demands for electricity, nonelectric commercial final energy, and noncommercial energy by villages, towns, and cities in the year 2030, and shares of S/D renewables (GWyr). The energy demands of villages are assumed to comprise requirements of rural households and agriculture. The energy demands of towns and cities are obtained by distributing the urban demand in proportion to population.

Demand sector	Low			High		
	IV	V	VI	IV	V	VI
Villages						
Electricity	19	65	11	26	88	17
Contribution from wind, small-scale hydropower	7	25	4	9	33	6
Nonelectric commercial energy	28	149	26	32	176	31
Contribution from						
Charcoal	–	–	–	–	–	–
Biogas	–	35	12	–	40	14
Noncommercial energy	87	344	10	98	344	10
Towns						
Electricity	45	42	26	72	84	53
Contribution from wind, small-scale hydropower	6	6	4	10	12	8
Nonelectric commercial energy	261	291	149	419	498	281
Contribution from						
Charcoal	37	52	23	57	87	40
Soft solar	4	3	2	6	6	4
Noncommercial energy	22	–	–	11	–	–
Cities						
Electricity	192	167	97	305	337	200
Contribution from wind, small-scale hydropower	–	–	–	–	–	–
Nonelectric commercial energy	1111	1163	560	1788	1991	1056
Contribution from						
Charcoal	159	207	85	244	347	149
Soft solar	18	12	7	26	24	13
Total						
Electricity	256	274	134	402	509	270
Contribution from wind, small-scale hydropower	13	31	8	19	45	14
Nonelectric commercial energy	1400	1603	735	2238	2665	1368
Contribution from						
Charcoal	196	259	108	302	433	188
Soft solar	22	15	9	33	30	16
Biogas	–	35	12	–	40	14
Noncommercial energy	109	344	10	109	344	10

TABLE 13 Share (%) of villages in commercial final energy demand.

Demand category	1975	2000		2030	
		Low	High	Low	High
Region IV					
Electricity	4.2	7.7	6.9	7.5	6.3
Nonelectric energy	3.0	2.2	1.7	2.0	1.4
Total final energy	3.1	2.8	2.3	2.8	2.2
Region V					
Electricity	8.9	21.9	18.0	23.9	17.2
Nonelectric energy	7.3	10.8	8.8	9.3	6.6
Total final energy	7.4	12.2	10.0	11.4	8.3
Region VI					
Electricity	4.1	5.8	5.5	8.1	6.3
Nonelectric energy	7.3	5.1	4.3	3.5	2.3
Total final energy	7.2	5.2	4.4	4.2	3.0
Regions IV + V + VI					
Electricity	6.1	13.1	11.2	14.4	11.0
Nonelectric energy	5.5	6.4	5.1	5.4	3.8
Total final energy	5.6	7.1	5.9	6.8	5.0

TABLE 14 Fraction (%) of commercial final energy demand met by the projected use of S/D renewables (High scenario).

Demand sector	2000			2030		
	Electricity	Nonelectric energy	Total final energy	Electricity	Nonelectric energy	Total final energy
Region IV						
Villages	6	0	2.1	35	0	15
Towns	2.3	10	9	14	15	15
Cities	–	10	8	–	15	13
Total	1.0	9	8	5	15	13
Region V						
Villages	7	36	29	37	22	27
Towns	2.5	12	11	14	19	18
Cities	–	12	11	–	19	16
Total	1.9	14	13	9	19	17
Region VI						
Villages	6	42	37	36	45	42
Towns	2.5	8	7	15	15	15
Cities	–	8	7	–	15	13
Total	1.0	9	8	5	16	14
Regions IV + V + VI						
Villages	7	33	27	37	22	27
Towns	2.4	10	9	14	17	16
Cities	–	10	9	–	17	14
Total	1.4	11	10	7	17	15

TABLE 15 Present and projected per capita consumption of electricity (W(e)yr yr^{-1}) as final energy delivered to consumers.

	1975	2030	
		Low	High
Region IV			
Total population average	77	321	504
Urban population average	123	349	555
Rural population average[a]	8	160	214
Region V			
Total population average	15	77	143
Urban population average	64	107	216
Rural population average[a]	1.7	41	55
Region VI			
Total population average	35	378	766
Urban population average	75	424	875
Rural population average[a]	2.6	170	269

[a]Comprises electricity consumption of rural households and of the agricultural sector; rural electrification is assumed to be 25% for Region IV and 15% for Regions V and VI in 1975, and 100% for all regions in 2030.

The supply of charcoal, the most important component of the S/D renewables, will be quite different. According to the present estimates the quantities of charcoal required in the High scenario will by 2030 amount to 302 GWyr yr^{-1} for Region IV, 433 GWyr yr^{-1} for Region V, and 188 GWyr yr^{-1} for Region VI (Table 12). If the efficiency of converting wood to charcoal is 45% (Earl 1975) and there are 5% losses in transportation of charcoal from production sites to towns and cities, the quantities of dry wood (in terms of the energy content of wood) required for meeting these demands will be 704, 1013, and 440 GWyr yr^{-1} for Regions IV, V, and VI respectively.

The regenerative capacity of natural forests in the developing regions, expressed in terms of the average annual increment of dry wood above ground, is estimated at about 3.5 t ha^{-1}, which corresponds to an annual energy production of about 1.95 kWyr ha^{-1} (Earl 1975). Thus the total regeneration in the natural forests of Regions IV, V, and VI amounts to about 2090, 1880, and 55 GWyr yr^{-1}, respectively. In view of the difficulties of access, transportation, management, and environmental safeguards, it is assumed that no more than about one-third of the natural forests in each of these regions would be harvested for producing energy. If concerted efforts are made it would perhaps be possible to reach such a level of exploitation within the next 50 years. This would then supply sufficient wood for the High scenario charcoal demand of Region IV, but not for those of Regions V and VI. In fact the extremely small forest area of Region VI (Table 3) would not be able to cope with even the estimated Low scenario demand in 2030 (108 GWyr yr^{-1}), even if the region's entire annual increment of wood were used for charcoal production.

Owing to the inadequate supply potential of their natural forests, Regions V and VI would have to resort to energy plantation schemes if they decided to use charcoal to the degree envisaged in this assessment. The available literature (e.g. Earl 1975, NAS 1976, Revelle 1979) appears to indicate that fairly high yields of wood (6–30 t ha^{-1} yr^{-1} or more)

TABLE 16 Resource utilization to achieve the projected S/D renewables commercial energy use in 2030. The values are expressed in terms of primary energy equivalents, using conversion efficiencies of 0.37 for electricity, 1.0 for soft solar, 0.45 for conversion of wood to charcoal, and 0.60 for conversion of agricultural and animal wastes to biogas.

	Maximum production capacity ($GWyr\ yr^{-1}$)	Capacity required ($GWyr\ yr^{-1}$)	
		Low	High
Small-scale hydropower			
IV	583[b]	23	33
V	761[b]	56	81
VI	68[b]	14	25
Windmill-generated electricity			
IV	N.A.	12	17
V	N.A.	28	41
VI	N.A.	7	13
Soft solar			
IV	N.A.	22	33
V	N.A.	15	30
VI	N.A.	9	16
Wood from forests			
IV	2090	458	704
V	1880	604	673
VI	55	18	18
Wood from plantations			
IV	N.A.	0	0
V	N.A.	0	340
VI	N.A.	234	421
Agricultural and animal wastes[a]			
IV	291–374	0	0
V	1054–1355	58	67
VI	98–126	20	23

[a]The production capacities correspond to 3.5–4.5 times the estimated production in 1975.
[b]These figures refer to total hydropower-generating capacity including centralized hydropower generation.
N.A.: Not available.

may be obtained from energy plantations by raising specific varieties of fast-growing trees on marginal farmland. If the average annual yield of dry wood is 15 $t\ ha^{-1}\ yr^{-1}$ (i.e. yielding about 8.4 $kWyr\ yr^{-1}$ in the form of wood or 3.8 $kWyr\ yr^{-1}$ as charcoal) energy plantations will be required to cover about 40×10^6 ha in Region V (High scenario only) and 28–50 $\times\ 10^6$ ha in Region VI. These figures should be compared with the present arable land areas in Regions V and VI, which amount to about 450×10^6 ha and 45×10^6 ha, respectively (Table 3). These are large operations, particularly for Region VI, but they may still be feasible if proper governmental support is provided to convert some of the permanent pastures and other land into energy farms.

Although our assumptions about the use of solar devices for space and water heating in buildings and hot-water/steam production for use in industry are optimistic, the share of soft solar energy in meeting the final energy demand in developing regions would not exceed much above 1% by 2030 (Table 12). If only manufacturing requirements of final energy are taken into consideration the share would be about 1.7% in each region. A rather high contribution of soft solar, at a level of about 5%, is also expected in the final energy demand of the household/service sector in Region IV where central heating is more common than in the other regions.

Table 16 reports both the estimated maximum production capacity in each of Regions IV, V, and VI and the capacity required in 2030 for meeting that part of the final energy demand in the two IIASA scenarios that is considered appropriate for S/D renewables in the present assessment.

7 CONCLUSION

The assessment in this report has shown that the soft/decentralized technologies based on renewable energy sources, if fully supported by national policy measures, may meet in 2030 about 7% of the electricity demand and about 17% of the nonelectric commercial final energy requirements of the developing regions. Our assumptions for identifying the potential markets for soft/decentralized renewables and for estimating the extent of their penetration into the appropriate potential markets within feasible limits (in our judgment) have been clearly stated. We believe that efforts to introduce these renewables at a higher scale would result in undue hardships to the users, and may also adversely affect economic development.

Of the electricity demand in 2030 in the developing regions about 13% would arise from the requirements of the rural population (including irrigation requirements) and about 17% from those of small towns. It is estimated that in 2030 about 37% of the electricity requirements of rural areas and some 14% in small towns may be supplied by decentralized, small hydropower units and windmills.

About 60% of the nonelectric final energy demand in the developing regions in 2030 would be for transportation, construction, mining, agriculture (mainly tractor fuel), and feedstock production activities (Khan and Hölzl 1981) and would be essentially all met by liquid fuels. The remaining 40% would be thermal requirements, which would define the main role of S/D renewables in the nonelectric sector. The use of commercial S/D renewables has been envisaged in 2030 to meet about 52% of the thermal requirements of the household/service sector and 36% of those attributed to manufacturing activities. In addition, the use of noncommercial fuels would meet about 18% of the household/service sector requirements of useful thermal energy in 2030.

The shares of biogas, soft solar, and charcoal in meeting the total nonelectric final energy demand of the developing regions in 2030 have been assessed as about 0.9, 1.3, and 14.7%, respectively, in the High scenario. (The shares are almost the same as in the Low scenario.) Thus the most important contribution would come from charcoal. The quantities of wood required in 2030 to produce the necessary amounts of charcoal would be in the range of 2.4–3.9 $\times$ 10^9 t for the two IIASA scenarios. These quantities should be seen against the total annual increment of wood in the forests of the regions, which is estimated as about 7.2 $\times$ 10^9 t. The situation is even more complicated if regional demands for charcoal and the regional resources of natural forests are considered separately. It turns out that Region VI would need to undertake intensive energy plantation even before the turn of the century, while Region V would be required to do the same later and only in the High scenario. Region IV, on the other hand, has sufficient resources of forests to meet its demand up to 2030 by utilizing not more than about one-third of the annual regenerative capacities of its forests, this being a practical upper limit, in our view, imposed by various constraints. If plantation activities are pursued in Regions V and VI, the land area under plantations in 2030 in the High scenario would be about 10% of the present arable land in Region V and roughly the same as the present arable area in Region VI. Region V, with very limited resources of fossil fuels, would probably have no other choice although the oil-rich Region VI may still consider it unnecessary to follow such a course.

The shares of biogas and soft solar in the total nonelectric final energy demand are rather low despite the incorporation into the assessment of some generous assumptions about their use. This is mainly a result of the limited sizes of their potential markets. Biogas is suitable for use only in rural areas, where a large fraction of the thermal energy requirements would still be met by noncommercial fuels even if the use of such fuels were assumed not to exceed the regional consumption levels of 1975. Soft solar is suitable only for water heating, space heating in detached centrally heated dwellings or low-rise buildings, and steam and hot-water production for manufacturing industries. The generally warm climates of the developing regions make their water- and space-heating requirements low compared with those in the developed regions, which mostly have cooler climates. The demand for hot water and steam by manufacturing industries accounts for only about 40% of their useful heat requirements. A considerable fraction of these industries are generally in or near major cities where scarcity of land precludes large solar installations for hot-water and steam generation.

In our opinion, this assessment of the possible use of soft/decentralized renewables is based on quite optimistic, although still not unrealistic, assumptions. It would call for a well organized, large, and persistent effort on the part of the developing regions if the use of renewables to the extent envisaged were to become a reality within, say, the next 50 years. The resource conditions, with respect to both conventional fuels and renewables, are not the same in all regions. There will even be large variations within each region, if individual countries are taken into consideration. It may well be that countries with abundant resources of oil, gas, or coal or with large potentials of centralized hydropower generation consider it unnecessary to change to soft/decentralized renewables to any significant extent in the next few decades. The unavailability of investment funds from individuals and small groups or unfavorable loan policies of governments may also retard the introduction of S/D renewables in areas where the resource conditions are most favorable to their use. These considerations only tend to lower the share of S/D renewables in meeting the future energy demands. Our present estimates should, therefore, be taken as an upper limit under generally favorable conditions.

ACKNOWLEDGMENTS

The author would like to express his sincere thanks to Professor W. Häfele for his valuable advice and encouragement during the course of this work. Useful discussions with several persons at IIASA, in particular with P. Basile, J. di Primio, A. Hölzl, J. Parikh, W. Sassin, and B. Spinrad, are also gratefully acknowledged.

REFERENCES

Cecelski, M., J. Dunkerley, and W. Ramsay (1979) Household Energy and the Poor in the Third World. Washington, D.C.: Resources of the Future.

Chant, V.G. (1981) Two global scenarios: Evolution of energy use and the economy to 2030. Laxenburg, Austria: International Institute for Applied Systems Analysis (forthcoming).

Earl, D.E. (1975) Forest Energy and Economic Development. Oxford: Clarendon Press.

Energy Systems Program Group of the International Institute for Applied Systems Analysis (1981) Energy in a Finite World: vol. 2, A Global Systems Analysis. Cambridge, MA: Ballinger.

Food and Agriculture Organization (1977) Production Yearbook vol. 30, 1976. Rome: FAO.

Goldemberg, J. (1979) Global options for short-range alternative energy strategies. Paper presented at the Conference on Energy Alternatives, The UN University, East–West Centre, Honolulu, HI, 9–12 January 1979.

Hölzl, A. (1981) A computer program for energy demand analysis by the MEDEE approach: Documentation of the MEDEE-2 version used in IIASA's Energy Systems Program. Appendix to Khan and Hölzl (1981). Laxenburg, Austria: International Institute for Applied Systems Analysis.

Khan, A.M., and A. Hölzl (1981) Energy demand to 2030 for six world regions: Estimates for two IIASA scenarios. Laxenburg, Austria: International Institute for Applied Systems Analysis (forthcoming).

Lapillonne, B. (1978) MEDEE 2: A model for long-term energy demand evaluation. RR-78-17. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Makhijani, A., and A. Poole (1975) Energy and Agriculture in the Third World. Cambridge, MA: Ballinger.

National Academy of Sciences (1976) Energy for Rural Development – Renewable Resources and Alternative Technologies for Developing Countries. Washington, D.C.: NAS.

Parikh, J. (1978) Energy systems and development – Constraints, demand and supply of energy for developing countries. Report to the World Bank, P.U. PUN 43. Washington, D.C.: World Bank.

Revelle, R. (1979) Energy sources for rural development. Paper presented at the Conference on Energy Alternatives, The UN University, East–West Centre, Honolulu, HI, 9–12 January 1979.

United Nations (1976) Compendium of Housing Statistics 1972–1974. New York, NY: UN.

United Nations (1977) World Energy Supplies 1971–1975. New York, NY: UN.

World Bank (1975) Rural Electrification. Washington, D.C.: World Bank.

World Energy Conference (1978) World Energy Resources 1985–2020. Executive Summaries of Reports on Resources, Conservation, and Demand to the WEC Conservation Commission. Guildford: IPC.

APPENDIX THE THREE MARKET-ECONOMY DEVELOPING REGIONS OF IIASA'S ENERGY SYSTEMS PROGRAM

Region IV: Latin America (LA)

Developing economies with some energy resources and significant population growth.

Argentina	El Salvador	Nicaragua
Bahamas	Guadeloupe	Panama
Belize	Guatemala	Paraguay
Bolivia	Guyana	Peru
Brazil	Haiti	Puerto Rico
Chile	Honduras	Surinam
Colombia	Jamaica	Trinidad and Tobago
Costa Rica	Martinique	Uruguay
Cuba	Mexico	Venezuela
Dominican Republic	Netherlands Antilles	Other Caribbean nations
Ecuador		

Region V: Africa (except Northern Africa and South Africa) and South and Southeast Asia (Af/SEA)
Slowly developing economies with some energy resources and significant population growth.

Africa

Angola	Kenya	Rwanda
Benin	Lesotho	Senegal
Botswana	Liberia	Sierra Leone
Burundi	Madagascar	Somalia
Cameroun	Malawi	Sudan
Cape Verde	Mali	Swaziland
Central African Republic	Malta	Tanzania
Chad	Mauritania	Togo
Congo	Mauritius	Tunisia
Ethiopia	Morocco	Uganda
Gabon	Mozambique	Upper Volta
Gambia	Namibia	Western Sahara
Ghana	Niger	Zaire
Guinea	Nigeria	Zambia
Guinea-Bissau	Réunion	Zimbabwe
Ivory Coast		

Asia

Afghanistan	Indonesia	Philippines
Bangladesh	Korea, South	Singapore
Brunei	Macao	Sri Lanka
Burma	Malaysia	Taiwan
Comoros	Nepal	Thailand
Hong Kong	Pakistan	Timor
India	Papua New Guinea	West South Asia, not elsewhere specified

Region VI: Middle East and Northern Africa (ME/NAf)
Developing economies with large energy resources.

Members of the Organization of Arab Petroleum Exporting Countries		*Others*
Algeria	Libya	Iran
Bahrain	Qatar	Jordan
Egypt	Saudi Arabia	Lebanon
Iraq	Syria	Oman
Kuwait	United Arab Emirates	Yemen
		Yemen, South

NONPROCEDURAL COMMUNICATION BETWEEN USERS AND APPLICATION SOFTWARE

Bořivoj Melichar
International Institute for Applied Systems Analysis, Laxenburg, Austria

SUMMARY

This report is a survey of nonprocedural communication between users and application software in interactive data-processing systems. It includes a description of the main features of interactive systems, a classification of the potential users of application software, and a definition of the nonprocedural interface. Nonprocedural languages are classified into a number of broad groups and illustrated with examples. Finally, future trends in user–computer interfaces and possible developments in manager-oriented languages are discussed.

1 INTRODUCTION

1.1 Basic Concepts

Advances in semiconductor technology during the past decade have dramatically increased the availability of low-cost computer hardware. One of the results of this greater availability has been the development of cheap but powerful small-scale computer systems.

According to Fick (1980), the power of computer systems has recently been doubling every two years, while the price of computer systems has remained approximately constant. With the "real" cost of computing capability declining, it is nevertheless apparent that low-cost computer hardware does not necessarily mean "cheap" computing – the cost of the software should also be considered. Computer software is a labor-intensive product, generally designed specifically for a small group of users or even for an individual (Fick 1980), and it is therefore more expensive than the mass-produced hardware. There are a number of areas in which development should take place if low-cost hardware is to be matched with suitable software, and these areas are outlined below.

1. Development of theoretical and methodological tools for software design in different fields of application.

2. Development of tools for software realization (programming languages, automatic program generation, program debugging and verification, etc.).
3. Development and production of media for software distribution (semiconductor read-only memories, magnetic tapes, magnetic discs, punched cards, punched paper tapes, books, journals, etc.).
4. Development of means for communication between users and the application software (input/output devices, new languages, etc.).

In this article we survey problems of communication between users and their application software. The manner of communication between users and application software is highly dependent on the users' access to the computer system. In recent years there has been much discussion of the issue of "indirect access" vs. "direct access", e.g., batch-processing systems vs. time-sharing systems. The communication between user and computer is very slow in a batch-processing system; the user can neither influence the way in which the program is run nor intervene while it is being run. The issue of batch-processing systems vs. time-sharing systems has therefore been resolved in favor of time-sharing systems. This means that communication between user and computer is now generally interactive in nature. The rest of this paper assumes the use of interactive systems and discusses the most interesting features of these systems.

1.2 Interactive Systems

Many of the problems associated with batch-processing systems may be overcome by the opportunity to communicate directly with the computer using an interactive system. However, the use of interactive systems has helped to create various new problems, which are now receiving considerable attention; the main requirements of interactive systems and the basic principles of the interactive dialogue are still under debate (Miller and Thomas 1977, Watson 1976, Fitter 1979, Gaines and Facey 1976). Here we provide a list of facilities which we think could and should be provided by interactive systems and some principles which should be followed in an interactive dialogue. Some of the following points may also be relevant to a batch-processing (noninteractive) environment, but we consider only their importance in interactive systems.

System response time. System response time is the time spent in processing the input and in producing a response. It is difficult to know exactly how long or short the response time should be, and there is no general agreement on this subject. There are several arguments for short response times:

- human response times are of the order of two seconds
- long response times decrease throughput
- long delays are usually disruptive and disturbing

On the other hand, there are arguments against short response times:

- short response times require high investment in the system
- long response times might be helpful for more complex tasks
- fast responses may encourage users to expect the same level of service at all times

NONPROCEDURAL COMMUNICATION BETWEEN USERS AND APPLICATION SOFTWARE

Bořivoj Melichar
International Institute for Applied Systems Analysis, Laxenburg, Austria

SUMMARY

This report is a survey of nonprocedural communication between users and application software in interactive data-processing systems. It includes a description of the main features of interactive systems, a classification of the potential users of application software, and a definition of the nonprocedural interface. Nonprocedural languages are classified into a number of broad groups and illustrated with examples. Finally, future trends in user–computer interfaces and possible developments in manager-oriented languages are discussed.

1 INTRODUCTION

1.1 Basic Concepts

Advances in semiconductor technology during the past decade have dramatically increased the availability of low-cost computer hardware. One of the results of this greater availability has been the development of cheap but powerful small-scale computer systems.

According to Fick (1980), the power of computer systems has recently been doubling every two years, while the price of computer systems has remained approximately constant. With the "real" cost of computing capability declining, it is nevertheless apparent that low-cost computer hardware does not necessarily mean "cheap" computing – the cost of the software should also be considered. Computer software is a labor-intensive product, generally designed specifically for a small group of users or even for an individual (Fick 1980), and it is therefore more expensive than the mass-produced hardware. There are a number of areas in which development should take place if low-cost hardware is to be matched with suitable software, and these areas are outlined below.

1. Development of theoretical and methodological tools for software design in different fields of application.

2. Development of tools for software realization (programming languages, automatic program generation, program debugging and verification, etc.).
3. Development and production of media for software distribution (semiconductor read-only memories, magnetic tapes, magnetic discs, punched cards, punched paper tapes, books, journals, etc.).
4. Development of means for communication between users and the application software (input/output devices, new languages, etc.).

In this article we survey problems of communication between users and their application software. The manner of communication between users and application software is highly dependent on the users' access to the computer system. In recent years there has been much discussion of the issue of "indirect access" vs. "direct access", e.g., batch-processing systems vs. time-sharing systems. The communication between user and computer is very slow in a batch-processing system; the user can neither influence the way in which the program is run nor intervene while it is being run. The issue of batch-processing systems vs. time-sharing systems has therefore been resolved in favor of time-sharing systems. This means that communication between user and computer is now generally interactive in nature. The rest of this paper assumes the use of interactive systems and discusses the most interesting features of these systems.

1.2 Interactive Systems

Many of the problems associated with batch-processing systems may be overcome by the opportunity to communicate directly with the computer using an interactive system. However, the use of interactive systems has helped to create various new problems, which are now receiving considerable attention; the main requirements of interactive systems and the basic principles of the interactive dialogue are still under debate (Miller and Thomas 1977, Watson 1976, Fitter 1979, Gaines and Facey 1976). Here we provide a list of facilities which we think could and should be provided by interactive systems and some principles which should be followed in an interactive dialogue. Some of the following points may also be relevant to a batch-processing (noninteractive) environment, but we consider only their importance in interactive systems.

System response time. System response time is the time spent in processing the input and in producing a response. It is difficult to know exactly how long or short the response time should be, and there is no general agreement on this subject. There are several arguments for short response times:

- human response times are of the order of two seconds
- long response times decrease throughput
- long delays are usually disruptive and disturbing

On the other hand, there are arguments against short response times:

- short response times require high investment in the system
- long response times might be helpful for more complex tasks
- fast responses may encourage users to expect the same level of service at all times

It has been observed that the variability of the system response time can be very annoying to the user.

Rohlfs (1979) proposed that systems should be designed so that their response time may be adjusted to the activity of the individual user:

>15 sec	intolerable
>4 sec	too long in most cases, possibly tolerable after termination of a major step
>2 sec	too long for very involved work
<2 sec	necessary for work consisting of more than one step
<1 sec	immediate reaction

Availability and reliability. The computer system should be available for use at any time; this would be possible if each user had his/her own computer. Because the user will be unhappy with any system performance error or degradation regardless of good normal performance, reliability is also a very important feature of an interactive system. For many computer applications almost no degradation or loss in availability can be tolerated.

Commonality. A software system is usually composed of a number of subsystems. In this case, the system should be organized such that terminology does not change between subsystems. This implies that the input language of each subsystem should be an extension of the common base language. Thus, the user will only need to learn additional functions or statements when using a new subsystem and not have to learn a completely new "foreign" language. When the user is in trouble, he/she can use standardized "help" functions to extricate him/herself from the situation.

Adaptability to user proficiency. It should be possible to design the interface between user and computer to suit users with different amounts of knowledge about a particular subsystem. A sophisticated user may prefer to use mathematical or formalized notation in his/her dialogue. On the other hand, a novice user is likely to prefer less formalized notation and use simpler system functions.

The newer systems have been adapted for users with various levels of proficiency by designing different user interfaces. As the user becomes more proficient, he/she can use more sophisticated functions or a more formalized interface.

Immediate feedback. A system should make an unambiguous response to each of the user's requests. This response should be sufficient to identify the activity and state of the system. In situations where system response times are longer than usual, it is highly desirable to confirm receipt of the user's command immediately. It is very useful to let the user know when the computer will produce a response, for example, by displaying a countdown clock on the terminal.

Observability and controlability. A system can be regarded as an automaton. It is important that the user should feel in control of the system, and in order to make this control possible he/she must have some knowledge about the current state of the system. Thus, when the user's input is processed, the user should be informed about the current state of the system. The display of this information may be regarded as a transition from one state to another.

Use the user's model. Everybody rationalizes their experiences in their own terms, and in the same way each user will model a computer system according to his/her experience of it. This cannot be prevented and should be made as easy as possible. The system should use a model of computer activity which corresponds to that perceived by the user, so that the interactive dialogue resembles a conversation between two users accepting the same model. Given that we can somehow determine the user's model of the computer system, we should make the underlying processes reflect it, and design the dialogue to reveal it as clearly as possible.

Validation. All input commands and data must be validated by checking syntax, semantics, and, if possible, values. The system must inform the user about any errors or ambiguities in the input data and let the user update the values in question before the system acts upon them.

Query-in-depth. Information and advice on the system should be categorized according to possible user requests and should be available to the user through a simple standard mechanism.

Extensibility. There will never be enough professional programmers and system developers to provide all the tools that users may desire for their work. It should therefore be possible for users to develop new tools or to extend the functions already present in the system.

Written documentation. In some cases it is necessary to produce high-quality documents as a result of interaction between user and terminal. Text processing is one of the most important examples.

System activities. It is necessary to maintain records of system performance and user's activities to evaluate and improve the behavior of the system.

1.3 Interface between the User and the Application Software

The nature of a user–application software interface is largely determined by the medium used for communication. The most basic means of communication are alphanumeric texts and graphics. More advanced methods of communication, such as speech, eye movement, brain-wave control, and handwritten script (Watson 1976) are currently being investigated.

In this survey we concentrate on alphanumeric texts as a medium for user–computer communication. It is assumed that a normal keyboard and alphanumeric display (with or without hard copy) are available to the user.

The user–software interface has two sides (Watson 1976, Sprague 1981): the input side, through which the user inputs information by means of an *action language*; and the output side, through which the computer provides feedback and other assistance to the user by means of a *display language*.

Let us first survey the action languages. A wide range of action languages has been designed to accommodate a wide variety of users. The selection of a particular action language determines the communication mode that should be used. We can classify action languages and/or communication modes as follows:

- low-level machine-like programming languages
- high-level universal programming languages
- high-level programming languages with new syntax and semantic forms (such languages can be used as special-purpose languages)

- self-contained special-purpose languages
- answer languages
- command languages
- query languages
- natural languages
- two-dimensional positional languages

This classification of communication modes covers the complete range from artificial machine-oriented languages to natural human-oriented languages.

Different types of user may wish to communicate with the computer in different ways, and so it is very important during software development and implementation to select the communication modes appropriate to the end-users.

According to Schneiderman (1978), we can categorize users into three groups with respect to their skills, the frequency with which they use application software, their degree of knowledge of the problem under investigation, and their professional fields.

1. *Nontrained intermittent users* who infrequently use application software. A user in this category is called a "casual user" by Codd (1977) and a "general user" by Miller and Thomas (1977). These people are not computer professionals, have no syntactic knowledge, and have little knowledge of the organization of the application software. At the same time it is assumed, however, that these users have sound professional knowledge in their own particular fields and, therefore, that the system should allow them to express themselves in their own specialized terminology (Lehmann 1978).
2. *Skilled frequent users* who make daily use of application software. These users can learn the simple syntax of a communication action language, but they are more interested in their own work than in computer programming. This category includes skilled secretaries, engineers, and managers.
3. *Professional users* whose main task is to develop and maintain the application software. They are highly trained in this field and are concerned with the efficiency and the quality of the computer system. This category includes database administrators and software system programmers.

Although programming and communicating with a computer in high-level programming languages like ALGOL, FORTRAN, COBOL, PL/1, and PASCAL represents a major advance over the use of machine-like low-level programming languages, high-level languages are becoming less appropriate now that cheap hardware is available. With the rapid diffusion of cheap computer hardware it is expected that the numbers of people in the nontrained intermittent and the skilled frequent user categories will grow very quickly. These users have little or no experience of data-processing, and would find it very difficult and very time-consuming to learn how an algorithm may be constructed and described in a programming procedure-oriented language. Therefore, it is highly desirable to find some means by which these users may communicate with application software in a nonprocedural manner.

There is very great interest in the development of nonprocedural languages, not only to facilitate communication between the user and the application software, but also in connection with the implementation of application software.

According to Winograd (1979), it is useful to distinguish three ways in which computational processes may be specified:

1. *Program specification.* A formal specification which can be interpreted as a set of instructions for a given computer. This is the imperative mode characteristic of traditional procedure-oriented programming languages.
2. *Result specification.* A process-independent specification of the relationship between the inputs (or initial state), internal variables, and outputs (or final state) of the program.
3. *Behavior specification.* A formal description of the activity of a computer over time. A description of this type selects certain features of the machine state and action without specifying in full the mechanisms which generate them.

We can divide an algorithm into two components (Kowalski 1979), a *logic component*, which specifies the knowledge to be used in solving the problem, and a *control component*, which determines the way in which the knowledge will be used. For example, consider an algorithm for computing factorials. The logic component of the algorithm is given by the definition of a factorial:

```
1 is the factorial of 0
u is the factorial of x if v is the factorial of
  x - 1 and
u is v times x
```

This is an example of result specification. For comparison, consider the following procedure in ALGOL 60:

```
procedure factorial (x); value x; integer x;
if x = 0 then factorial : = 1 else
              factorial : = factorial (x - 1)*x
```

In this procedure the logic component is blended with the control component.

According to McCracken (1978), we can characterize nonprocedural languages in two ways.

1. The user cannot control the storage of data. Decisions that relate only indirectly to the calculation are considered to be part of the internal functioning of the system. These include decisions about the internal representation of numbers (fixed point, floating point, octal, decimal), dimensions of quantities that occur only as intermediate results, and input and output formats. The representation of data is selected by the system itself, and the description of the data representation is stored with the data. This is called *data independence*.
2. The user cannot tell the computer *how* to obtain the desired results. Rather, he/she tells the computer only *what* he/she wants. This means that the user does not become involved in the loops and branches which make up most of the computational steps in a program written in procedural language. This we can call *control independence*.

The following query on the data stored in a data base provides a nice illustration of the nonprocedural approach.

```
RETRIEVE (AGE > 40 AND < 65) AND SALARY >3000;
FOR EACH
      IF WEIGHT > TABLE (HEIGHT - 50)
      THEN SET OVERWEIGHT = "TRUE"
      PENSION = SALARY/3;
      ELSE SET PENSION = SALARY/2
```

We can now give a working definition of a nonprocedural language:

> In a nonprocedural language the computational process is specified by the desired result (or behavior). This specification is data independent and control independent.

We shall consider a nonprocedurally *oriented* language to be a language which does not fulfill all of the conditions necessary for classification as a nonprocedural language, but which does not require program specification. Of the nine communication modes listed earlier, we can consider answer languages, command languages, query languages, natural languages, two-dimensional positional languages, and some special-purpose languages as nonprocedurally oriented action languages.

In the next section we examine the nonprocedural action languages available for communication between nontrained intermittent or skilled frequent users and application software. Communication in the reverse direction (from the system to the user) takes place through display languages, the main features of which are described in Section 3.

2 NONPROCEDURALLY ORIENTED ACTION LANGUAGES

In the last section we defined nonprocedural action languages and sketched the arguments for using them for user–application software communication. In this section we shall give the basic characteristics of each type of language and illustrate them with examples taken from the literature.

2.1 Answer Languages

An answer language is the set of words, phrases, or sentences which may be used to answer questions asked by the computer. This type of language is introduced first because it is the simplest means of user–computer communication.

The answer languages used as action languages are very closely related to display languages. The display language in this case contains, among other things, the set of questions which are asked and which the user is obliged to answer. We can categorize answer languages with respect to the number of different questions that can be answered in each user response: this may be one, or more than one. Languages in which the user may answer only one question at a time can be divided into three classes: binary answer systems, menu selection systems, and instruction and response systems.

Binary answer systems only recognize two answers, YES and NO, often represented by the abbreviations Y and N. The binary answer languages are used in software systems in which the internal structure corresponds to a binary oriented graph. In the binary oriented graph two branches leave each edge. Edges correspond to states of the system; in each such state the system asks a question, and according to the answer (NO or YES) a branch is selected which leads to a new state. The binary answer language is used mostly in simple systems such as computer games.

As an example we use the popular game blackjack (Thompson and Ritchie 1975). The dealer (simulated by the computer) might ask the following questions:

```
NEW GAME?

HIT?

INSURANCE?

SPLIT PAIR?

DOUBLE DOWN?
```

Each question is answered by YES or NO.

It is clear that binary answer language may be used only in systems in which a limited number of questions may be asked. For cases in which the answers YES or NO are not sufficient to answer all the questions which may arise, we can use a *menu selection system* (n-ary answer language).

In a menu selection system the set of possible answers to each question is defined. Each set must be finite, and from a practical point of view should consist of only a small number of answers.

The set of answers to a particular question is called the "menu". There are two ways in which the menu may be presented to the user. First, the menu of answers for each question may be given in the description of the software system, and the user is thus obliged to learn these menus before using the system. Much better is the second way, in which the menu and the question are provided together. This method has a number of advantages, the most important of which is that the user need not learn the menus for all possible questions.

Menu selection systems, like binary answer languages, are used in software systems in which the internal structure corresponds to an ordinary oriented graph. In such a graph a varying number of branches leave from each edge, the edge representing the question and the branches corresponding to the set or "menu" of possible answers.

As an example we use some menus from Teitelman (1979), who describes a system designed to help the user to develop programs. In this system, for example, the user may be presented with the following choice:

```
MENUS:
WINDOW
DOCUMENT
EDIT
LOOK
HISTORY
BREAK
OPERATIONS
```

This menu is then used to select further menus.

```
WINDOW:

READ

MOVE

GROW

SHRINK

PUT ON TOP

PUT ON BOTTOM

KILL

MAKE INVISIBLE
```

```
EDIT:

INSERT

APPEND

DELETE

REPLACE

MOVE

--->

<---

DONE
```

Questions with menus are displayed on a screen, and the user can select an answer by means of the cursor.

When the number of possible answers to a particular question is very large, it is inefficient (or sometimes impossible) to display all possible answers with the question. This situation may arise if the answer contains a number. In such cases we may use an *instruction and response system*.

In an instruction and response system an explanation of the answer required is included in the question. The following example of an instruction and response dialogue is taken from Hebditch (1979).

```
ORDER OR CREDIT? O

CUSTOMER NUMBER? 848923

CUSTOMER IS BROWN'S STATIONERS LTD.

            HIGH STREET

            WATFORD

PLEASE CONFIRM (Y/N)? Y

DELIVERY ADDRESS IS AS ABOVE

PLEASE CONFIRM (Y/N)? Y
```

```
ORDER NUMBER? 77/34
DISCOUNT? 12.5
***12.5 PERCENT IS HIGHER THAN NORMAL TERMS
PLEASE CONFIRM BY RE-ENTRY? 12.5
ENTER PRODUCT CODE. QUANTITY (END AFTER
LAST ITEM)
?B04,24   24 DOZ PENCILS (HB)
?A68,10   10 REAMS BANK PAPER
?B61,36   36 DOZ BALL-POINT PENS
***36 DOZ IS ABNORMAL QUANTITY FOR THIS ITEM.
PLEASE CONFIRM
?B61,3   3 DOZ BALL-POINT PENS (BLUE)
?Z15,1   1 DISPLAY STAND (BALL-POINT PENS)
?END     ORDER COMPLETED (4 ITEMS)
DO YOU WISH TO SEE INVOICE PRIOR TO PRINTING
(Y/N)? N
```

Systems in which a user can answer more than one question at a time require some type of fixed format user input. This format can be used, for example, to separate single answers in user input, and may be described in the question. Such a system is called a *displayed format system*.

The following simple and self-explanatory example of a book-ordering system (Hebditch 1979) shows a question containing the format description and the resulting answer.

```
ENTER AUTHOR / TITLE / PUBLISHER / ISBN /
      NO. OF COPIES / CUSTOMER NAME /
      CUSTOMER ADDRESS / POST OR COLLECT?
```

```
HEBDITCH / DATA COMMUNICATIONS: AN
INTRODUCTORY GUIDE / ELEK SCIENCE LTD. /
NK / 4 / A. WISEMAN / NA / COLLECT
```

2.2 Command Languages

Command languages in one form or another have been in use since the earliest operating systems first came into existence in the late 1950s. The name "command languages" was used in the past to describe the job control languages used as interfaces between users and operating systems. Today the term includes a wide variety of languages employed as user–computer interfaces in many types of software systems.

A command language consists of a set of commands. A typical command is composed of four elements: the command prefix, the operation specification, the parameter part, and the command completion.

The first part of the command, the *command prefix*, contains

- a command indication (a symbol or string of symbols to distinguish the command from other inputs)
- a command identification (a label or number used for reference purposes in other commands)
- a condition, which must be satisfied before the command is executed (for example: `IF TIME < MAXTIME THEN`)

A command word is frequently reserved for use as an *operation specification* (Miller and Thomas 1977). Watson (1976) proposes an operation specification in the form of a verb–noun pair. In this case we obtain a verb–noun matrix as, for example, in an editing system:

	character	line	page
delete			
insert			
change			
move			

Each element of this matrix is a normal English command.

Displaying the operation specification in this form seems to be very helpful for users with no experience in data processing (Keen and Hackathorn 1979). Hebditch (1973) proposes a more structured format, using adjectives (J) in addition to nouns (N) and verbs (V), to create an operation specification with three forms:

```
V J N     (PRINT CONDENSED RECORD)
V N J     (FIND EMPLOYEES WITH A DEGREE)
V J N J   (PRINT ALL LINES BEGINNING WITH +)
```

Further, Hebditch (1973) proposes that a set of basic verbs could be used as an interface with a data base, as shown in Table 1.

TABLE 1 Computer functions, verbs commonly used to specify the function, and alternative specifications.

Function	Verb (abbreviation)	Alternative forms
Initiation	START (S)	Begin, Sign-on, Initiate, Go, Set-up, Evoke
Location of logical record	FIND (F)	Locate, Get, Search, Read, Obtain, Pick (good for inventory data base?)
Display of data item	PRINT (P) (for hard copy)	Display (for VDUs), Show, Query, Give, List, Present
Amendment of data item	ALTER (A)	Change, Amend, Modify, Replace, Convert, Set
Addition of new record or item	INSERT (I)	Add, New, Assign, Include, Originate, Form
Movement of record (or data) from one logical location to another	MOVE (M)	Transfer, Shift, Relocate, Convey, Reallocate, Transpose
Obtain assistance	EXPLAIN (E)	Assist, Why?, Expand, Clarify, Help, Interpret
Termination	HALT (H)	End, Finish, Done, Close, Terminate, Conclude

The *parameter part* specifies the operand and suggests various ways in which the command could be executed. There are two distinct methods of formatting the arguments for commands: *positional format* and *keyword format*.

In the positional format, particular pieces of data must appear in fixed relative or absolute positions in the parameter part.

In the keyword format the parameter part is a permutable string of arguments, each argument containing a *keyword* which indicates the argument type and, sometimes, its value.

Both types of argument format occur in current systems. From the user's point of view the keyword format is more acceptable, because it requires the user to memorize less information.

The arguments in the parameter part may be composed of several different elements; these may include keywords, constants (numerical, boolean, etc.), text strings, and expressions (regular, arithmetical, boolean, etc.).

There remains the question of what to do when the user does not specify some information that either could or should have been provided. There are several ways of prompting the user for missing information:

1. Listing the missing argument names with all their possible values so that the user may choose the correct values.
2. Assigning a *default* value automatically to some of the missing arguments and asking the user for agreement.
3. Supplying missing information on the basis of the arguments provided to previous commands.

The problem of choosing argument delimiters is related to argument specification. Delimiting functions may be delimiting command words, delimiting arguments, or delimiting optional arguments (arguments with default values) (Watson 1976). It is generally advisable to use the same symbol for delimiting command words and arguments and to use a different symbol for delimiting optional or default arguments.

The last part of the command is the *command completion*. According to Watson (1976), there are three types of command completion:

1. Command accept: a command completion indicating that the command should be executed and the system should then return to the base state to receive the next command.
2. Repeat: a command completion indicating that the command should be executed and the system should then return to an intermediate command state for quick repetition of the command with or without request. This mode is useful when an operation must be repeated several times.
3. Insert: a command completion indicating that the command should be executed, the system should then enter insert-command mode for insertion of some new parameters, and then the command should be repeated.

A different symbol should be used for each type of command completion.

2.3 Query Languages

According to Olle (1973), there are four levels on which a user might come into contact with a data base. At the highest level is the data administrator, that is, the person responsible for the data base. The applications programmer occupies the second level, and at the next level is the application specialist, who is able to formulate questions about the data stored in the data base but is not a programmer. Finally, people who are unable to formulate questions occupy the lowest level in this hierarchy.

Data administrators and applications programmers generally use programming languages in their work. Users who are unable to formulate questions may use the simpler answer languages discussed previously. Query languages are designed to be used by the intermediate group of people, users who are not programmers but who understand how to formulate questions for a particular application.

Query languages are high-level nonprocedural data-base languages, which allow the user to perform operations such as insertion, deletion, and retrieval on the data base. Strong emphasis is placed on retrieval operations and, in view of this fact, a finer categorization of retrieval operations seems appropriate (Schneiderman 1978). There are four main retrieval operations which may be performed on a data base:

1. Simple verification of the presence, absence, or acceptability of a specified item.
2. Retrieval of a single record when a key is provided.
3. Retrieval of a number of records when a key or boolean predicate is provided.
4. Total listing of all information stored.

The list below indicates how query languages may be used to sort and retrieve data, and gives some examples of the type of queries which may be asked (Schneiderman 1978).

1. Simple mapping produces data values from one field when a known value for another field is supplied. Example: Find the names of all employees in department 50.
2. It is possible to select all of the data associated with a specified key. Example: Give the entire record for the employee whose name is John Jones.
3. In a relational model it is possible to select any domain of a relation. Example: Print the names of all employees.
4. Boolean queries permit AND/OR/NOT connections. Example: Find the names of those employees who work for Smith and who are not employed in department 50.
5. Set operation queries involve set operations such as intersection, union, and symmetric difference. Example: Find the names of the employees who work for Smith and the addresses of the employees who work for Black.
6. Built-in functions such as MAXIMUM, MINIMUM, AVERAGE, SUM, make it easier for the user to formulate questions. Example: Print the sum of salaries of employees in department 50.
7. Combination queries are produced by using the output of one query as the input for another. Example: Find the names of all departments which have more than 30 employees and then find the names of the department managers.

8. It is possible to group items with a common domain value. Example: Print the names of departments in which the average salary is greater than $15,000.
9. Universal quantification corresponding to the "for all" (∀) concept of the first-order predicate calculus. Example: Find the addresses for all employees.

The features listed above are available in most query languages designed for data bases using relational, hierarchical, or network models of data.

As an example, consider the following data base, which is built on a relational model (Chamberlin 1976):

PRESIDENTS

NAME	PARTY	HOME-STATE
Eisenhower	Republican	Texas
Kennedy	Democrat	Massachusetts
Johnson	Democrat	Texas
Nixon	Republican	California
Carter	Democrat	Georgia
Reagan	Republican	California

This relation PRESIDENTS has domains NAME, PARTY, and HOME-STATE.

ELECTIONS-WON

YEAR	WINNER-NAME
1956	Eisenhower
1960	Kennedy
1964	Johnson
1968	Nixon
1972	Nixon
1976	Carter
1980	Reagan

This relation ELECTIONS-WON has domains YEAR and WINNER-NAME. The two relations PRESIDENTS and ELECTIONS-WON are the only relations in our sample data base.

According to Chamberlin (1976), there are four classes of query language: relational calculus-based languages; relational algebra-based languages; mapping-oriented languages; and graphics-oriented languages. Languages in the first three categories may be distinguished by their mathematical basis. The fourth category includes certain two-dimensional languages.

One example of a *relational calculus-based* query language is the language QUEL (Stonebraker *et al.* 1976). A typical query in QUEL has three parts:

- a *result name*, which is the name of the relation from which data will be retrieved
- a *target list*, which specifies the particular domains of the relation from which data will be retrieved
- a *qualification*, which specifies certain conditions that the retrieved data must fulfill

A QUEL interaction must include at least one RANGE statement to specify the relation over which each variable ranges. Two examples of queries in QUEL are given below.

1. What was the home state of President Kennedy?

```
RANGE OF P IS PRESIDENTS

RETRIEVE INTO X (STATE = P.HOME-STATE)

WHERE P.NAME = "KENNEDY"
```

2. List the years in which a Republican from Illinois was elected President!

```
RANGE OF E IS ELECTIONS-WON

RANGE OF P IS PRESIDENTS

RETRIEVE INTO Y (YEARS = E.YEAR)

WHERE P.PARTY = "REPUBLICAN"

  AND P.HOME-STATE = "ILLINOIS"

  AND P.NAME = WINNER-NAME
```

Relational algebra-based query languages use a variety of operators that deal with relations, yielding new relations as a result. Among the most important of these operators are projection, restriction, and set-theory (union, intersection, and symmetric difference) operators. Translating the two queries given above into relational algebra-based query language we obtain:

1. What was the home state of President Kennedy?

```
PRESIDENTS [NAME = "KENNEDY"] [HOME-STATE]
```

The above example uses projection and restriction operators.

2. List the years in which a Republican from Illinois was elected President!

```
(ELECTIONS-WON [WINNER-NAME = NAME] PRESIDENTS)
[PARTY = "REPUBLICAN"] [HOME-STATE =
"ILLINOIS"] [YEAR]
```

In this example we use union, projection, and restriction operators.

The basis of *mapping-oriented* query languages is the operation of "mapping", in which a known domain or set of domains is "mapped" into a desired domain or set of domains by means of some relation. Our two examples are now in the mapping-oriented language SEQUEL (Astrahan *et al.* 1976).

1. What was the home state of President Kennedy?

```
SELECT HOME-STATE
FROM PRESIDENTS
WHERE NAME = "KENNEDY"
```

2. List the years in which a Republican from Illinois was elected President!

```
SELECT YEAR
FROM ELECTIONS-WON
WHERE WINNER-NAME =
      SELECT NAME
      FROM PRESIDENTS
      WHERE PARTY = "REPUBLICAN"
      AND HOME-STATE = "ILLINOIS"
```

Graphics-oriented query languages are mentioned later in this survey in the section dealing with two-dimensional positional languages.

The mapping- and graphics-oriented query languages are designed for users with no experience in data-processing and offer power equivalent to relational algebra- or relational calculus-based languages while avoiding difficult mathematical concepts.

2.4 Natural Languages

The idea of communicating with computers using a natural language has provoked much discussion from the early years of machine translation. However, though this concept is obviously very attractive to the user, the implementation of a natural language interface presents considerable difficulties to the programmer.

Natural language is the technique of verbal communication between people. According to Addis (1977), natural languages have an extremely complex structure because they reflect the way in which people think.

The use of a natural language for user–computer communication has several major advantages (Infotech International Ltd. 1979).

1. It provides a familiar way of forming questions. This means that the natural language interface would be available to a large number of users without the need for special training.
2. There are often many ways to extract the same data. The user can usually communicate his/her knowledge in a natural language augmented by specific notation and vocabulary characteristic of his/her specialist field.
3. It may be easier to formulate complicated queries using a natural language than using formal languages or menu selection methods.
4. The user does not have to learn a formal syntax and his/her departures from accepted grammar may be tolerated without comment.

At the same time we must note the following disadvantages:

1. The use of a natural language interface encourages an unrealistically high expectation of system power.
2. The linguistic limitations of such a system are not as well-defined as they are within a formal language. Confusion can arise as the result of an unknown word, an unknown grammatical construction, or a misunderstanding.
3. Sentences in natural languages are frequently ambiguous. Implementation is difficult if all the possible meanings of a sentence must be considered.
4. Because much of the vocabulary may be specific to the particular application, the system has to be partially recast to deal with each new application.
5. A system with a natural language interface is inherently much more complicated to implement than systems using more formal languages.

Besides these disadvantages there is one additional problem: the great variety of natural languages in the world. These problems do not imply that a natural language interface is useless, only that it is not as generally applicable as is sometimes thought.

Natural language communication with computers falls into two major categories (Kraegeloh and Lockemann 1978): communication with systems designed to understand natural language as part of the development of artificial intelligence, and communication with data-base systems.

In developing artificial intelligence the main goal is for the system to understand a natural language by the simulation of natural language perception. The resulting systems are highly complex, because the semantics of the natural language must remain unrestricted. One of the characteristics of Man is his ability to produce a mental image of his environment, a model of the real world. During a conversation, statements about his environment are related to this cognitive model and can modify it. Simulation of this process requires some system in which the real world, or part of it, may be modeled in a computer.

One system capable of understanding a natural language is GUS (Genial Understander System), which is described in more detail by Bobrow (1977). This system is restricted to the role of a travel agent in a conversation with a client who wants to make a trip to a single city in California. The following is a simple dialogue between a client and GUS.

```
GUS:     Hello. My name is GUS. I can help you
         plan a simple trip by air. Where do
         you want to go?
Client:  I want to go to San Diego on May 28.
GUS:     What time do you want to leave?
Client:  I must be in San Diego before 10 am.
```

GUS:	Would you like P.S.A. flight 102 that arrives at 9.15 am?
Client:	What is the next flight?
GUS:	Air California flight 310 that leaves at 8.30 am and arrives at 10.00 am.
Client:	I'll take the first one.
GUS:	In what name should I make a reservation, please?
Client:	Dan Bobrow.
GUS:	I have confirmed the following flight: P.S.A. flight 102 on Wednesday May 28 from San Jose to San Diego that leaves at 7.30 am and arrives at 9.15 am. What date do you want to return on?
Client:	On Friday in the evening.
GUS:	Would you like the flight that leaves at 7.45 pm?
Client:	That's fine.
GUS:	I have confirmed the following flight: P.S.A. flight 307 on Friday May 30 from San Diego to San Jose that leaves at 7.45 pm and arrives at 9.30 pm. Thank you for calling. Goodbye.

The second important area in natural language communication with computers is the use of natural languages to provide an interface between the user and data-base systems. This is a subject under much discussion because it could provide easy data access to users unfamiliar with query languages. In this case the computer's understanding of a natural language is not based on a simulation of the cognitive process; the data base itself serves as a model for understanding. Consequently, the input language should be restricted to the vocabulary necessary to use the data base. Under these conditions the natural languages used as interface languages in data-base systems may still be considered formal languages.

As an example of a natural language interface between user and data base we use the system PLANES (Waltz 1978). This system was developed with the following aims:

1. To accept the user's natural English
2. To provide explicit, clear answers to questions (possibly using graphics)
3. To tolerate minor errors (such as misspelling and grammatical errors)
4. To use clarifying dialogue for several purposes:
   - to increase its understanding of the user's request
   - to ask the user about parts of the request which it does not understand
   - to add new words, phrases, and sentences to its vocabulary
   - to provide appropriate assistance to the user
   - to provide information about the capabilities of the system
5. To be convenient to use, i.e.,
   - to be on-line and interactive
   - to operate rapidly
   - to require a minimum of typing
   - to answer a substantial percentage of questions
6. To be relatively easy to develop further

An ordinary dialogue between a user and the PLANES system is given below.

```
PLEASE ENTER YOUR QUESTION....................

>>How many flights did plane 3 make in Jan 73?
```

PLANES searches the `MONTHLY FLIGHT AND MAINTENANCE SUMMARIES` and paraphrases the answer required as the value of `TOTAL FLIGHTS` for plane `SERIAL #3` during January 1973.

```
SHOULD I EVALUATE THE QUERY? TYPE Y OR N
>>y
EVALUATING..................................
(SUM TOTFLTS) = 17
PLEASE ENTER YOUR QUESTION...............
```

For our second example we give a clarifying dialogue involving a spelling correction.

```
>>Which Skyhawks logged less than 10 hours?
PLEASE ENTER A SPECIFIC TIME PERIOD
>>Between Jan. 1 and Jule 1 '73
I DON'T KNOW THE MEANING OF JULE.
PERHAPS IT'S MISSPELLED?
PLEASE ENTER A NUMBER CORRESPONDING TO THE
FOLLOWING:
1.   JUNE
2.   JULY
3.   NONE OF THE ABOVE
>>2.
O.K.
```

A natural language may not provide the best interface in every situation. Schneiderman (1978) describes a "natural versus artificial query language experiment" concerning communication with data bases and concludes that the user must have some knowledge of the application domain if a natural language interface is to be used; i.e., the user has to be familiar with the semantics of the information in the data base. When a user learns a query language, he/she automatically learns the semantics of the information stored in the data base at the same time.

2.5 Special-Purpose Languages

In some fields, specific formalisms are used to describe particular problems. It seems reasonable to use these formalisms directly as *special-purpose languages* to interface with specialized software systems.

As an example, we consider one class of formalisms used widely for language design and implementation. Special-purpose languages based on formalisms of this type are used as interfaces in written translation systems. These languages are based on the idea of a context-free grammar, i.e., a set of rules of the form:

left part : right part

where the *left part* is one nonterminal symbol called a syntactic category and the *right part* is a string of nonterminal and terminal symbols.

The way in which sentences, composed of terminal symbols, may be derived from one particular nonterminal symbol, known as the start symbol, is first defined. The set of all sentences which can be derived from the start symbol may be described as a formal language generated by the grammar.

The following example illustrates the use of a language based on context-free grammar. In this case nonterminal symbols are represented by mnemonic names between angular brackets ⟨ ⟩; the terminal symbols are 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and /. The rules are:

```
<date>   : <number>/<number>/<number>
<number> : <digit>
<number> : <number>/<digit>
<digit>  : 1
<digit>  : 2
<digit>  : 3
         .
         .
         .
         .
<digit>  : 9
<digit>  : 10
```

The start symbol is <date>, and we shall use the symbol → to represent one step in the derivation.

```
<date> ——> <number>/<number>/<number>
       ——> <digit> /<number>/<number>
       ——>    1    /<number>/<number>
       ——>    1    /<digit> /<number>
       ——>    1    /   2    /<number>
       ——>    1    /   2    /<number><digit>
       ——>    1    /   2    /<digit><digit>
       ——>    1    /   2    /    7<digit>
       ——>    1    /   2    /    79
```

The language generated by this grammar is a set of sentences of the form number/number/number/, which can be read as dates.

Context-free grammars are often used to describe the syntax of formal and natural languages and, as mentioned above, they can also be used as a basis for the special-purpose languages used in written translation systems. The following is an example of text input to the YACC written translation system (Johnson and Lesk 1978).

```
% token DIGIT
%%
date : number '/' number '/' number
          {date ($1, $3, $5) ;};
number : DIGIT
          {$$ = $1;};
number : number DIGIT
          {$$ = 10 * $1 + $2;};
```

The nonterminal symbols in this grammar are `date` and `number`; the terminal symbols are `DIGIT` and /. A program fragment is given at the end of each grammar rule, and these program fragments compute the meaning or value of the nonterminal symbols. The variable $$ refers to the nonterminal symbol on the left-hand side of each rule, while $1, $2, . . . , $*n* refer to the first, second, or *n*th symbols on the right-hand side of the rule, respectively.

This input text may then be processed by a YACC parser generator, which generates a program able to read dates, convert them into a suitable form, and store them in the computer, provided that the digits are first read by another program returning the value of each digit.

2.6 Two-Dimensional Positional Languages

In two-dimensional positional languages the input information corresponds to a given position in two-dimensional space. This space is displayed on a screen. The correct position of the information is generally indicated by means of a cursor controlled by a keyboard, through a joystick, or a mouse. Other methods include use of a lightpen or touch-sensitive screens.

Two-dimensional positional languages have many uses, the most important of which include systems for filling in forms, systems for screen editing, and two-dimensional query systems.

In form-filling systems the user is provided with a format map displayed on a screen and he/she can then insert the appropriate data in free areas. The format map is protected and cannot be inadvertently altered from the keyboard. After filling in the form the user presses a special key and all input data are transmitted to the computer. This type of technique is very easy to use. A typical "form" is shown below; note that the user can only put data between the square brackets [ ].

```
NAME [                         ]
FIRST NAME [                   ]
BIRTHDATE
     DAY [     ]  MONTH [          ]  YEAR [      ]
PERSONNEL CODE [               ]
```

The second type of two-dimensional positional system involves editing on a screen. A screen editing system displays part of a file on the user's screen and allows him/her to make changes at the position indicated by the cursor. There are three principal types of command in a typical screen editing action language (Pearson 1980): cursor movement commands, text movement commands, and text modification commands. In some cases cursor movement commands can be replaced by the use of special keys on the keyboard (Altair Software Distribution Company 1977).

As our final example in this section we consider a two-dimensional query system, known as the Query-By-Example system (Zloof 1976), which is used as an interface between user and data base. In order to query the data base the user inserts a possible answer in the skeleton of the data base displayed on the screen.

As an example, the skeleton of the data base used earlier is given below.

PRESIDENTS	NAME	PARTY	HOME-STATE

Here PRESIDENTS is the name of the relation, and NAME, PARTY, and HOME-STATE are the names of the domains. To obtain information the user should fill in the skeleton using an example element (a variable), which must be underlined, and a constant element, which should not be underlined. In addition, the function "P." (print) must be inserted before the example element to indicate that this class of data forms the output.

As an example, assume that the user wishes to print out the names of the Democratic Presidents of the USA since 1956 using the relation PRESIDENTS; he/she just fills in the skeleton with P.<u>NIXON</u> (the name of any President would do) and DEMOCRAT.

PRESIDENTS	NAME	PARTY	HOME-STATE
	P. <u>NIXON</u>	DEMOCRAT	

The answer of the system should be:

NAME
KENNEDY
JOHNSON
CARTER

3 MAIN FEATURES OF DISPLAY LANGUAGES

In this section we consider the output side of the user- application software interface. The information produced by the computer to provide the user with feedback and other assistance we shall call the *display language*.

The display language must be able to perform a number of distinct functions. It should be able to:

- format the dialogue document (the printed record or screen display of the statements made by both user and computer)
- assist the user to input data and commands
- respond to the user after receiving valid input
- provide error messages
- provide "help" facilities

In this section we discuss the ways in which a display language can best fulfill these functions.

3.1 Formatting the Dialogue Document

There are a number of factors which help to produce a well-formatted dialogue document (Hebditch 1979).

1. Logical sequencing. The dialogue document contains several different types of information, and this information should be arranged in as logical a sequence as possible. One example of bad sequencing would be to blend input and output text.
2. Distinguishing input from output. It is very useful and improves legibility to distinguish inputs (action language phrases) from outputs (display language phrases). The ways in which this can be done depend on the type of terminal available. Possible methods include the use of lower-case characters for input and upper-case characters for output; underlining either input or output; using different colors or different line densities for displaying input and output, and so on.
3. Spaciousness. The whole two-dimensional space of the dialogue document can be used for output. Use of a tabular format can improve legibility; for example, if a menu is included in the output it could be presented as a table.

3.2 Assisting the User to Input Data and Commands

The assistance given to the user depends on the user–computer interface. For example, if an answer language is used as an action language, the form of the desired input is specified in the question asked by the computer. Another possibility is that the input language may contain keywords; in this case the system can be designed to assist the user through rapid keyword recognition. There are five forms of keyword recognition.

1. The whole keyword mode. In this case the user is obliged to type the whole keyword.
2. The anticipatory mode. This mode requires the user to type just enough characters for the command to be uniquely specified. The system then automatically fills in the remainder of the keyword.
3. The fixed mode. The keywords are chosen such that it is possible to recognize each keyword in a fixed number of characters.
4. The demand mode. This mode requires the entry of a special character to initiate recognition after the first part of the keyword has been typed.
5. The single-character mode. This mode allows high-speed single-character recognition of the most commonly used keywords. This mode may be used only when the keywords begin with different characters.

Another method of system assistance involves the use of *noise words*. When the system recognizes the first part of an input phrase, it can generate some words, called noise words, to tell the user what information is awaited by the system. For example, in the input command

```
CREATE LINE from x1 to x2
```

the words *from* and *to* could be generated by the system as noise words on recognizing the phrase `CREATE LINE`. The noise words prompt the user into entering data in the correct manner.

As mentioned earlier, the system may be designed to help the user by assigning default values to missing arguments, or by supplying missing information on the basis of previous commands. Whenever this happens the system should inform the user and ask him/her for confirmation.

3.3 Responding to the User after Receiving Valid Input

The system should provide regular feedback to the user on receiving valid input. The response should contain the following information:

1. Confirmation that input has been received. The system should confirm that the input is valid and has been accepted. In cases where misunderstanding is possible, as, for example, with natural language interfaces, the system can output a question and ask the user for confirmation.
2. Information about the unavailability of resources. If a process requested by a user involves the use of resources such as files or peripheral devices, the user must be informed if these resources are not available and, if necessary, *why* they are not available.
3. Output data. The output data can either be displayed on the user's terminal or by means of some other output device. In the latter case the user should be told where and how to obtain his/her results.

4. End information. When the process initiated by the user comes to an end, information about the mode of termination (normal, failure of the system, error in input, etc.) is useful.

3.4 Providing Error Messages

The system must anticipate errors in every piece of the input; sophisticated techniques must be used to handle these errors. There are three possible levels on which errors can be handled:

1. Error detection. The system must take great care to ensure that every error is detected.
2. Error recovery. After an error has been detected in the input text it is desirable to continue processing the remainder of the input without "pseudo-error" indications.
3. Error correction. Some errors may be corrected automatically, but in this case the system must ask the user for confirmation because it is possible to introduce insoluble problems through automatic error correction.

After an error has been detected the system must inform the user exactly and clearly of the nature of the error. Hebditch (1979) provides some guidelines on error-reporting techniques:

1. Avoid giving error messages in code and thus the need to refer to manuals.
2. Make error messages as self-explanatory as possible.
3. Error messages should be specified by the system designer, and the ease with which they may be understood and used checked with the potential users.
4. Errors should be detected as quickly as possible.
5. Avoid the need to rekey valid input during the error-correction process.
6. Recheck everything after correction.

3.5 Providing "Help" Facilities

Any software system must be properly documented in order to be usable (Cohen 1976). To document a large system is not an easy task, and it is made more difficult if the system is designed to be expanded by its users. Any printed documentation of such a system would be outdated before it was published and therefore the system itself must be capable of providing documentation that is guaranteed to be up to date.

In general, the user needs to know three things (Watson 1976):

- what he/she has already done
- what he/she is doing now
- what he /she can do next

The system should therefore provide information in the following three areas:

1. *Information space*. The user needs to know where he/she is in information space and which part of the information available is being displayed to him/her. The user arrived at his/her present position from a series of previous positions, and he/she may want to be able to return to these positions as well as to be able to move on. It is possible to achieve this by organizing help facilities in a tree structure. Each information node in the tree contains an explanation of a specific part of the system. The tree structure provides easy access to information about a specific topic.
2. *Subsystem or tool space*. In systems containing many tools (or subsystems), the user needs to know which tools are active, which tools he/she has used previously, and which subsystems can be entered from the present position. Each subsystem has a name and contains a number of related commands. In an ideal case all of the tools would operate on information in the same file because this would make it easier to move from one tool to another.
3. *Input syntax space*. Several ways in which the computer may help the user to formulate input have been described in a previous section. If, however, there is still some uncertainty about the basic concepts or the vocabulary, the user can employ the help facilities described above, either by specifying the concept causing difficulty or by making a more general request for help. In the latter case the system could make use of the information input up to this point to select the information required by the user.

In data-base management systems the user should be kept informed about the semantics of the data stored in the data base. In the data-base management system INGRES (Stonebraker *et al.* 1976) information about relations is available and may be used in the same way as help facilities which specify the names of the relations only.

4 CONCLUSION

Although we have discussed many of the issues concerned with user–application software interfaces, there are numerous aspects which we have not mentioned. This is largely because we have concentrated on interfaces in which alphanumeric texts are used as a means of communication. The main problem in communicating with a computer using alphanumeric texts is the great difference between the speed of the input and the speed of the output. While the output speed can be very high (thousands of characters per second), the speed of input via a keyboard is very low (less than ten characters per second). This drawback can be partly reduced by using single letters in an action language, for example, 'F' for `FIND`, 'P' for `PRINT`, and so on. However, this reduces the legibility of the dialogue document and can only be used by frequent users.

Graphics provide another promising medium for user–computer interfaces. Graphics can be used within display languages, action languages, or both. We have already discussed two-dimensional positional languages, in which simple graphics are used as part of a display language. Communication in such systems is both simpler and faster than using alphanumeric texts, as can be seen by comparison of a line-oriented editing system with a screen-oriented editing system.

The second problem in communication between users and application software is the selection and design of an appropriate language. Computers, especially small-scale computers, are increasingly being used as everyday tools in offices, businesses, and management. Most people using these systems have little or no knowledge of data processing. It is therefore desirable to design software systems with a nonprocedural interface for these applications, and a natural language seems to be the most appropriate. However, because of the problems involved in implementing natural language interfaces even on large-scale computers, we must suppose that formal languages will remain widely used in the future. Thus it is very important to design any language to be used by nonskilled operators so that it follows the natural language as closely as possible. The interested reader may find a more extensive discussion of languages designed for use in offices in Rohlfs (1979); the design of languages to be used in managerial systems is treated in more detail in Keen and Hackathorn (1979) and Blanning (1979).

ACKNOWLEDGMENTS

I am grateful to Göran Fick for many useful and valuable suggestions during the preparation of the manuscript, to Ronald Lee and Michael Dempster for critically reviewing the manuscript, and to Miyoko Yamada for her help in preparing this report.

REFERENCES

Addis, T.R. (1977) Machine understanding of natural language. International Journal of Man–Machine Studies 9:207–222.

Altair Software Distribution Company (1977) Altair Word Processing Package. Altair Software Distribution Company, Atlanta, Georgia, USA.

Astrahan, M.M., M.W. Blasgen, D.D. Chamberlin, K.P. Eswaran, J.N. Gray, P.P. Griffiths, W.F. King, R.A. Lorie, P.R. McJones, J.W. Mehl, G.R. Putzolu, J.L. Traiger, B.W. Wade, and V. Watson (1976) System R: Relational approach to database management. ACM Transactions on Database Systems 1:97–137.

Blanning, R.W. (1979) A language for describing decision support systems. Informal Workshop on Decision Support, Department of Decision Sciences, The Wharton School, University of Pennsylvania, USA.

Bobrow, D.G. (1977) GUS, a frame-driven dialog system. Artificial Intelligence 8:155–173.

Chamberlin, D.D. (1976) Relational data-base management systems. Computing Surveys 8:43–63.

Codd, E.F. (1977) Seven steps to rendezvous with the casual user. In Data Base Management, edited by J.W. Klimbie and K.L. Koffeman, pp. 179–199. North-Holland Publishing Company, Amsterdam.

Cohen, S. (1976) Speakeasy – A window into a computer. AFIPS National Computer Conference and Exposition, Conference Proceedings 45:1039–1047.

Fick, G. (1980) The challenge of low-cost computers to the organization. In Human Choice and Computers 2, edited by A. Mowshowitz, pp. 17–19. North-Holland Publishing Company, Amsterdam.

Fitter, M. (1979) Toward more "natural" interaction systems. International Journal of Man–Machine Studies 11:339–350.

Gaines, B.R. and P.V. Facey (1976) Programming interactive dialogues. In Computing and People, proceedings of a conference held at Leicester Polytechnic. Edward Arnold, Leicester, England.

Hebditch, D.L. (1973) Terminal languages for data base access. In Data Base Management, Infotech State of the Art Report 15. Infotech International Ltd., Maidenhead, Berkshire, England.
Hebditch, D.L. (1979) Design of dialogues for interactive commercial applications. In Man/Computer Communication, Infotech State of the Art Report 2, pp. 171–192. Infotech International Ltd., Maidenhead, Berkshire, England.
Infotech International Ltd. (1979) Man/Computer Communication, Infotech State of the Art Report 1: Analysis and bibliography. Infotech International Ltd., Maidenhead, Berkshire, England.
Johnson, S.C. and M.E. Lesk (1978) Unix time-sharing system: Language development tools. Bell System Technical Journal 57:2155–2175.
Keen, P.G. and R.D. Hackathorn (1979) Decision support systems and personal computing. Working Paper 79-01-03. Department of Decision Sciences, The Wharton School, University of Pennsylvania, USA.
Kowalski, R. (1979) Algorithm = Logic + Control. Communications of the ACM 22:424–436.
Kraegeloh, K.-D. and P.C. Lockemann (1978) Access to data base systems via natural language. In Natural Language Communication with Computers, Lecture Notes in Computer Science 63, pp. 49–86. Springer Verlag, West Berlin.
Lehmann, H. (1978) Interpretation of natural language in an information system. IBM Journal of Research and Development 22:560–572.
McCracken, D.D. (1978) The changing face of applications programming. Datamation 24: 25–30.
Miller, L.A. and J.C. Thomas (1977) Behavioral issues in the use of interactive systems. International Journal of Man–Machine Studies 9:509–536.
Olle, T.W. (1973) A summary of the state of the art in data base management. In Data Base Management, Infotech State of the Art Report 15, pp. 215–223. Infotech International Ltd., Maidenhead, Berkshire, England.
Pearson, M. (1980) Using the Computer to Communicate: An Introduction to Text Processing at IIASA – The "edx" and "nroff" Programs. WP-80-111. International Institute for Applied Systems Analysis, Laxenburg, Austria.
Rohlfs, S. (1979) User interface requirements. In Convergence: Computers, Communications and Office Automation, Infotech State of the Art Report 2, pp. 165–199. Infotech International Ltd., Maidenhead, Berkshire, England.
Schneiderman, B. (1978) Improving the human factors aspect of data base interactions. ACM Transactions on Database Systems 3:417–439.
Sprague, R.H. (1981) A framework for research on decision support systems. In Decision Support Systems: Issues and Challenges, edited by G. Fick and R.H. Sprague, pp. 5–22. Volume 11 in the IIASA Proceedings Series, Pergamon Press, Oxford, England.
Stonebraker, M., E. Wong, P. Kreps, and G. Held (1976) The design and implementation of INGRES. ACM Transactions on Database Systems 1:189–222.
Teitelman, W. (1979) A display oriented programmer's assistant. International Journal of Man–Machine Studies 11:157–187.
Thompson, K. and D.M. Ritchie (1975) Unix Programmer's Manual. Bell Telephone Laboratories.
Waltz, D.L. (1978) An English language question answering system for a large relational database. Communications of the ACM 22:526–539.
Watson, R.W. (1976) User interface design issues for large interactive systems. AFIPS National Computer Conference and Exposition, Conference Proceedings 45:357–369.
Winograd, T. (1979) Beyond Programming Languages. Communications of the ACM 22:391–407.
Zloof, M.M. (1976) Query-By-Example – Operations on hierarchical data bases. AFIPS National Computer Conference and Exposition, Conference Proceedings 45:845–853.

THE ENERGY SUPPLY MODEL MESSAGE

Leo Schrattenholzer
International Institute for Applied Systems Analysis, Laxenburg, Austria

SUMMARY

This paper describes the general form of an energy supply model, called MESSAGE (Model for Energy Supply Systems And their General Environmental impact). MESSAGE is a dynamic linear programming model minimizing total costs of energy supply over a given time horizon. It balances secondary energy demand, which is disaggregated into sectors and exogenous to the model, with primary energy supply, given as a set of resource availabilities which are disaggregated into an optional number of cost categories by choosing from a given set of energy conversion technologies. Model constraints reflect the limited speed of build-up of technologies, the limited annual availability of resources, and technological relationships – among other aspects of the energy system. The model permits the definition of load regions (e.g., for electricity demand), distinguishes between indigenous and imported resources, takes account of the environmental impact of energy supply strategies, and optionally includes this impact in the objective function.

The most important application of MESSAGE so far has been in the description of the IIASA global energy scenarios. A set of models and formalized procedures, each describing one part of the energy system, is arranged in a loop. Iterative application of this model loop leads to globally consistent scenarios in the development of the global energy system. The most important connection for MESSAGE in this loop is with the energy demand model MEDEE-2 that provides the demand data for MESSAGE. Information on energy costs flows back to the demand model.

With this application in mind it was felt that the model description could be made more specific if some generality of the model features was omitted in favor of a clearer exposition. Thus, some variables that can be chosen by the model user to suit a particular application are replaced here by the values they assumed in the global study. A case in point is the description of the energy chain; here it stops at the level of secondary energy as in the global scenarios, whereas the generic model formulation also permits consideration of final or useful energy.

These replacements make it easier to understand the description of the model relations. However, the description does not allow decisions on the relevance of individual model runs. Such a judgment can be arrived at only by using this description together

with a full documentation of the model including the input data for a particular run. The IIASA runs for the global scenarios are documented in a separate forthcoming paper (Schrattenholzer 1982a). Another related publication is a user's guide for the computer program that generates the computer image of the model (Schrattenholzer 1982b).

Experience with the model has been useful in several respects. It appears that MESSAGE can also be used as a stand-alone model if the information, which in the global study came from other models of the IIASA model set, is collected by different means. The model can also be applied to geographical regions other than the world regions of the global study. In general, such applications may necessitate modifications of the original model but the basic features remain the same.

---

1 INTRODUCTION

Future energy supply is a problem of truly global concern and the subject of national and international debates. These debates diverge enormously. One reason is that there are many different perceptions of the energy system and these perceptions often pertain only to isolated aspects of its future development. This was the setting when the Energy Systems Program (ENP) at the International Institute for Applied Systems Analysis (IIASA) undertook to formulate consistent scenarios of the development of the global energy system. The scope of this work has included:

- The construction of plausible and consistent paths of development of the global energy system.
- The conceptualization of the dependence of the various subsystems on each other, and of the system boundaries.
- The embedding of the energy system into the economy.
- The economic and environmental impact of energy supply strategies.

On this basis, two scenarios, High and Low, of the global energy system over the next 50 years were formulated using the following guidelines:

- Consistency, to be understood as the establishment of a global balance between supply (by primary energy sources) and demand (for useful energy) and a balance between global exports and imports of primary energy.
- Reasonableness, according to expert judgment of the scenario assumptions and of the scenarios themselves. Obviously each step in the definition and selection of an expert, in their judgment on a first round of scenario results, and in the incorporation of this opinion is subjective. However, what remains at least establishes a reference point with which everybody can compare his own assumptions and perspectives.
- Continuity, so that the scenarios evolve smoothly from historical development and are themselves smooth. This does not mean, however, that the scenarios are extrapolations of past trends.
- Degree of variation. The difference between the High and the Low scenarios was to be large enough to cover an area within which the actual development is expected to fall with reasonably high probability.

These general considerations resulted in a global energy model or, better, in a set of submodels which together describe the global energy system. MESSAGE, a linear programming energy supply model, is one part of this model set.

The main purpose of this report is to document an important part of IIASA's global energy model enabling those who are interested to understand the tool with which the corresponding parts of the global scenarios were built up. Together with the documentation of the MESSAGE computer program and the model input data that were used for the description of the global scenarios (Schrattenholzer 1982a, b), this report should provide sufficient information for those with the appropriate experience to reproduce the MESSAGE results for the global scenarios in any desired degree of detail.

Thus, the report mainly addresses those already familiar with energy models. For the sake of conciseness much of the description of the real-world energy supply systems, that is the background of MESSAGE, has been omitted.

Section 2 gives a short summary of the model system used for the global scenarios; Section 3 describes the MESSAGE model; Section 4 briefly discusses general aspects of applications of MESSAGE and reports on experiences with its usage for countries and regions other than those used in the global study. Appendices give a summary of some high-level input data used for the global model runs, the description of a standard form of a dynamic linear programming model, and a classification of the model variables.

2 A GLOBAL ENERGY MODEL SET

To give an idea of the model environment in which MESSAGE was primarily used, this section describes briefly the model set used for the global energy scenarios. The actual application of this model set, the scenarios, and independent studies that have been integrated into the scenarios are given in the comprehensive report of the activities of IIASA's Energy Systems Program Group (1981).

A schematic description of the model set and the information flow between its submodels is given in Figure 1. The part above the dotted line, describing models and formalized procedures, was applied to each of seven world regions (defined in Appendix A), which together cover the globe. The parts of the large model for each world region are arranged in the form of a loop; the seven loops are connected by a procedure that balances interregional trade of primary energy.

Historically, the first step in formulating the scenarios was to make assumptions about the overall development of the economy (expressed as GDP per capita) and about population. However, after a first round of model runs, the built-in feedback mechanism changed the original assumptions so that there is no real "beginning" of the model loop. The output of this procedure as well as assumptions about the economic structure (GDP per sector), about lifestyle (e.g., dwelling space per capita), and about technical efficiencies (e.g., fuel efficiency of a car) are inputs to the energy consumption model MEDEE-2, which in turn produces scenarios of energy demand. (For a description of MEDEE-2 see Lapillone 1978.) These are converted into scenarios of secondary energy demand (by fuel type) using scenario assumptions for the allocation of demand for substitutable fuels to the various fuel types. The demands for secondary energy, maximum build-up rates, and cost figures are inputs into the MESSAGE model. MESSAGE results on the marginal

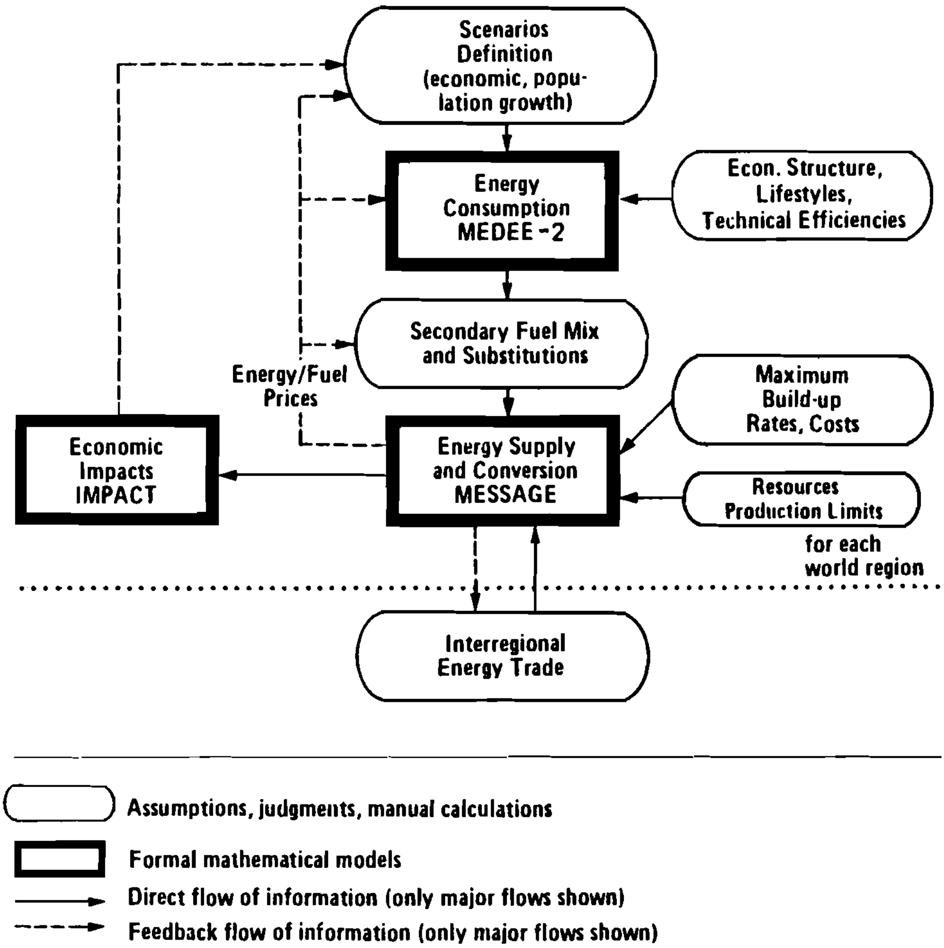


FIGURE 1 IIASA's global energy model.

costs of supplying secondary energy feed back into MEDEE-2. The other connection between MESSAGE and the model loop is with the procedure for the balance of global primary energy trade. Information about energy requirements flows from the world regional MESSAGE models into this procedure for energy trade, resulting in available imports for each region. The output of this procedure that is most important for the global scenarios is the development of the global oil market. Global availabilities of crude oil imports are based upon the assumption that the oil-exporting countries maximize the revenues from their exports and assumptions about the resource situation in the seven world regions. The results of MESSAGE are fed into the IMPACT model (Kononov and Pór 1979) which calculates the economic impact of energy supply strategies in terms of some economic variables (direct and indirect investments, land requirements, requirements for skilled labor, etc.) which are compared with the original assumptions on general economic development, thus closing the main model loop.

This loop is applied iteratively until satisfactory consistency is achieved.

3 GENERAL DESCRIPTION OF MESSAGE

The energy supply model MESSAGE (Model for Energy Supply Systems And their General Environmental impact) is a dynamic linear programming (DLP) model which minimizes total discounted costs of energy supply over a given time horizon. The main subject of the model is the balancing of demand for secondary (or final) energy and supply of primary energy resources via divers technologies. The most important model constraints reflect limits on the speed of build-up of technologies, the availabilities of indigenous and imported resources, and technological relationships. Major distinctive features of the model are the consideration of load regions for electricity demand, the disaggregation of resources into cost categories, and the consideration of the environmental impact of energy supply strategies. The model output is used to describe scenarios of energy supply. The description comprises the physical flows of energy between primary energy and eventual use as specified by the demand data, shadow prices of supply and demand constraints, and the environmental impact of energy supply paths, expressed as emissions and concentrations of pollutants. The energy flows give a consistent picture of the supply/demand balance; and the shadow prices allow for an assessment of the incremental benefit of additional resources, the incremental benefit of new technologies, and the marginal costs of meeting additional demand. The environmental module may be used to model the influence of emission or concentration standards (upper limits) on the model solution. Another possibility is the inclusion of emissions and/or concentrations of pollutants in the objective function.

Since this report gives only a general description of the model, it does not provide a solid basis for judging its results. Such a basis would have to include the whole set of input data which is documented elsewhere (Schrattenholzer 1982a). This documentation supplements the comprehensive report of the activities of IIASA's Energy Systems Program Group (1981). Another related publication is the user's guide for the computer program that generates the LP matrix in a standard format (Schrattenholzer 1982b). [The FORTRAN code generating a standard-format (MPS) LP matrix is available through the Energy Systems Program at IIASA.] In order to give at least some sense of the application of MESSAGE within the global scenarios, some higher level input data are summarized in Appendix A.

3.1 Model History

MESSAGE is the last in a series of LP energy models that were developed at IIASA. The first one was the Häfele–Manne model (Häfele and Manne 1974); this was followed by Suzuki's model (Suzuki 1975). The most significant differences distinguishing MESSAGE from its predecessors are the following: the inclusion of an optional number of primary energy resource categories allowing for the modeling of the nonlinear relation between extraction costs and available amount of a resource; the explicit consideration of demand load curves, in order to take account of e.g., the variation of demand for electricity; the calculation of residual discharges to the environment; increased program flexibility allowing e.g., for easy modular inclusion and removal of technologies. This flexibility has the disadvantage of somewhat blurring the distinctions between model, model inputs, and model input data. However, these distinctions will be discussed in some detail below.

The model history also includes stages in the development of MESSAGE as documented in Agnew et al. (1978, 1979). Agnew et al. (1979) describes an earlier concept of the model encompassing a part of the energy supply system significantly larger than considered for the actual global model runs. The restriction to the size described here was a consequence of the purpose of its application, the lack of sufficient data, and the level of detail to be incorporated. Agnew et al. (1978) is a user's guide for an earlier version of the computer program for the implementation of MESSAGE.

3.2 Model Relevance

In this section, questions of model credibility and model validity are discussed. These points precede the main model description as they establish the frame of reference for a part of the description. This section, containing some theoretical aspects, is an important part of the model description. However, it is not relevant to those using this report as an introduction to MESSAGE or as a reference, and may be omitted.

The validation of a model or, more generally, a discussion of model relevance would be a routine task if standard procedures were available and simply needed applying. Such procedures would have to be based on a general theory of systems and models. However, general and abstract concepts in mathematical system theory are not sufficiently familiar, nor widely accepted, among model builders and users nor are these concepts detailed enough to be used or referred to. Therefore two important aspects influenced the shape and content of this section. First, we emphasize the importance of model validation and second, as there are still no standard procedures of general systems and model theory to be followed, we take the freedom available to rearrange and modify existing theory in order to best combine general theory and particular model description. With this in mind, the following definitions should be considered merely as working definitions which are not exhaustive in the sense that they do not describe a complete theory. The reader who is further interested in theory is referred to Casti (1980) and Kalman (1974).

The first step is definition of a model as a description of a system. Accordingly, whenever a system is mentioned in connection with a model, the system being modeled is referred to. Where no confusion can arise, parts of the model are denoted by the same name as the relevant parts of the system.

The next step is a classification of models (Häfele 1980). This classification is not intended to be exhaustive, i.e., applicable to all models. Rather, it should distinguish between three important model types for which validation is quite different. The characteristic features of each type of model should therefore be viewed more as ingredients than as complete descriptions. This classification explains the background for the kind of model validation that was adopted for this report.

Three types of models are distinguished:

1. Models formulating laws of nature.
2. Models formulating regular behavior.
3. Models formulating concepts of controlling man-made systems.

The model attribute underlying this classification could – in accordance with Lewandowski (1980) – be called "model background." The description of the respective backgrounds for the three types of models would be natural, behavioral, and man-made. With respect to application, a more important model attribute is predictive capability. The above three types of models are ranked in decreasing order of predictive capability. The following discussions give a more detailed characterization of the three types of models from the point of view of the two attributes just described.

Models of Type 1 directly reflect what are accepted as laws of nature. Their predictive power is evident and prediction is the main purpose of their usage. These models are found rather than constructed; their validity is demonstrated by reproducible experiment. An example of this type of model is the motion of a mass point in a gravity field.

Models of Type 2 are some models in economics, sociology, biology, etc. They usually reflect theories – natural laws, if they exist, have not been found. Their predictive value is local in the mathematical sense, i.e., for points in the state space which are sufficiently far away from the initial condition, these models are sometimes quite wrong. Models of this type typically reflect macro phenomena based on the behavior of micro agents. They are validated by generating output from historical input data and comparing model outputs with the observed development of the underlying system. An example is the development of a species in a limited environment.

Models of Type 3 often have a rather simple conceptual basis entailing a rather trivial mathematical formulation (often including a large number of relations between their variables). Many relations expressed in the model are immediately plausible. An important nontrivial aspect of such a model is for example, the degree of detail incorporated, reflected by the number of relations contained in the model and by the boundary assumed for the system modeled. The predictive capability of this type of model is often irrelevant, although often overestimated. The typical point of their application is to study consequences of alternative lines of action and alternative sets of constraints to decisions. They are used to describe consistent scenarios providing a quantitative conceptualization of complex systems. Much like physical models, they help visualization of the relation between a large number of assumptions. As MESSAGE belongs to this group of models, the validation of this type of model will be discussed in more detail below. Before, some more definitions are needed.

The dynamics of the system are described by its states at time t. Restricting our considerations to models that are described in mathematical form, the state of a system at time t is described by the value of the state variables of the model for that point in time. The state variables are therefore those variables that describe the dynamics of the system modeled. All model variables that are not state variables are control variables. They describe the control that can be applied to the state variables in order to move the system in a desired way. Using MESSAGE as an example, the capacity of, say, coal-fired power plants is a state variable which is influenced by the control variable construction of coal-fired power plants.

The system equivalent of the control variables is called system inputs. These model inputs are not to be confused with the input data to the mathematical model (the computer code representing it).* The difference between the two is easily seen in the example

$$0 \leqslant u(t) \leqslant 1$$

in which the control variable $u(t)$ is the representation of a system input whereas the parameters defining the lower and upper bounds are input data to the model.

In addition, it is helpful to further classify the input data to the computer program (representing the model) in the following way:

- Data that should be considered part of the model because they refer to system boundaries and to the degree of detail of the description of the system, e.g., the number and the names of energy demand sectors.
- Data constraining the system inputs (control variables) and system outputs, e.g., the amounts of resources available over the model time horizon.
- Model parameters, a part of the internal description, also called the "wiring diagram" in a "black-box" description (Casti 1980) of a system, e.g., the thermal efficiencies of power plants.

These groups should not be taken as an unambiguous classification but as a summary of the most important characterizations of data. In the context of this section, they are ranked in ascending order according to the degree to which they can be wrong. This means that data in the first group deal with the amount of detail that is to be incorporated in a model run and that they are therefore based mainly on judgment, whereas the last group of data refers to observations of the real system being based more on objective facts.

By system outputs we understand the real-world equivalents of the mathematically described model outputs.

Now, we turn to the validation of Type-3 models. Along with an overestimation of their predictive capabilities often goes an inadequate demand for their validation. A distinction should be drawn between forecasts and scenarios. The emphasis of a forecast is on the description of a single future development (with only statistical deviations) of the underlying system; scenarios (the result of Type-3 models), on the other hand, draw a consistent picture of the consequences of a given set of assumptions, the probability of which is not necessarily being assessed. As a consequence, a single result of Type-3 models is always much less relevant than comparisons between several model runs which typically attempt to investigate the influence of (unknown) input parameters on the model solution. Validating such a model therefore means demonstrating that the mapping from the set of assumptions onto the set of consequences is done correctly by the model. However, the question of correctness cannot be answered by yes or no. Rather, all an answer can reasonably give is the degree to which the mapping is correct. Since there is no established scale by which this degree can be measured, the answer can be given only in qualitative terms.

---

*For the sequel it will be assumed that, as is the case for MESSAGE, there is a computer program that corresponds to a model.

Turning to the validation of MESSAGE, the correctness of the mapping done by the model depends on the degree of simplification of a model run.* This degree of simplification is defined by the model formulation as described below *and* by the input data for a particular model run. Hence, validation of MESSAGE in the framework of this report means discussing – in qualitative terms – the degree of simplification which is implied by the very formulation of the model relations. The reader can then judge whether he agrees that MESSAGE is valid enough for its application within the global model or not.

To aid understanding a transparent presentation of the model relations was aimed at. One means to achieve this was the style chosen for the model description, consisting in a standardized formulation of dynamical LP models (Propoi 1977) for describing the model relations. One important advantage of this formulation is that it differentiates LP variables into state variables and control variables as well as distinguishing between constraints that are state equations and actual constraints. The mere separation of these terms significantly helps comprehension of the model dynamics.

The second part of the model validation, the discussion of the simplification of the mapping as defined by the input data, belongs to the data documentation and is therefore contained in Schrattenholzer (1982a). Here, we conclude these general considerations of model validation by emphasizing an important practical part of model validation – the repeated model runs. These can provide a judgmental check of the model output for reasonableness. However, this does not mean that any results that seem unreasonable point to an invalid part of the model. They may equally well reveal an important point in the system dynamics that had been overlooked. To differentiate between these two kinds of "unreasonable" results, the model user traces the input data through the model to see which model part causes the particular result.

3.3 Model Description

This section describes the general form of MESSAGE. To give an adequate description of a particular model run, the general outlines must be supplemented by the input data for the particular runs. In cases where it helps to understand the general equations, some of the input parameters are denoted by the values that were used in the global analysis. Further input data for the seven IIASA world regions are summarized in Appendix A. For a full documentation of all input data used in the global scenarios, see Schrattenholzer (1982a).

3.3.1 General Model Description

This description is as follows. The generic form of the model relations (i.e., functions, constraints, equations) will be aggregated into groups and written down in matrix/vector notation. Each group of relations is described in a separate subsection. The grouping of model relations follows a standardized formulation of a Dynamic Linear Program (DLP) as defined in Propoi (1977) and summarized in Appendix B.

*This is assuming that the corresponding computer program is correct, i.e., exactly generating the mathematical formulation of the model. Although, in practice, this is not a trivial aspect, for the theoretical description here we make this assumption.

Each group of relations will be followed by an explanation of those symbols that have not previously been introduced. The relations will be interpreted where not readily clear. Next, a description of the input requirements for this group of relations will be given. The last paragraph of each subsection contains a discussion of their relevance in the theoretical terms described in Section 3.2. These final paragraphs are not relevant for those readers that omitted that section.

For the formulation of the dynamic equations, the time horizon of the model is divided into n time intervals of equal length. These time intervals lie between the grid points $t_0, \ldots, t_n$. In the runs for our global scenarios, the number of time periods was chosen to be 11, each one containing 5 years. Although these numbers are not fixed for the computer program corresponding to MESSAGE, they will – for the sake of clarity of the description – be treated as if they were fixed. This remark is valid for some other model parameters, too. In cases where it matters for the scenario runs, these parameters will be identified.

3.3.2 Demand/Supply Balance

$$\mathbf{D}x(t) \geqslant d(t) + \mathbf{H}x(t) \qquad t = 1, \ldots, 11 \tag{1}$$

where

D is the matrix describing supply/demand paths (constants)
x is the vector of annual supply activities (control variables)
t is the index of the current time period
d is the vector of annual secondary energy demand (exogenous inputs)
H is the matrix with the coefficients for the inputs of secondary energy required by technologies (constants)

This group of constraints links the output of energy conversion technologies to the vector of exogenously given energy demand. The matrix **D** contains ones, if a technology contributes to the supply of a demand sector,* and zeroes where a technology does not contribute to the supply of a demand sector. For the sake of clarity, the matrix **H**, which defines inputs of secondary energy into conversion technologies (and thereby increases the exogenously given demand within the model), is defined separately from the matrix **D**. (As they have the same dimensions, they could have been added into one matrix which would have to be interpreted accordingly.) It should be noted here that the consideration of demand for secondary energy is a consequence of the model application. There is no implication by the model formulation that the modeling of the energy chain has to stop at the level of secondary energy. If desired, the energy chain could be considered until its end (i.e., end-use of energy) using the same general model. In that case, the demand vector contains secondary energy carriers (e.g., electricity) and end-use sectors (e.g., space heating). The only demand sectors for which demand is exogenously specified are then the end-use sectors. Demand for secondary energy is then calculated endogenously via the intermediate demand of the technologies supplying end use, expressed by the matrix **H**.

*In some cases the output x_i of a technology may be defined in different units than the demand d_j; then, the respective matrix coefficient is the applicable conversion factor rather than unity.

Input data for this group of constraints describe the conversion factors of technologies (**D**), input requirements that consist of secondary fuels (**H**), and the energy demand projection (d). A demand sector can be divided into load regions (see Subsection 3.3.3).

These constraints set a lower limit to the variables x which are control variables. These lower limits are exogenous and can safely be called driving parameters of the model.

3.3.3 Capacity Utilization

$$\mathbf{B}_1 x(t) \leqslant c(t) \qquad t = 1, \ldots, 11$$

$$\mathbf{B}_n x(t) \leqslant c(t) \tag{2}$$

where

$\mathbf{B}_i$ are the matrices defining load regions and availability of technologies in the load regions; $i = 1, \ldots, n$ (input data)
c is the vector of installed capacities (state variables)

Since this form of the utilization constraints is not very instructive, they will be derived using the example of an ordered load curve of electricity demand and three load regions. The general case is then easily seen by analogy.

The upper part of Figure 2 shows an ordered load curve for electricity demand. This curve is approximated by a step function consisting of three steps. The load duration (width) of each step is optional and is part of the model input data. It should be chosen so as to "optimally" approximate the given load curve. The height of the step is determined so that the areas under the load curve and its approximation are identical. The supply activities (the vector x) for those technologies which supply electricity are disaggregated according to the load durations as shown in the lower part of Figure 2. In this graph the upper horizontal line represents the installed capacity, the dashed horizontal line represents the upper limit of the actual utilization of a power plant, determined by the plant factor.

Thus, for the j-th step of the demand curve, the capacity constraint for a given load region and for a given technology (now in scalar notation) is

$$x_{i,j}(t) \leqslant c_i(t) \cdot (h_j - h_{j-1}) \cdot pf_i \qquad t = 1, \ldots, 11 \tag{3}$$

where

i is the index of the technology
pf_i is the maximal plant factor, expressed as a fraction of installed capacity of the technology
$(h_j - h_{j-1})$ is the duration of the j-th load region ($h_0 = 0$)

In order to be able to express all these constraints in the form of eqn. (2), the constraints of eqn. (3) were divided by the constant factors of the right-hand side.

A few remarks are in order. This definition of a load region differs from the usual definition, in which a load region is defined as the area under a function of utilization

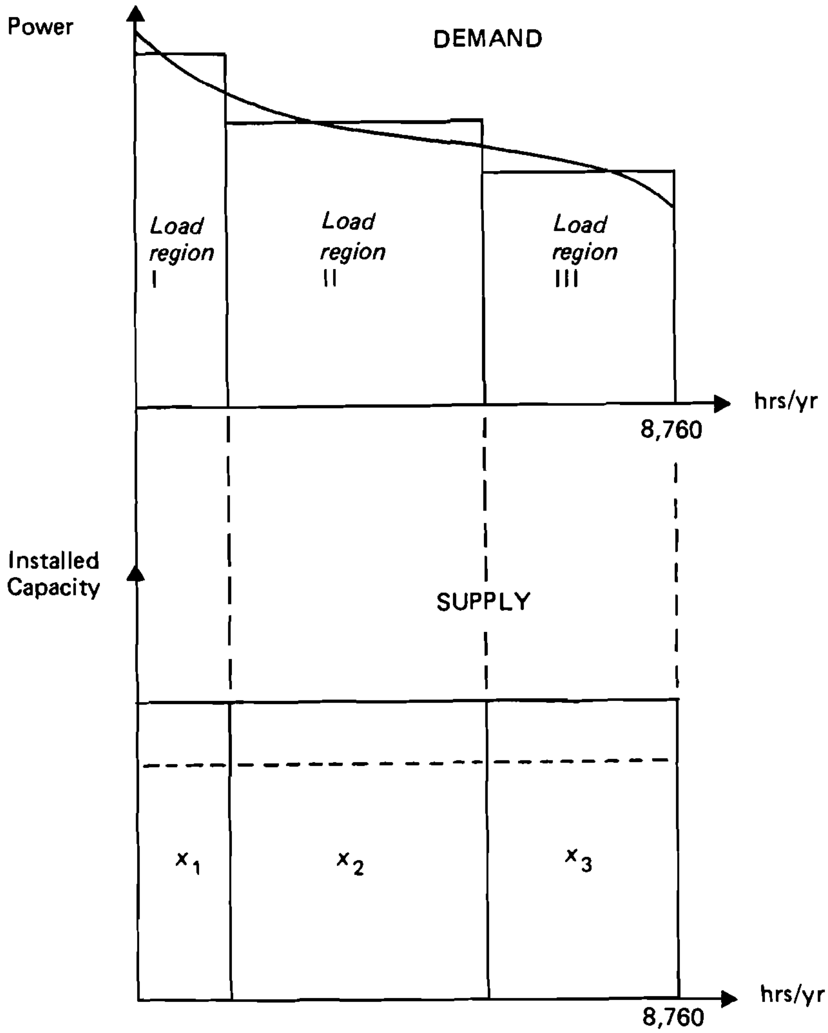


FIGURE 2 Disaggregation of supply activities.

hours per year. (In Figure 2 the usual definition would be represented by a horizontal division of the total area.) This difference is not drastic because the model solution allows results to be shown in terms of either definition. The advantage of our formulation is that it is more flexible for example, permitting variable load durations of technologies and endogenous adjustments of the demand load curve.

Input data required for these constraints are the durations of the load regions $(h_j - h_{j-1})$ and the maximal plant factors of the technologies (pf_i).

These constraints are constraints on control variables by state variables. The fact that the formulation of the model assumes that the usage of a load region can be chosen by the model seems to be in conflict with reality, where technical properties of power plants put constraints on their usability in the load regions. However, this formulation was preferred because some of the technologies included in the model will only be developed in the future. For these technologies the model results provide criteria for the design of these power plants. The drawback of such freedom of the model – it can yield technically infeasible solutions – is compensated by the possibility of restricting the use of technologies to certain load regions. Details about the implementation of these restrictions in the MESSAGE computer program are found in Schrattenholzer (1982b).

3.3.4 Capacities of Technologies

$$c(t) = c(t-1) + 5z(t) - 5z(t-6) \qquad t = 1, \ldots, 11 \tag{4}$$

where

z is the vector of annual additions to capacity (control variables)
$t - 6$ reflects a 30-year service life: after 6 (variable in the program, but fixed here) periods of service, an energy conversion facility is phased out

Input data belonging to this group of equations are a list of technologies (defining the length of the vectors c and z), their initial capacities $c(0)$, and the historical construction rates $[z(t-6)$ for $t-6 \leqslant 0]$ as initial conditions.

These equations are state equations. They relate the variables c (state variables) to the variables z (control variables). Since the single z-variables can take any nonnegative value, this means that the size of installed energy conversion facilities can be arbitrarily small. Because of the large size of the geographical regions modeled and because of the large time horizon, this did not pose a problem in the global analysis. If the model is applied to smaller geographical regions or for shorter time spans, this point must be kept in mind.

3.3.5 Build-Up Constraints

$$z(t) \leqslant \gamma z(t-1) + g \qquad t = 1, \ldots, 11 \tag{5}$$

where

γ is a (diagonal) matrix of growth parameters (input data)
g is a vector of start-up values allowing z to reach positive values after having been zero before (input data)

$$\sum_{i \in I_1} z_i(t) \leqslant GUB(t) \qquad t = 1, \ldots, 11 \tag{6}$$

where

$GUB(t)$ is a time series of absolute upper limits (input data)
I_1 is a subset of the set of technologies

The first group of constraints limits the growth rates of build-ups of single technologies, the second one puts an absolute upper limit on the total installation of a group of technologies. The first group is particularly important for new technologies, the second group is so far only used for limiting the total annual installation of nuclear capacity.

Input data are the growth parameters ($\gamma_{i,i}$) of the first group of constraints and the time series of installation limits [$GUB(t)$] for the second.

The functioning of eqn. (5) can be illustrated for the case of a new technology, for which these constraints are binding for some time periods, t_i, t_{i+1}, etc. In this case, total installed capacity of this technology is proportional to the parameter g and roughly proportional to $\gamma_{i,i}^n$ where n is the number of time periods in which the constraint is binding.

The build-up constraints just constrain the control space making sure that some foresight is employed by the model – the need for technology in the future is anticipated early enough, keeping growth rates within limits. This function is not the only reason why these constraints are to be considered to belong to the most important ones of the model. Another reason is suggested by Marchetti and Nakićenović (1979), where the penetrations of new energy carriers into existing energy systems are investigated. There, many examples are given and a theory is described that support the conjecture that such penetrations follow internal laws. In MESSAGE, a less constraining form of such a law, an inequality instead of an equality, has been incorporated.

3.3.6 Resource Balances

$$s(t) = s(t-1) - 5r(t) \qquad t = 1, \ldots, 11 \tag{7}$$

where

s is the vector of reserves (stocks) of primary energy carriers or man-made fuels (state variables)
r is the vector of annual consumption of primary energy carriers (control variables)

The lengths of the vectors s and r depend on the number of natural and man-made resources incorporated in the model. These resources can be subdivided into different cost categories in which case the vectors s and r are extended accordingly. Such a disaggregation can be interpreted as an approximation of the nonlinear relation between the availability and the unit cost of a resource by a step function.

Input data belonging to this group of constraints are the total resource availabilities $s(0)$. It is worth remembering here that LP variables are nonnegative by default. This is an important constraint for the vector s making sure that not more than the initial availabilities $s(0)$ are consumed in the model. For those activities in the vector r that refer to man-made fuels, this nonnegativity constraint is removed so as to also allow for production (not only consumption) of these materials.

Renewable energy sources (solar, hydro, etc.) are not included in these constraints as their total availability is unlimited for the purpose of the model. However, the *rate* of utilization of renewable sources is limited. This limitation is introduced as the characteristic of a technology converting renewable energy and is described in Subsection 3.3.9 where the bounds of the model variables are discussed.

Eqn. (7) is a state equation relating annual consumption of primary energy r (control variables) to the total stock of primary energy s (state variables). An implicit assumption behind this formulation is that any share of available resources can be used at any time within the time horizon and thus arbitrarily fast. This is in contrast to the real world, where a good part of the resources used in a given time interval of 50 years are only gradually discovered or deployed. However, MESSAGE contains a means for compensating this drawback to some extent as the model formulation allows the limitation of annual production of primary energy. These constraints are described below.

3.3.7 Resource Consumption

$$\mathbf{G}r(t) \geqslant \mathbf{Q}_1 x(t) + \mathbf{Q}_2 z(t) - \mathbf{Q}_3 z(t-6) \qquad t = 1, \ldots, 11 \tag{8}$$

where

$\mathbf{G}$ is a binary matrix (containing only zeroes and ones) aggregating all categories of a resource (input data)

$\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{Q}_3$ are matrices of parameters describing specific consumption of resources by conversion technologies (model input data belonging to the definition of technologies)

As mentioned above, resources can be divided into different cost categories. Since they are nevertheless meant to serve the same purpose, these categories are aggregated, by the matrix **G**, when balanced with the resource consumption of the energy conversion facilities. This consumption is expressed by the matrices $\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{Q}_3$ which describe resource consumption per unit of output ($\mathbf{Q}_1$), per unit of new capacity ($\mathbf{Q}_2$), and recovery of a resource at the end of a service life of a technology ($\mathbf{Q}_3$). For the time being, $\mathbf{Q}_2$ and $\mathbf{Q}_3$ are exclusively used to describe inventory requirement and recovery of nuclear fuels.

Input data for these constraints are the specific consumption of fuels (matrices $\mathbf{Q}_i$) and the disaggregation of resources into categories (matrix **G**).

As already pointed out, the disaggregation of resources into cost categories and their subsequent aggregation for the purpose of their consumption can be interpreted as a nonlinear cost function for a resource. The way the model is set up, the independent variable in this function is cumulative use of a resource. An alternative formulation would be to have such a function depend on the use of a resource in each time period, in which case the model could determine the cost level up to which it is optimal to deploy a resource. Such a formulation seems to be more realistic but at the same time more data intensive. This trade-off has been resolved in favor of using the above formulation. The linearity of the right-hand side of the above relation assumes a linear relationship between fuel input and energy output for the energy conversion facilities. This simplification was considered appropriate for the applications carried out so far.

3.3.8 Resource Extraction

$$\mathbf{G}_1 r(t) \leqslant p(t) \qquad t = 1, \ldots, 11 \tag{9}$$

where

$\mathbf{G}_1$ is the matrix for the aggregation of indigenous resource categories (input data)
p is the vector of annual production limits for each resource kind (exogenous inputs)

Optionally, one category per resource may be defined as import category. (Since there can be at most one import category for each resource, the costs of import categories may be defined to be time dependent. See also the description of the objective function.) Such a definition does not change the purpose of its use as a resource as described in eqn. (8). The separate definition of an import category applies in this group of constraints, where the total annual extraction of only indigenous resource categories is constrained. The annual amount available for the import of any resource is constrained separately as described in Subsection 3.3.9.

Input data required for these constraints are time series of upper limits for the annual extraction of indigenous resources [$p(t)$].

This group of constraints sets an upper limit to a part of the control variables r. The way these constraints are formulated assumes that each cost category of a particular resource can be exploited up to its maximum availability before the next category is tapped. In the LP solution this is reflected by abrupt transition between adjacent categories. It may be argued that for example, a separate bound on each category for each time period would be more realistic because depletion rates of resource categories could be taken into account thus allowing for smoother transitions between categories. At the same time, such a formulation would be one way to approximate the nonlinear cost function for a resource for each time period rather than over the time horizon . Because of these advantages, the decision to choose this particular formulation was very close, just slightly in favor of the decision finally taken.

3.3.9 Bounds on Single Variables

Since bounds on single variables in many LP computer packages are taken together in one set, they are treated here under one heading; and since the mathematical formulation is too trivial to add any further information, only a qualitative description is given.

Lower, upper, or fixed bounds may be set on the supply of a single technology (x-variables) and on the installation of new capacity (z-variables). Bounds on x-variables are usually used to limit the harvesting of renewable energy sources; bounds on z-variables are a means to constrain capacities of technologies also in absolute terms (not only their growth as described in Subsection 3.3.5). Upper bounds may be set on the annual availability of an import resource category. The treatment of renewable energy in MESSAGE deserves some further explanation. As is the case for nonrenewable energy, increasing utilization, in general, entails increasing unit costs. This is usually taken into account in the model, by defining two (or more) technologies that have the same technical characteristics but different costs. Bounds are then set for the output of these technologies (the corresponding x-variables) making sure that the model cannot choose more of a technology than that available at a certain cost.

Input data to this part of the model are the kinds of bounds (lower, upper, fixed) and the corresponding time series.

Although the practical relevance of the bounds is trivial, the fact that these bounds constrain the control space of the modeled system and thus significantly influence the model output, makes the bounds a characteristic ingredient of Type-3 models and thus of MESSAGE.

3.3.10 Objective Function

$$\sum_{t=1}^{11} \beta_1(t)(a_1 x(t)) + \beta_2(t)(a_2 z(t)) + \beta_1(t)(a_3(t), r(t)) \rightarrow \min \qquad (10)$$

where

β_i are discount factors (input data)
a_i are vectors of annual cost coefficients (input data)

Because of the special formulation of eqns. (2) and (4), which imply that the annual utilization of a capacity already includes the build-up in the same time period, the build-up variables z have to be interpreted as activities occurring before the supply activities x which utilize this capacity. Consequently, the discount factors belonging to the respective LP variables have to be different to take account of this time lag. Besides the time lag, all the discount factors $\beta_i(t)$ are uniformly calculated using constant annual discount rates. In addition, the parameters $\beta_2(t)$ contain a correction factor expressing the value of capacity that keeps operating beyond the model time horizon. Thus the objective function excludes costs, the benefits of which do not accrue within the model boundaries. The cost coefficients are – with the exception of the costs for an import resource category – constant over time, the usual interpretation being that the costs remain constant in real terms. Accordingly, the discount factors are interpreted as real discount factors excluding inflation. An expected change of costs of technologies can be reflected in the model by defining two model technologies with different costs and different availabilities over time.

The input data belonging to the objective function are the annual discount rate and the cost coefficients.

Taking costs as the function to be minimized assumes economically rational behavior of future decision makers. Although it is implied in what has been said already, it should be pointed out here that this formulation of the objective function does not mean that costs are the only criterion determining the model output. Other criteria are imposed by the model constraints. (Subsection 3.3.12 on the environmental submodule of MESSAGE introduces still another criterion.) Also it is repeated here that in the applications for the global runs the location of the feasible region in the state space, determined by the scenario variables, has always had a larger effect on the solution than the optimal point in the state space, determined by the objective function. Nevertheless, some model features based on this function will be discussed in more detail.

One important drawback of an LP model – in the case of uncertain input data – is the discontinuous dependence of the solution on the objective function: a small change in cost data can eliminate a technology's contribution to energy supply replacing it by a different one. Thus, if no consideration is given to this feature, a "second-best" solution can remain undiscovered although it would be optimal under an only slightly different objective function. Usually, this problem is solved by analyzing the model sensitivity. However, the sensitivity analysis depends on the model application much more than on the model itself and therefore falls outside the scope of this report. The particular form of sensitivity analysis that was thought to be appropriate for the global analysis is described in Energy Systems Program Group (1981). A related problem is the sudden replacement of one energy source by another as soon as a "cheaper" one becomes available. MESSAGE contains two features that limit this unsteady behavior of the solution: the build-up constraints for technologies prevent sudden build-up of a technology; and the explicit consideration of capacities the underutilization of which entails economic penalty thus working against a sudden drop-out of a technology.

3.3.11 Initial Conditions

The initial conditions have already been described together with the corresponding variables. Here, they are summarized under a common heading.

The initial capacities $c(0)$, and the historical build-up activities $z(0), \ldots, (z-5)$ of all technologies have to be specified consistent with the conditions

$$c(0) = z(0) + \ldots + z(-5)$$

and

$$z(t) \geqslant 0, \qquad t = 0, \ldots, -5 \tag{11}$$

The only other set of initial conditions are the numbers of the availabilities of natural resources, which in fact are an upper limit on the amount of resource which is available in a category within the time horizon of the model.

The theoretical relevance of the initial conditions for MESSAGE does not seem to justify separate discussion. (The initial conditions are much more critical for the evolution of the system for those types of models in which control plays a lesser role.)

3.3.12 Environmental Submodule

The environmental impact is one of the most intensively discussed aspects of energy supply strategies. Along with widespread discussion goes a large number of vastly different views of various impacts, expressed by significant differences in the corresponding data. Taking these discrepancies into account would have meant performing a considerable number of model runs in order to match the technical and economic parts of the global energy scenarios with an adequate scenario on environmental impacts. Even an extended study in this respect would have carried the risk of being inconclusive. In view of these difficulties the two global energy scenarios restrict the consideration of the environmental impact of energy supply strategies to one global kind of impact – the global CO_2 concentration in the atmosphere. With respect to the importance of the environmental impact, the corresponding part of MESSAGE is somewhat elaborate, allowing for a more thorough treatment of this problem than in the global study.

As it describes a part of the model which was only used to some extent in the description of the global scenarios, this subsection on the environmental submodule of MESSAGE is separated from the rest of the model description. Although this somewhat breaks the format it does summarize related parts of the model in one place.

The following two equations are the basis for the environmental submodule:

$$e(t) = \mathrm{E}x(t) \qquad t = 1, \ldots, 11 \tag{12}$$

where

e is the vector of emissions of pollutants (state variables)
E is the matrix of specific emissions (input data)

$$b(t) = \lambda(-t)b(0) + \sum_{\tau=1}^{t} 5\lambda(\tau - t)e(\tau) \qquad t = 1, \ldots, 11 \tag{13}$$

where

b is the vector of concentration of pollutants (state variables)
λ is a (diagonal) matrix of coefficients expressing the rest time of pollutants in the environment (input data)

These two groups of equations account for emissions of pollutants [eqn. (12)] and ambient concentrations of selected pollutants [eqn. (13)]. There are three levels at which these variables can be used within MESSAGE. Firstly, they may be used as monitoring variables only. In this case, they merely quantify the environmental impact of an energy supply strategy in natural units. Secondly, these variables may be constrained thus reflecting emission or concentration standards in the model. Finally, they may be included in the objective function thus directly participating in the optimization. The last of these three levels can be interpreted as multiobjective optimization. An example of such a joint optimization of economic and environmental aspects in a mathematical programming model is found in Jansen (1977).

Input data requirements for the environmental submodule of MESSAGE depend on the extent to which it is used for a particular model run. If it is fully used, the data required are on specific emissions of pollutants by energy conversion facilities, rest times of pollutants, the initial concentrations of pollutants, emission and concentration standards, and "cost data" (objective function coefficients) for emission and concentration variables.

The formal relevance of the emission and concentration variables is an extension of the state space of the model and the corresponding description of the results in these terms. Constraining the range of these variables means a reduction of the feasible region; their inclusion into the performance index means a change in the preference ranking of the energy conversion alternatives. As far as the inclusion in the objective function is concerned, it should be noted that the assumption of linearity means rather oversimplifying the real-world system.

4 EXPERIENCE AND CONCLUSIONS

MESSAGE was developed for the application to geographical regions the size of continents. It may also be applied to smaller regions or countries, provided that some care is taken in supplying the input data and in interpreting the model results. A particular problem that may arise comes from the continuity of the model variables that – for small countries – may very likely result in sizes of energy conversion facilities that are unrealistically small. Also, in some regions or countries the energy system may have some peculiarities which have not been considered in the general model formulation.

Both problems appeared in applying the model to Austria, because of her small size and her heavy reliance on hydro storage. The results of this application showed that the difficulty of continuous variables was less serious than might have been expected. Even continuous solutions yielded a good enough first set of scenarios (Schrattenholzer 1979). The continuous solutions were then adjusted by trial and error to arrive at reasonable block sizes of power plants. This method was considered preferable to using mixed integer programming, a method similar to linear programming, but permitting restriction of variables to discrete values. The difficulty relating to Austria's heavy reliance on hydro storage was solved by a modification of the original model.

Another important application of MESSAGE besides its usage within the global model was the one for the Commission of the European Communities (CEC). In contrast to the application for Austria, where the emphasis was on the transfer of methodology, the CEC application emphasized the disaggregation of global results. This was achieved by splitting IIASA's Region III into "Europe of the Nine" and "Rest of the Region" using a modified model loop. The results of the IIASA models were then compared with "bottom-up" model runs performed by the CEC. A description of this work contrasting regional aspirations with globally consistent scenarios, is contained in Commission of the European Communities (1980).

Other applications underway (such as for Brazil, Bulgaria, the FRG, and Hungary) seem to prove that the definition of MESSAGE is general enough to serve as a basis for a great variety of applications.

Work is currently being undertaken at IIASA to extend MESSAGE. The extensions are the result of modeling the energy chain up to useful energy, of running the model for a longer time period, and of modeling countries for which detailed data are available. The extended model is called MESSAGE II and is documented in Messner (1982).

ACKNOWLEDGMENTS

The author is indebted to M. Agnew and A. Voss who were the co-builders of the original MESSAGE model. In discussing earlier drafts of this report helpful suggestions were obtained from J. Casti, A. Lewandowski, A. Papin, and A. Wierzbicki. Helpful comments were also provided by W. Häfele, K. Hoffmann, and H. Vos. It is impossible to name all those who in some way have contributed to the finalization of this paper; instead the author wishes to emphasize the fruitful environment provided by IIASA in general and by the Energy Systems Program, led by Professor W. Häfele, in particular.

REFERENCES

Agnew, M., L. Schrattenholzer, and A. Voss (1978) A User's Guide for the MESSAGE Computer Program. RM-78-26. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Agnew, M., L. Schrattenholzer, and A. Voss (1979) A Model for Energy Supply Systems and their General Environmental Impact. WP-79-6. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Casti, J. (1980) Mathematical System Theory and System Modeling. WP-80-68. Laxenburg, Austria: International Institute for Applied Systems Analysis.
Commission of the European Communities (1980) Crucial Choices for the Energy Transition. Bâtiment Jean Monnet, Luxembourg: Directorate-General, Scientific and Technical Information and Information Management.
Energy Systems Program Group (1981) Energy in a Finite World, A Global Systems Analysis. Report by the Energy Systems Program Group of the International Institute for Applied Systems Analysis, Wolf Häfele, Program Leader. Cambridge, Massachusetts: Ballinger.
Häfele, W., and A.S. Manne (1974) Strategies for a Transition from Fossil to Nuclear Fuels. RR-74-4. Laxenburg, Austria: International Institute for Applied Systems Analysis.
Häfele, W. (1980) Putting the results of the IIASA Energy Systems Program to work. IIASA Reports 3(1):47–68. Laxenburg, Austria: International Institute for Applied Systems Analysis.
Jansen, P.J. (1977) Decision under Several Objectives. Technical Report SOL 77-20. Stanford, California: Systems Optimization Laboratory, Department of Operations Research, Stanford University.
Kalman, R.E. (1974) Comments on the scientific aspects of modeling, in Towards a Plan of Action for Mankind. M. Marois, ed., Amsterdam: North-Holland Publishing Co.
Kononov, Y.D., and A. Pór (1979) The Economic IMPACT Model. RR-79-8. Laxenburg, Austria: International Institute for Applied Systems Analysis.
Lapillone, B. (1978) MEDEE-2: A Model for Long-Term Energy Demand Evaluation. RR-78-17. Laxenburg, Austria: International Institute for Applied Systems Analysis.
Lewandowski, A. (1980) Issues in Model Validation. Laxenburg, Austria: International Institute for Applied Systems Analysis.
Marchetti, C., and N. Nakićenović (1979) The Dynamics of Energy Systems and the Logistic Substitution Model. RR-79-13. Laxenburg, Austria: International Institute for Applied Systems Analysis.
Messner, S. (1982) A User's Guide for MESSAGE II. Forthcoming. Laxenburg, Austria: International Institute for Applied Systems Analysis.
Propoi, A.I. (1977) Dual Systems of Dynamic Linear Programming. RR-77-9. Laxenburg, Austria: International Institute for Applied Systems Analysis.
Schrattenholzer, L. (1979) Modelluntersuchungen langfristiger Strategien der Energieversorgung Österreichs. Doctoral thesis. Vienna, Austria: Technical University.
Schrattenholzer, L. (1982a) Documentation of the World Regional MESSAGE Runs. Forthcoming. Laxenburg, Austria: International Institute for Applied Systems Analysis.
Schrattenholzer, L. (1982b) User's Guide for the MESSAGE Computer Program. Forthcoming. Laxenburg, Austria: International Institute for Applied Systems Analysis.
Suzuki, A. (1975) An Extension of the Häfele–Manne Model for Assessing Strategies for a Transition from Fossil Fuels to Nuclear and Solar Alternatives. Laxenburg, Austria: International Institute for Applied Systems Analysis.

APPENDIX A

This appendix defines the seven world regions and presents some high-level input data for these regions, giving some idea of the model boundaries of the MESSAGE runs for the global scenarios.

The seven world regions are:

Region I North America
Region II Soviet Union and Eastern Europe
Region III Western Europe, Japan, Australia, New Zealand, South Africa, and Israel
Region IV Latin America
Region V Africa (except Northern and South Africa), South and Southeast Asia
Region VI Middle East and Northern Africa
Region VII China and Other Asian Countries with Centrally Planned Economies

The input data in this appendix give, for each region, the names of the demand sectors and the names of the technologies supplying these sectors. The list of primary energy carriers is the same for each region and will therefore be given only once (at the end of this appendix). It does not contain renewable energy sources as these are treated differently (under the description of the technologies). For complete information about all input data, the reader is referred to Schrattenholzer (1982a).

Region 1

Demand sectors:
Electricity
Liquid fuels
Solid fuels
Gaseous fuels
Soft solar

Technologies:

1. For electricity generation:
   Light water reactor (LWR)
   Fast breeder reactor (FBR)
   Coal-fired power plant
   Coal-fired power plant (advanced)
   Hydroelectric power plant
   STEC (centralized solar power plant)
   Liquid fuel power plant
   Gas-steam power plant
   Gas turbines
2. Production of liquid fuels:
   Crude oil refinery
   Coal liquefaction
3. Solid fuels:
   Coal
4. Gaseous fuels:
   Natural gas
   Coal gasification
5. Soft solar:
   Local solar energy conversion facilities

Region II

Demand sectors:
Electricity
Liquid fuels
Solid fuels

Gaseous fuels
Soft solar
Heat I
Heat II

The difference between the two heat sectors is that "Heat I" means demand for heat which can be supplied by central sources. "Heat II" is heat which can only be supplied by smaller units.

Technologies:

1. For electricity generation:
   Light water reactor (LWR)
   Fast breeder reactor (FBR)
   Coal-fired power plant
   Coal-fired power plant (advanced)
   Hydroelectric power plant
   STEC (centralized solar power plant)
   Liquid fuel power plant
   Gas-steam power plant
   Gas turbines
2. Combined production of Heat I and electricity:
   Light water reactor (LWR)
   Coal-fired power plant
   Liquid fuel power plant
   Gaseous fuel power plant
3. Production of liquid fuels:
   Crude oil refinery
   Coal liquefaction
4. Solid fuels:
   Coal
5. Gaseous fuels:
   Natural gas
   Coal gasification
6. Soft solar:
   Local solar energy conversion facilities
7. Heat II:
   Light water reactor (LWR)
   Coal-fired power plant
   Liquid fuel power plant
   Gaseous fuel power plant

Region III

Demand sectors:

Electricity
Liquid fuels

Solid fuels
Gaseous fuels
Soft solar
Heat

Technologies:

1. For electricity generation:
   Light water reactor (LWR)
   Fast breeder reactor (FBR)
   Coal-fired power plant
   Coal-fired power plant (advanced)
   Hydroelectric power plant
   STEC (centralized solar power plant)
   Liquid fuel power plant
   Gas-steam power plant
   Gas turbines
2. Production of liquid fuels:
   Crude oil refinery
   Coal liquefaction
3. Solid fuels:
   Coal
4. Gaseous fuels:
   Natural gas
   Coal gasification
5. Soft solar:
   Local solar energy conversion facilities
6. Heat:
   Geothermal heat
   Combined production of heat and electricity

Region IV

Demand sectors:
Electricity
Liquid fuels
Solid fuels
Gaseous fuels
Soft solar
Heat
Renewables

Technologies:

1. For electricity generation:
   Light water reactor (LWR)
   Fast breeder reactor (FBR)

Coal-fired power plant
Coal-fired power plant (advanced)
Hydroelectric power plant
Hydroelectric power plant (expensive)
STEC (centralized solar power plant)
Liquid fuel power plant
Gas-steam power plant
Gas turbines

2. Production of liquid fuels:
   Crude oil refinery
   Coal liquefaction
3. Solid fuels:
   Coal
4. Gaseous fuels:
   Natural gas
   Coal gasification
5. Soft solar:
   Local solar energy conversion facilities
6. Heat:
   Local solar energy conversion facilities
7. Renewables:
   Fuel wood

Region V

Demand sectors:
Electricity
Liquid fuels
Solid fuels
Gaseous fuels
Soft solar
Heat
Renewables

Technologies:

1. For electricity generation:
   Light water reactor (LWR)
   Fast breeder reactor (FBR)
   Coal-fired power plant
   Coal-fired power plant (advanced)
   Hydroelectric power plant
   Hydroelectric power plant (expensive)
   STEC (centralized solar power plant)
   Liquid fuel power plant
   Gas-steam power plant
   Gas turbines

2. Production of liquid fuels:
   Crude oil refinery
   Coal liquefaction
3. Solid fuels:
   Coal
4. Gaseous fuels:
   Natural gas
   Coal gasification
5. Soft solar:
   Local solar energy conversion facilities
6. Heat:
   Combined production of heat and electricity

Region VI

Demand sectors:
Electricity
Liquid fuels
Solid fuels
Gaseous fuels
Soft solar
Heat

Technologies:
1. For electricity generation:
   Light water reactor (LWR)
   Fast breeder reactor (FBR)
   Coal-fired power plant
   Coal-fired power plant (advanced)
   Hydroelectric power plant
   STEC (centralized solar power plant)
   Liquid fuel power plant
   Gas-steam power plant
   Gas turbines
2. Production of liquid fuels:
   Crude oil refinery
   Coal liquefaction
3. Solid fuels:
   Coal
4. Gaseous fuels:
   Natural gas
   Coal gasification
5. Soft solar:
   Local solar energy conversion facilities
6. Heat:
   Combined production of heat and electricity

Region VII

Demand sectors:
Electricity
Liquid fuels
Solid fuels
Gaseous fuels
Soft solar
Heat

Technologies:

1. For electricity generation:
   Light water reactor (LWR)
   Fast breeder reactor (FBR)
   Coal-fired power plant
   Coal-fired power plant (advanced)
   Hydroelectric power plant
   STEC (centralized solar power plant)
   Liquid fuel power plant
   Gas turbines
2. Production of liquid fuels:
   Crude oil refinery
   Coal liquefaction
3. Solid fuels:
   Coal
4. Gaseous fuels:
   Natural gas
   Coal gasification
5. Soft solar:
   Local solar energy conversion facilities
6. Heat:
   Combined production of heat and electricity

Resources and Fuels (Same for all Regions)

Coal
Crude oil
Natural gas
Natural uranium
Plutonium

APPENDIX B

This appendix gives the general definition of a dynamic linear programming (DLP) model as described in Propoi (1977) and as used in this report. In the general definition, each part of a DLP is formulated by alternative descriptions; however, the description here is restricted to only those variants that apply for MESSAGE.

I. State equations

$$x(t+1) = \sum_{i=1}^{\nu} \mathrm{A}(t-n_i)x(t-n_i) + \sum_{j=1}^{\mu} \mathrm{B}(t-m_j)u(t-m_j) \qquad t = 0, \ldots, T-1$$

where

x is the vector of state variables
u is the vector of control variables
A, B are matrices (model input data)

II. Constraints

$$\mathrm{G}(t)x(t) + \mathrm{D}(t)u(t) \leqslant f(t)$$

where

G, D are matrices (input data)
f is a vector (input data)

III. Boundary Conditions

$$x(0) = x^0$$

IV. Planning Period

T is fixed.

V. Performance Index (Objective Function)

$$J(u) = [a(T), x(T)] + \sum_{t=0}^{T-1} \{[a(t), x(t)] + [b(t), u(t)]\} \rightarrow \max\ (\min)$$

where

a, b are vectors of cost coefficients (input data)

The variables of MESSAGE, disaggregated between state and control variables are:

Control variables:
$r(t)$ (annual consumption of resources)
$x(t)$ (energy production)
$z(t)$ (annual additions to capacity)

State variables:
$b(t)$ (concentrations of pollutants)
$c(t)$ (capacities of technologies)
$e(t)$ (annual emissions of pollutants)
$s(t)$ (stocks of resources)

ABSTRACTS OF OTHER IIASA PUBLICATIONS

Kunreuther, H., and P. Kleindorfer, Guidelines for Coping with Natural Disasters and Climatic Change. IIASA Research Report RR-81-15, July 1981.
Reprinted from *Zeitschrift für Umweltpolitik*, Vol. 3, 1980, pp. 887–902.

This paper presents a conceptual framework for evaluating alternative programs that deal with natural hazards and weather-related risks. The framework stresses the importance of integrating descriptive or behavioral analysis with prescriptive measures. In particular there is a need to understand the institutional arrangements and decision processes of the interested parties as a prelude to formulating and evaluating alternative plans.

Empirical evidence on the decision processes associated with the purchase of disaster insurance suggests ways of linking descriptive and prescriptive analysis for dealing with natural hazards. A parallel type of analysis is undertaken for evaluating what action, if any, should be taken with respect to the manufacturing of aerosol spray cans, which use chlorofluorocarbons (CFC), given that CFC destroys ozone in the atmosphere.

Sauberer, M., Migration and Settlement: 10. Austria. IIASA Research Report RR-81-16, July 1981.

In this report, Dr. Michael Sauberer, Director of the Austrian Institute for Regional Planning, shows how multiregional population analysis may be used to describe and analyze the changes taking place in the size and distribution of populations over time. Adopting a two-level multiregional system (nine provinces and four regions) for his study of Austria, he demonstrates that it is important to consider the interdependence between the provinces when studying populations at the subnational level.

Foell, W.K., et al., The Wisconsin–IIASA Set of Energy/Environment (WISE) Models for Regional Planning and Management: An Overview. IIASA Research Report RR-81-17, August 1981.

This report presents an overview of the analytical framework and quantitative models used in the IIASA case studies on Regional Energy/Environment Management and Planning. Its purpose is to summarize the structure of the models, to provide a complete listing of the sources of more detailed model and data descriptions, and to indicate how the models are integrated to provide a foundation for regional energy/environment policy analysis. Within this context the term 'region' denotes geographic or administrative units, ranging from small countries such as Austria to subnational regions such as the state of Wisconsin in the USA. The audience for the report includes managers, planners, technical advisors, and modelers.

Tomlinson, R., Some Dangerous Misconceptions Concerning Operational Research and Systems Analysis. IIASA Research Report RR-81-19, September 1981.
Reprinted from the *European Journal of Operational Research*, Vol. 7(2), 1981, pp. 203–212.

After defining the field of interest as operational research and/or applied systems analysis (ORASA), this paper examines ORASA by describing and then qualifying seven 'near-truths' about the subject; each is sufficiently accurate to be accepted by many ORASA practitioners, but equally each contains the seeds of dangerous misconceptions and distortions if its limitations are not recognized. The seven near-truths and the corresponding qualifications thought necessary may be summarized as follows. 1, "ORASA is problem solving": certainly an important aspect but achievement of understanding is the more fundamental goal. 2, "Models are central to ORASA": what is truly 'central' is complex reality even though the models may be 'essential' to the process. 3, "Problems can and must be defined (uniquely and invariantly)": definition is vital but redefinition in the light of increased understanding or the views of different analysts is equally necessary. 4, "Models are partial representations of reality": often the aim, but it is important that the exact meaning and range of validity of the models must be understood by their interpreters and the relationship between models and reality very closely examined. 5, "Tactics and strategy are entirely separate": it is misleading to postulate an absolute qualitative distinction – both overviews and problem-solving applications are needed and successful ORASA analysts must combine the two. 6, "All rigorous thought can be expressed in mathematical terms": rigor is necessary but, given the nature of real-world problems, it is essential to combine 'hard' mathematical analysis with equally valid insight from the 'soft' sciences. 7, "ORASA is a science": scientific in its approach but multidisciplinary and cross-cutting in the expertise on which it draws.

Dziewoński, K., and P. Korcelli, Migration and Settlement: 11. Poland. IIASA Research Report RR-81-20, October 1981.

In this report, Professors Kazimierz Dziewoński and Piotr Korcelli, of the Institute of Geography and Spatial Organization, Polish Academy of Sciences, analyze the changing population patterns in Poland and their relations to spatial policy. The analysis focuses on regional interdependence and the role of major urban agglomerations in the spatial population system.

Philipov, D., Migration and Settlement: 12. Bulgaria. IIASA Research Report RR-81-21, October 1981.

In this report, Dimiter Philipov analyzes recent changes in Bulgaria's patterns of population redistribution and studies in detail the demographic dynamics of seven economic planning regions.

Csáki, C., A National Policy Model for the Hungarian Food and Agriculture Sector. IIASA Research Report RR-81-23, October 1981.

The development at IIASA of the Hungarian Agricultural Model (HAM), as a prototype of models of centrally planned food and agriculture systems was completed at the end of 1979. The model is a joint undertaking of the Food and Agriculture Program at IIASA and three institutes in Hungary. The results of the entire three-year HAM project are summarized in this paper. HAM is a descriptive, recursive simulation model describing the Hungarian food and agriculture system as a disaggregated part of an economic system closed at the national and the international levels. The model, which will ultimately become one of a system of interconnected models, is structured according to the major elements of the centrally planned food and agriculture systems. Two spheres are differentiated within the model. The government economic management and planning submodel describes the decision making and control activities of the government. The production submodel deals with the fulfillment of central plan targets, covering the whole national economy. The general structure of the model and its mathematical description are discussed first. Two versions of HAM have in fact been produced. HAM-1 is a relatively aggregated model (10 food and agricultural commodities are considered); HAM-2 is more disaggregated (45 commodities are considered) and further refined. The two models are described in separate parts of the report, together with the results of the validation procedure and the conclusions of the actual calculations.

Rempel, H., Rural–Urban Labor Migration and Urban Unemployment in Kenya. IIASA Research Report RR-81-24, October 1981.

The starting point of this study is a model of rural household decision-making, which generates a set of testable hypotheses regarding the determinants and consequences of rural–urban migration. A survey of one of Kenya's eight largest urban centers was carried out in December 1968 to provide data that were then combined with census data to test these hypotheses. The questionnaire that was distributed was designed to obtain the migration, employment, and income history of each migrant from 1 year before his move to the time of the survey as well as the migrant's opinion on why he moved, how long he intended to stay, and what he thought of life in urban centers. This volume is an analysis of those data. The basic thesis is that rural–urban migration is a rational response to development in Kenya. Migration does not shape this development; it is merely one symptom of growth. On the basis of the results obtained, the study concludes with a general discussion of several aspects of the urbanization process that can be influenced by policy actions.

Sassin, W., On Energy and Economic Development. IIASA Research Report RR-81-25, October 1981.
Reprinted from *Scientific American*, Vol. 243(3), 1980, pp. 118–131.

This paper presents a complementary view of IIASA's energy analysis. Based on the realization that technoeconomic approaches often "solve" problems by splitting them and transferring their parts elsewhere, it searches for the deeper roots of obvious failures in the evolution of the energy segment of man's vital infrastructure, and seeks to identify the limits of sound principles that govern past and present decisions and that become evident when we examine such failures.

From this point of view, the author describes a serious conflict between macroeconomic and microeconomic decisions that will occur if scientific and technological progress does not compensate for a decline in the quality of the world's energy resources. Resolving this conflict is an important matter for institutional and political forces to deal with.

Fedra, K., G. van Straten, and M.B. Beck, Uncertainty and Arbitrariness in Ecosystems Modelling: A Lake Modelling Example. IIASA Research Report RR-81-26, October 1981.
Reprinted from *Ecological Modelling*, Vol. 13, 1981, pp. 87–110.

Mathematical models of ecosystems are considerable simplifications of reality, and the data upon which they are based are usually scarce and uncertain. Calibration of large complex models depends upon arbitrary assumptions and choices, and frequently calibration procedures do not deal adequately with the uncertainty in the data describing the system under study. Since much of the uncertainty and arbitrariness in ecological modelling is inevitable, because of both practical as well as theoretical limitations, model-based predictions should at least reveal their dependence on, and sensitivity to, uncertainty and arbitrary assumptions.

This paper proposes a method that explicitly takes into account the uncertainty associated with data for modelling. By reference to a partly qualitative and somewhat vague definition of system behaviour in terms of allowable ranges, an ensemble of acceptable parameter vectors for the model may be identified. This contrasts directly with a more conventional approach to model calibration, in which a quantitative (squared-error) criterion is minimized and through which a supposedly 'unique' and 'best' set of parameters can be derived. The ensemble of parameter vectors is then used for the simulation of a multitude of future systems behaviour patterns, so that the uncertainty in the initial data and assumptions is preserved, and thus the predicted future systems response can be interpreted in a probabilistic manner.

Arthur, W.B., The Analysis of Causal Linkages in Demographic Theory. IIASA Research Report RR-81-27, December 1981.

Many seemingly different questions that interest demographers can be phrased as the same technical question: how, within a given demographic model, would variable y change if the age- or time-specific function f were to change arbitrarily in shape and intensity? At present demography lacks the machinery to answer this question in analytical and general form.

This paper suggests a method, based on modern functional calculus, for deriving closed-form expressions for the sensitivity of demographic variables to changes in input functions or schedules. It uses this "causal linkage" method on three bodies of theory:

stable population analysis, nonstable or transient population analysis, and techniques for the estimation of incomplete demographic data.

In stable theory, closed-form expressions are obtained for the response of the intrinsic growth rate, birth rate, and age composition to arbitrary marginal changes in the age patterns of fertility and mortality.

In nonstable theory, expressions are obtained for the transient response of the age composition to time-varying changes in the birth sequence, and to changing age-specific fertility and mortality patterns. The problem of "bias" in period vital rates is also looked at.

In incomplete-data analysis, a general format for robustness or error analysis is suggested; this is applied to a standard Brass estimation technique.

Leonardi, G., A Unifying Framework for Public Facility Location Problems. IIASA Research Report RR-81-28, November 1981.
Reprinted from *Environment and Planning A*, Vol. 13, 1981, pp. 1001–1028.

This paper, a condensed report of the present state of the work in the Public Facility Location Task (formerly the Normative Location Modeling Task) at IIASA, has three main aims: first, to build a general framework for location problems; second, to use this framework to unify existing location models; and third, to use the framework to develop new, more general, and more meaningful location models. Suggestions are also given on how to introduce multiple services and multiple time periods in location problems. The multiactivity dynamic location models that this perspective generates are the subject of future research in the Public Facility Location Task.

This first part of the paper gives a nontechnical description of the proposed general framework for analyzing location problems. The second part will describe mathematical models for static, single-service, facility location problems and their possible extensions and improvements, and will appear in the next issue.

Marchetti, C., Society as a Learning System: Discovery, Invention, and Innovation Cycles Revisited. IIASA Research Report RR-81-29, November 1981.
Reprinted from *Technological Forecasting and Social Change*, Vol. 18, 1980, pp. 267–282.

The very simple heuristic suggestion that society as a whole and its numerous subsets operate like learning systems, basically governed by Volterra–Lotka equations, has been extremely valuable in organizing a most variegated collection of statistical sets of time series, ranging from the structure of energy markets to the efficiency of machinery and the expansion of empires. In this paper an attempt is made to treat invention and entrepreneurship, generally perceived as the most "free" of human activities but actually subject to iron rules. Invention and innovation during the last 250 years appear in precisely structured waves that lend themselves to robust prediction. The present wave will reach its maximum momentum around 1990. Furthermore, the introduction, maximum market penetrations, and prices of new primary energies show a very strong link to these innovation waves. This stresses once more that economic features may be the expression of deeper "physical" phenomena related to the basic working of society and thus become predictable up to a point through a very abstract and noneconomic analysis.

This work has been done in the frame of IIASA's Energy Systems Program and can be considered as an outgrowth of and complement to the research on the evolution of energy systems described in IIASA Research Reports 79-12, 79-13, and 77-22. There it was found that a new primary energy coming into the market must be observed for 10 or 20 years if one is to extract the basic features necessary to predict its long-term market behavior. Specifically, it was concluded that the dates at which new primary energies come into play cannot be predicted. In this paper innovations are considered not one by one but as an abstract set, whose behavior is analyzed. In this frame possible birth dates for new energy sources can be identified, thus enhancing the quality of very long-term forecasting in the energy field. Also, prices appear predictable, at least in their gross features.

BIOGRAPHIES

Robert J. Anderson, USA

Robert J. Anderson, a vice-president and Director of Economics at MATHTECH Inc. in Princeton, New Jersey, joined IIASA's Resources and Environment Area in December 1979. He was appointed Deputy Chairman of the Area in August 1980. Dr. Anderson graduated from Carleton College in 1965 and received his Ph.D. in Economics from the University of Pennsylvania in 1969. He served in the U.S. Public Health Service with the National Air Pollution Control Administration from 1967 to 1969. He was Assistant Professor of Economics at Purdue University from 1969 to 1972. In 1971 and 1972 he was a senior economist with CONSAD Research Corporation. In 1972 and 1973 he was Acting Associate Professor of Economics and Administration at the University of California, Riverside and in 1973 and 1974 he was Associate Professor of Economics and Director of the Center for the Study of Environmental Policy at Pennsylvania State University. Dr. Anderson's research interests include applied welfare economics and measurement of the benefits of environmental quality improvement.

Luis J. Castro, Mexico

Luis Castro came to IIASA from the Universidad Nacional Autonoma de Mexico to work on a comparative study of migration and settlement in the 17 nations associated with the Institute and on a case study of urbanization and development in Mexico. His current research is focused on age patterns of migration and their underlying cause-specific components.

Arshad M. Khan, Pakistan

Arshad M. Khan gained his Ph.D. in Physics at the University of Birmingham, UK in 1964. In 1968 he joined the Pakistan Institute of Nuclear Science and Technology and did experimental work on the nuclear physics of neutron capture reactions. At the same time, Dr. Khan was closely associated with the science and technology planning activities in Pakistan. He joined IIASA's Energy Systems Program in 1978 to study the long-term energy requirements of developing countries, and the options and alternative strategies that could be applied to their particular circumstances. He is now Head of the Applied Systems Analysis Group at the Pakistan Atomic Energy Commission.

Bořivoj Melichar, Czechoslovakia

Bořivoj Melichar studied electrical engineering and control engineering at the Czech University of Technology and received a diploma in electrical engineering in 1964. In 1978 he completed a dissertation on the translation of cooperative lists into programming languages. Since 1964 he has been an assistant professor in computing at the Czech University of Technology in Prague. Dr. Melichar's present interests include the syntax and semantics of computer languages and the design and implementation of computer language compilers.

Andrei Rogers, USA

Andrei Rogers has been Chairman of the Human Settlements and Services Area at IIASA since October 1976. He came to IIASA from Northwestern University, Illinois, USA. His current research focuses on migration patterns and the evolution of human settlement systems in both developed and developing countries.

Leo Schrattenholzer, Austria

Leo Schrattenholzer joined IIASA's Energy Systems Program in November 1973. His work is mainly on linear programming models. Dr. Schrattenholzer received his degree in Mathematics in 1973 and his Ph.D. in 1979, both from the Technical University of Vienna. His Ph.D. thesis was on modeling long-term energy supply strategies for Austria. From 1972 to 1974 he was a research and lecture assistant with the Institute of Mathematics, I, of the Technical University. His scientific interests include linear programming, dynamical systems, probability theory, statistics, and information and data processing.

IIASA
REPORTS

Volume 4 Number 2 October-December 1981
Completing Volume 4

Published quarterly

CONTENTS

R.J. ANDERSON, JR. An Economic Analysis of Supplementary Irrigation in Skåne 199
A. ROGERS AND L.J. CASTRO Model Migration Schedules 239
A.M. KHAN The Possible Share of Soft/Decentralized Renewables in Meeting the Future Energy Demands of Developing Regions 291
B. MELICHAR Nonprocedural Communication between Users and Application Software 319
L. SCHRATTENHOLZER The Energy Supply Model MESSAGE 353

Abstracts of other IIASA Publications 381
Biographies 387

IN FORTHCOMING ISSUES:

JOHN C. STONE, M. GADKOWSKI, A. SALEWICZ, W. SIKORSKI, AND F. DAIL SINGLETON, JR. Water Demand for Generating Electricity in Poland: A Programming Approach

ARSHAD M. KHAN AND ALOIS HOELZL Energy Demand to 2030 for Six World Regions: Estimates for Two IIASA Scenarios

F.D. McCARTHY AND W.M. MWANGI Kenyan Agriculture: Toward 2000

Printed by NOVOGRAPHIC, Vienna, Austria.