SERIE RESEARCH MEMORANDA

A MODEL FOR THE EMPLOYMENT PATTERN OF CONSTRUCTION PROJECTS

by

Pieter H.F.M.van Casteren

and

Arnold H.Q.M. Merkies

Research Memorandum 1989-78

December 1989

VRIJE UNIVERSITEIT

FACULTEIT DER ECONOMISCHE WETENSCHAPPEN

EN ECONOMETRIE

A M S T E R D A M

A MODEL for the EMPLOYMENT PATTERN of CONSTRUCTION PROJECTS

Pieter H.F.M.van Casteren
Faculty of Economics and Econometrics, University of Amsterdam,
1011 NH Amsterdam, The Netherlands

Arnold H.Q.M. Merkies
Faculty of Economics and Econometrics, Vrije Universiteit,
1007 MC Amsterdam, The Netherlands

A commonly used instrument for employment policies is the government's construction program. Efficient use of this instrument requires reliable predictions of the generated employment patterns through time. In this paper we present a model for the employment pattern of construction projects. It gives model employment as the result of inflow and outflow of workers, which appears to be an improvement of earlier models that generate employment directly. The model is applied on data of a small, a medium-size and a large project. In each case we estimate and test several specifications and select the efficient ones, using a new selection criterion (ADC). The performance of the inflow-outflow model is compared with that of a simpler direct model of employment.

KEY WORDS: Employment, Construction industry, Specification, Polynomial Curve Fitting, Model Selection.

1. INTRODUCTION

Full employment policies are often performed by stimulating the construction industry. In such a context it is of relevance to know when exactly the employment will manifest itself. This depends upon the composition of the construction program, upon the time patterns of employment of the various types of construction projects and upon the exact starting points of these projects. Merkies and Bikker (1981) have presented a model that describes how these elements are connected. The present paper attempts to improve on their description of the time pattern of employment of individual construction projects.

The analysis of Merkies and Bikker was based upon rather cursory measurements by the Dutch Central Statistical Office (C.B.S.) of accumulated costs. The C.B.S. has discontinued the collection of such data. For this paper we have collected our own sample. Thus we were also able to create more appropriate data: direct information about employment through time.

Due to this superior information we could improve on the model referred to above. The earlier model approximated the patterns by a gamma density. Below, we present a kind of spline model. For a treatment of spline functions see for instance Poirier (1976). Employment is given as the difference between (accumulated) inflow and outflow, which are both polynomial functions of time on the relevant intervals. The model is appropriate for classes of relatively homogeneous projects such as the dwelling projects we have selected for our application.

The plan of the paper is as follows. In section 2 we describe various possible specifications of the inflow-outflow model. In section 3 we discuss the estimation and search for efficient model specifications. Section 4 compares the performance of the inflow-outflow model with that of a direct polynomial model and presents an improved version of the inflow-outflow model. Section 5 summarizes and gives conclusions.

2. THE MODEL

2.1 BACKGROUND

The description of employment as the difference between inflow and outflow of workers originates from the way labor input is planned in the construction industry. Constructors consider projects as the completion of a number of operations to be consecutively executed on each dwelling or building. Ideally operations can be defined such that they require each the same operation time and a single operation team. In practice operation times tend to increase (or decrease) as time proceeds. If such changes are linear - which we will assume later - accumulated inflow and outflow of workers are similar functions of time. As illustrated in Figure 1, their difference along the vertical axis is the employment level, their difference along the horizontal axis is the operation time. Since the inflow-outflow model explains how employment arises, it is theoretically more appealing and easier to interpret than direct descriptions of employment patterns through time. It also gives better instruments for control.

2.2 THE BASIC MODEL

Let $\mu(t|\theta)$ be the number of construction workers that theoretically should be present at the site at time t, with θ a vector of parameters that characterize the type of the construction project. This number is the net result of the accumulated theoretical inflow of construction workers $\vec{\mu}(t|\theta)$ and the accumulated theoretical outflow $\overleftarrow{\mu}(t|\theta)$. So

$$\mu(t|\theta) = \vec{\mu}(t|\theta) - \overleftarrow{\mu}(t|\theta) \tag{2.1}$$

The most important parameters in θ are the marking points τ_1, τ_2, τ_3 and τ_4, where

- τ_1 = the starting time of the project (start of inflow)
- τ_2 = the earliest moment at which some operation ends (start of outflow)
- τ_3 = the latest moment at which an operation starts (end of inflow)

τ_4 = the ending time of the project (end of outflow).

Clearly we have $\tau_1<\tau_2<\tau_4$ and $\tau_1<\tau_3<\tau_4$ (there is always more than one operation), and normally we have also $\tau_2<\tau_3$. We define $[\tau_1,\tau_3]$ and $[\tau_2,\tau_4]$ as the inflow and the outflow interval respectively. Both are closed intervals. All τ's are observable and therefore known.

Inflow and outflow can be described by the following functions of time.

$$\vec{\mu}(t|\theta) := \begin{cases} 0 & t\in(-\infty,\tau_1) \\ \tfrac{1}{2}\,\bar{\beta}\left(1+\dfrac{2t-\tau_3-\tau_1}{\tau_3-\tau_1}\right)+\sum\limits_{k=0}^{\bar{k}}\vec{\beta}_k\left(\dfrac{2t-\tau_3-\tau_1}{\tau_3-\tau_1}\right)^k & t\in[\tau_1,\tau_3] \\ \bar{\beta} & t\in(\tau_3,\infty) \end{cases} \qquad (2.2a)$$

$$\overleftarrow{\mu}(t|\theta) := \begin{cases} 0 & t\in(-\infty,\tau_2) \\ \tfrac{1}{2}\,\bar{\beta}\left(1+\dfrac{2t-\tau_4-\tau_2}{\tau_4-\tau_2}\right)+\sum\limits_{k=0}^{\bar{k}}\overleftarrow{\beta}_k\left(\dfrac{2t-\tau_4-\tau_2}{\tau_4-\tau_2}\right)^k & t\in[\tau_2,\tau_4] \\ \bar{\beta} & t\in(\tau_4,\infty) \end{cases} \qquad (2.2b)$$

where

$$\bar{\beta},\ \vec{\beta}_k,\ \overleftarrow{\beta}_k \in R \qquad \text{for } k=0,1,\ldots,\bar{k}$$

$$\theta = (\tau',\beta')' = (\tau',\bar{\beta},\vec{\beta}',\overleftarrow{\beta}')' = (\tau_1,\tau_2,\tau_3,\tau_4,\bar{\beta},\vec{\beta}_0,\vec{\beta}_1,\ldots,\vec{\beta}_{\bar{k}},\overleftarrow{\beta}_0,\ldots,\overleftarrow{\beta}_{\bar{k}})'.$$

Both functions become independent of the values of τ if we normalize them on the interval $[-1,1]$, which allows comparison of β-estimates between inflow and outflow and among projects. Hence

$$\vec{\mu}(x|\theta) = \tfrac{1}{2}\bar{\beta}(1+x) + \sum_{k=0}^{\bar{k}}\beta_k x^k \qquad x\in[-1,1] \qquad (2.3)$$

with

$$x = \frac{2t-\tau_3-\tau_1}{\tau_3-\tau_1} \quad \text{and} \quad \beta_k = \vec{\beta}_k$$

and similarly for outflow $\overleftarrow{\mu}(x|\theta)$

with

$$x = \frac{2t-\tau_4-\tau_2}{\tau_4-\tau_2} \quad \text{and} \quad \beta_k = \overleftarrow{\beta}_k.$$

The first term of (2.3) is called the basic development. It is completely determined by $\bar{\beta}$. The parameter vectors $\vec{\beta}$ and $\overleftarrow{\beta}$ reveal polynomial deviations from this basic pattern up to degree $\bar{k}$.

As an example of (2.2) we have taken $\bar{k}=2$. Taking $\tau_1=0$ and $\tau_3=2$ we get the inflow function

$$\vec{\mu}(t|\theta) = \tfrac{1}{2}\bar{\beta}t + \vec{\beta}_0 + \vec{\beta}_1(t-1) + \vec{\beta}_2(t-1)^2 \qquad t\in[0,2] \tag{2.4a}$$

and taking $\tau_2=1$ and $\tau_4=3$ we get the outflow function

$$\overleftarrow{\mu}(t|\theta) = \tfrac{1}{2}\bar{\beta}(t-1) + \overleftarrow{\beta}_0 + \overleftarrow{\beta}_1(t-2) + \overleftarrow{\beta}_2(t-2)^2 \quad t\in[1,3] \tag{2.4b}$$

These inflow and outflow functions with arbitrary values of the β parameters are portrayed in figure 1b and the connected basic development is drawn in Figure 1a. The resulting employment functions are given in Figures 1c and 1d.

As mentioned before the values of τ are observable and the parameters $\bar{\beta}$, $\vec{\beta}$ and $\overleftarrow{\beta}$ can therefore be estimated from comparisons of $\mu(t|\theta)$ with y_t the actual number of construction workers observed at the site at time t. We will not do so, however, before we have extended our model with some preferable theoretical restrictions.

2.3 THEORETICAL RESTRICTIONS

Although (2.2a) and (2.2b) may generate a wide variety of different patterns not all of these can represent inflow and outflow patterns of construction workers. Theoretically inflow and outflow functions must satisfy the following requirements.

1) $\vec{\mu}'(t|\theta)\geq 0$, $\overleftarrow{\mu}'(t|\theta)\geq 0$ for all t (non-decreasing inflow and outflow).
   Consequently in view of (2.2): $\vec{\mu}(t|\theta)$, $\overleftarrow{\mu}(t|\theta)\in[0,\bar{\beta}]$ for all t.
2) $\vec{\mu}(t|\theta) \geq \overleftarrow{\mu}(t|\theta)$ for all t (outflow cannot exceed inflow).

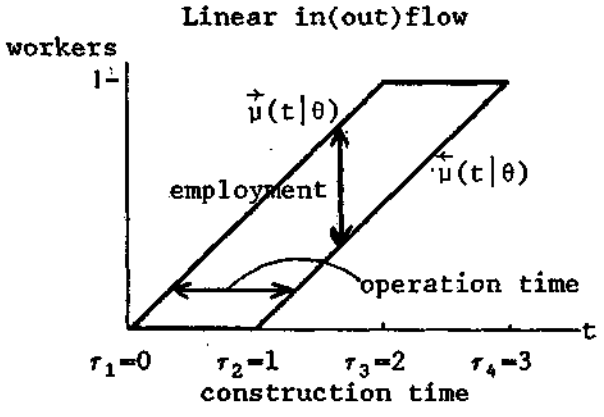
Figure 1. Inflow and outflow functions of workers in Construction Projects

Figure 1a

$\tilde{\beta}$=1 $\vec{\beta}_0$=$\vec{\beta}_1$=$\vec{\beta}_2$=0
$\overleftarrow{\beta}_0$=$\overleftarrow{\beta}_1$=$\overleftarrow{\beta}_2$=0

Figure 1b

$\tilde{\beta}$=1 $\vec{\beta}_0$=-1/4 $\vec{\beta}_1$=0 $\vec{\beta}_2$=1/4
$\overleftarrow{\beta}_0$=-1/4 $\overleftarrow{\beta}_1$=0 $\overleftarrow{\beta}_2$=1/4



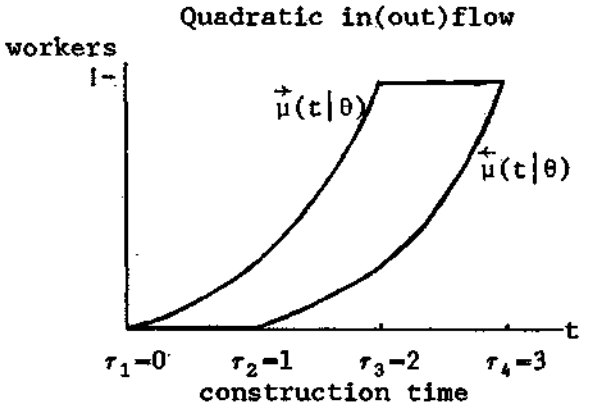


Figure 1c

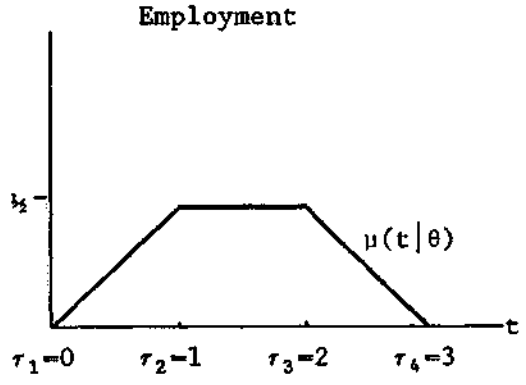
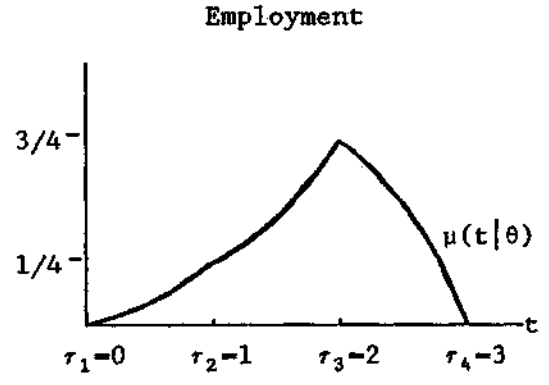


Figure 1d



To prevent that functions (2.2) violate these requirements in an empirical setting we consider some theoretical restrictions that may be imposed on the inflow and outflow functions.

A. Smoothness

The inflow and outflow functions are called smooth of degree k, if they have continuous derivatives up to order k. If we require smoothness of degree 0, the inflow and outflow functions must be continuous, so we have the following constraints:

$$\vec{\mu}(\tau_1|\theta) = 0 \tag{2.5a}$$
$$\vec{\mu}(\tau_3|\theta) = \bar{\beta} \tag{2.5b}$$
$$\overleftarrow{\mu}(\tau_2|\theta) = 0 \tag{2.5c}$$
$$\overleftarrow{\mu}(\tau_4|\theta) = \bar{\beta} \tag{2.5d}$$

Given model (2.2) this leads to

$$\sum_{k=0}^{\bar{k}} \vec{\beta}_k (-1)^k = 0 \tag{2.6a}$$
$$\sum_{k=0}^{\bar{k}} \vec{\beta}_k = 0 \tag{2.6b}$$
$$\sum_{k=0}^{\bar{k}} \overleftarrow{\beta}_k (-1)^k = 0 \tag{2.6c}$$
$$\sum_{k=0}^{\bar{k}} \overleftarrow{\beta}_k = 0 \tag{2.6d}$$

These constraints can be imposed when we estimate the β's and we may alternatively test whether their incorporation in the model is justified. In reality both inflow and outflow are discontinuous. In each of the marking points we may observe a jump, which will be ignored if we impose (2.6). This also prevents negative jumps to enter the model.

The number of possible specifications can be reduced further by demanding smoothness of degree 1. This adds the restriction that the functions must have continuous derivatives. We can go further by demanding smoothness of higher degree. Then our functions must also have continuous derivatives of higher order. Increasing the degree of smoothness from δ-1 to $\delta(-1,\ldots,\bar{k}-1)$ thus leads to the following additional restrictions on the δth derivatives $\vec{\mu}^{(\delta)}$ and $\overleftarrow{\mu}^{(\delta)}$ or equivalently on β:

$$\lim_{t\downarrow\tau_1} \vec{\mu}^{(\delta)}(t|\theta)=0 \quad \text{or} \quad \sum_{k=\delta}^{\bar{k}} \vec{\beta}_k \; k(k-1)\ldots(k-\delta+1)(-1)^{k-\delta}=0 \tag{2.7a}$$

$$\lim_{t\uparrow\tau_3} \vec{\mu}^{(\delta)}(t|\theta)=0 \quad \text{or} \quad \sum_{k=\delta}^{\bar{k}} \vec{\beta}_k \; k(k-1)\ldots(k-\delta+1) \qquad =0 \qquad (2.7b)$$

$$\lim_{t\downarrow\tau_2} \overleftarrow{\mu}^{(\delta)}(t|\theta)=0 \quad \text{or} \quad \sum_{k=\delta}^{\bar{k}} \overleftarrow{\beta}_k \; k(k-1)\ldots(k-\delta+1)(-1)^{k-\delta}=0 \qquad (2.7c)$$

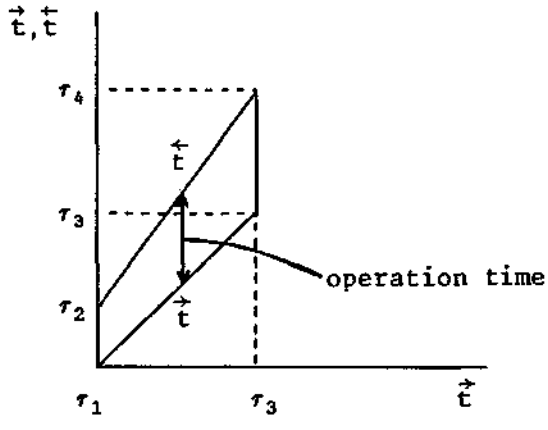
$$\lim_{t\uparrow\tau_4} \overleftarrow{\mu}^{(\delta)}(t|\theta)=0 \quad \text{or} \quad \sum_{k=\delta}^{\bar{k}} \overleftarrow{\beta}_k \; k(k-1)\ldots(k-\delta+1) \qquad =0 \qquad (2.7d)$$

except that for $\delta=1$ the term $\frac{1}{2}\vec{\beta}$ must be added to the left hand side of each second expression. The degree of smoothness is chosen such that the inflow and outflow functions appropriately meet the actually observed patterns as well as the theoretical requirements of non-decreasing inflow and outflow and positive employment.

B. Equality of shape.

Secondly we have the possible restriction that inflow and outflow are similar functions of time. This is the case if operation times increase linearly as construction time proceeds. An example is given in Figure 2, where the vertical differences $\overleftarrow{t}-\vec{t}$ are the operation times. Such figures are familiar to employment planners in the construction industry. Note that figure 1 above presents the special case of constant operation times.

Figure 2. Linear increasing operation times



The operation time of an operation starting at $\vec{t} \in [\tau_1,\tau_3]$ is a linear function of $\vec{t}$ if the corresponding ending time $\overleftarrow{t} \in [\tau_2,\tau_4]$ is given by a linear function of $\vec{t}$

$$\overleftarrow{t} - \overrightarrow{t} = p\overrightarrow{t} + q \quad \text{or} \quad \overleftarrow{t} = (p+1)\overrightarrow{t} + q$$

with p and q constants. In particular we have $\tau_2 = (p+1)\tau_1 + q$ and $\tau_4 = (p+1)\tau_3 + q$ so that p and q can be solved and the linear function is in fact

$$\overleftarrow{t} = \frac{\tau_4 - \tau_2}{\tau_3 - \tau_1}\overrightarrow{t} + \frac{\tau_2\tau_3 - \tau_1\tau_4}{\tau_3 - \tau_1} \quad \text{or} \quad \frac{2\overrightarrow{t} - \tau_3 - \tau_1}{\tau_3 - \tau_1} = \frac{2\overleftarrow{t} - \tau_4 - \tau_2}{\tau_4 - \tau_2} . \tag{2.8}$$

Thus we have a linear changing operation time, if any $\overrightarrow{t}$ and $\overleftarrow{t}$ satisfying (2.8) are the starting and ending time of the same operation, in other words if (2.8) implies $\overrightarrow{\mu}(\overrightarrow{t}|\theta) = \overleftarrow{\mu}(\overleftarrow{t}|\theta)$. In terms of the normalized variables this means that for any given $\overrightarrow{x}$, $\overleftarrow{x} \in [-1,1]$:

$$\overrightarrow{x} = \overleftarrow{x} \;\Rightarrow\; \tfrac{1}{2}\bar{\beta}(1+\overrightarrow{x}) + \sum_{k=0}^{\bar{k}} \overrightarrow{\beta}_k \overrightarrow{x}^k = \tfrac{1}{2}\bar{\beta}(1+\overleftarrow{x}) + \sum_{k=0}^{\bar{k}} \overleftarrow{\beta}_k \overleftarrow{x}^k \tag{2.9}$$

This is equivalent to

$$\overrightarrow{\beta}_k = \overleftarrow{\beta}_k \qquad k = 0, 1, \ldots, \bar{k} \tag{2.10}$$

C. Degree of the polynomial

By choosing polynomials we restrict ourselves to polynomials up to degree $\bar{k}$. This can be interpreted as a third kind of restriction. In terms of the coefficients it is:

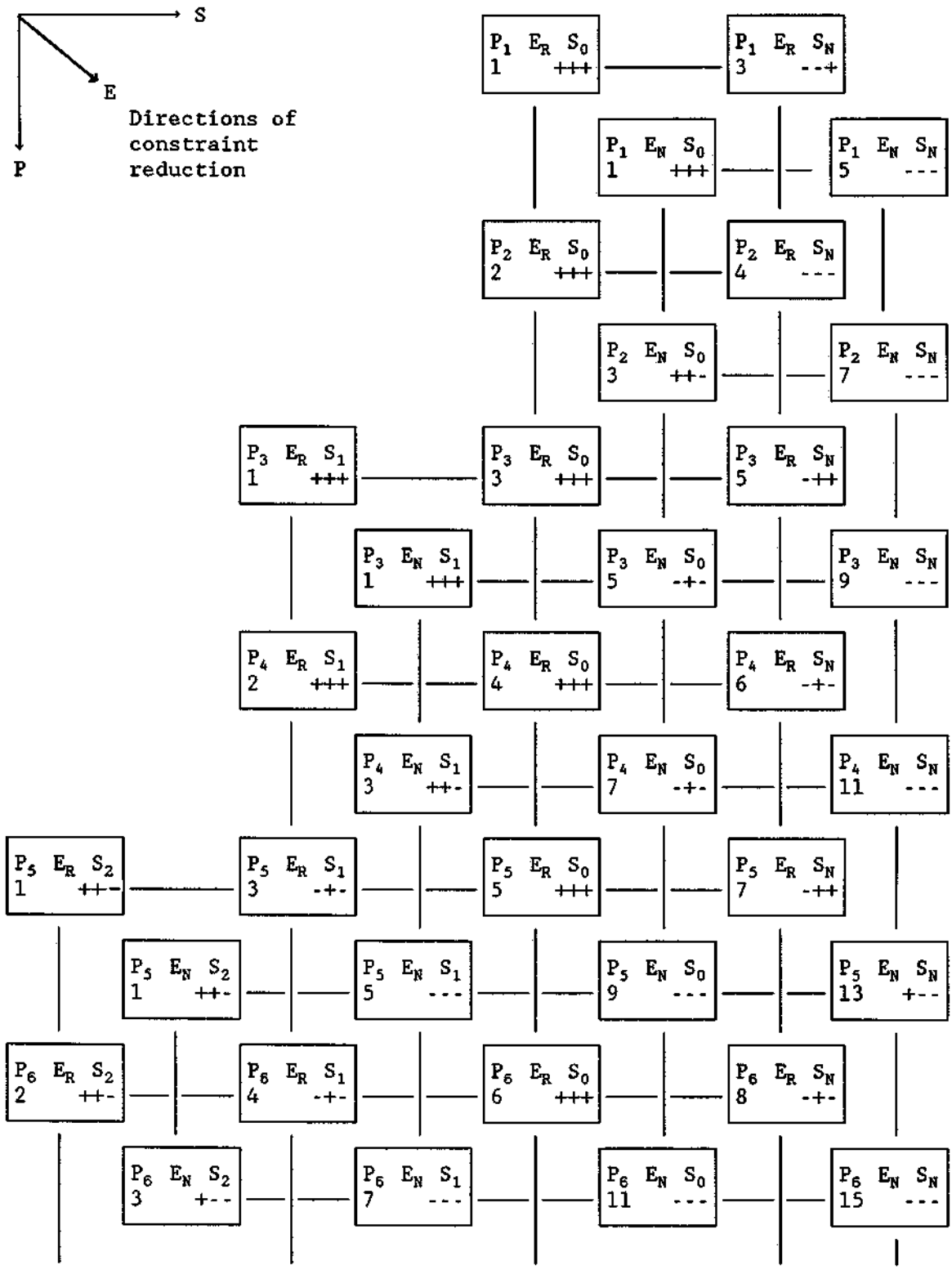
$$\overrightarrow{\beta}_k = 0 \text{ and } \overleftarrow{\beta}_k = 0 \qquad \text{for } k=\bar{k}+1, \bar{k}+2, \ldots \tag{2.11}$$

To choose a proper degree $\bar{k}$ one must weigh simplicity against greater flexibility offered by higher degree polynomials. Simpler models are easier to interpret (0=constant, 1=change, 2=convexity, 3=bend). It depends upon the empirical setup how much descriptive power can be sacrified to gain such nice interpretations.

D. Combining the restrictions

Because we have three kinds of restrictions we can vary the parameter structure in three directions. The resulting models can be classified in a three-dimensional figure, as is done below.

Figure 3. Model specifications with number of free parameters and fulfilment of theoretical requirements +/- (small, medium, large project).



In this figure each model is indicated by a box with a code $P_k E_i S_j$, where k=1, 2, 3... gives the degree of the Polynomials, i=R,N indicates whether Equality of shape is Required or Not, and j=N, 0, 1, 2,... gives the degree of Smoothness. In the S.W. corner of the boxes we have given the number of free parameters. The + and - symbols in the S.E. corner of the boxes will be dealt with in section 3.2. Note that $P_{2j+1}E_R S_j = P_{2j+1}E_N S_j$ for all j=0, 1, 2... . Each of these models defines the pattern up to the parameter $\tilde{\beta}$. The basic linear pattern is given by $P_1 E_R S_0 = P_1 E_N S_0$. Note further that k and j must satisfy $k \geq 2j+1$.

There are many more conceivable models than contained in the three dimensional framework. One could e.g. take different degrees for the inflow and outflow polynomials or be selective in setting smoothness restrictions. We have left out these models, assuming that the degree of the polynomial and the degree of smoothness are rough characteristics which should apply equally to inflow and outflow, and equally to the beginning and end of inflow and outflow respectively. At a later stage one may leave these assumptions and go into fine tuning.

3. ESTIMATION AND MODEL SELECTION

3.1 DATA AND ESTIMATION PROCEDURE

We have applied the model described in section 2 to three pilot construction projects of dwellings: a small, a medium-size and a large one. To do so we amended our model in view of the way the variables are measured. The construction period - the period from start until finish of the project - is measured in number of working days n: for each working day i the average number of workers on the site y_i is given. On some working days there is an exogeneously given amount of "nonproductive time" in which the workers on the site cannot continue their work (for instance during bad weather). Therefore the amount of "productive time" on each working day is observed too. If we take this time as a connected interval, the productive time period of workday i is given by $[t_{i-1}, t_i]$ with $t_{i-1} \leq t_i$. The total construction period can now be defined as

$$[t_0, t_n] = \bigcup_{i=1}^{n} [t_{i-1}, t_i]$$

Forecasting in actual time thus requires inversely appropriate spacing of (un)productive time over physical time by using information over weather conditions, holiday periods and the like. Apart from n, y_i and t_i we also have information on the marking points τ_1, τ_2, τ_3 and τ_4. The inflow period $[\tau_1, \tau_3]$ and outflow period $[\tau_2, \tau_4]$ are measured in productive time as well. Note that $\tau_1 = t_0 = 0$ and $\tau_4 = t_n$. Since t_i is known for all i we can derive the average number of construction workers $\mu_i(\theta)$, that theoretically should be present on workday i=1,...,n as

$$\mu_i(\theta) = \mu(t_i^* | \theta) \qquad t_i^* = \tfrac{1}{2}(t_{i-1} + t_i) \tag{3.1}$$

Adding a random disturbance term u_i gives for the actual employment

$$y_i = \mu_i(\theta) + u_i \qquad i = 1,\ldots,n \tag{3.2}$$

We assume the errors u_i follow an AR(1) process:

$$u_i = \rho\, u_{i-1} + v_i \qquad i=1,\ldots,n \tag{3.3}$$

where the v_i are independently and identically distributed random variables with zero mean and variance σ^2. We decompose $\mu_i(\theta)$ into the median state of inflow and the median state of outflow of construction workers at day i

$$\mu_i(\theta) = \vec{\mu}_i(\theta) - \overleftarrow{\mu}_i(\theta) = \vec{\mu}(t_i^* | \theta) - \overleftarrow{\mu}(t_i^* | \theta) \tag{3.4}$$

Applying some substitutions, we can replace model (3.2) by

$$y_i = x_i'\beta + u_i \qquad i=1,\ldots,n \tag{3.5}$$

where x_i is a $3 + 2\bar{k}$ vector whose elements are determined by t_i^*:

$$x_i = (\bar{x}_i, \vec{x}_i', \overleftarrow{x}_i')' = (\bar{x}_i, \vec{x}_{i0}, \ldots, \vec{x}_{i\bar{k}}, \overleftarrow{x}_{i0}, \ldots, \overleftarrow{x}_{i\bar{k}})'$$

$$\bar{x}_i = \left[\tfrac{1}{2} \left(1+\frac{2t_i^*-\tau_3-\tau_1}{\tau_3-\tau_1}\right)\right]_{\text{if } t_i^*\in[\tau_1,\tau_3]} + \left[1\right]_{\text{if } t_i^*\in(\tau_3,\tau_4]}$$

$$- \left[\tfrac{1}{2} \left(1+\frac{2t_i^*-\tau_4-\tau_2}{\tau_4-\tau_2}\right)\right]_{\text{if } t_i^*\in[\tau_2,\tau_4]}$$

$$\vec{x}_{ik} = \left[\left(\frac{2t_i^*-\tau_3-\tau_1}{\tau_3-\tau_1} \right)^k\right]_{\text{if } t_i^*\in[\tau_1,\tau_3]}$$

$$\overleftarrow{x}_{ik} = -\left[\left(\frac{2t_i^*-\tau_4-\tau_2}{\tau_4-\tau_2} \right)^k\right]_{\text{if } t_i^*\in[\tau_2,\tau_4]}$$

In matrix notation we have

$$y = X\beta + u \tag{3.6}$$

If $t_{i-1} < t_i$ a more refined way to compute $\mu_i(\theta)$ would be

$$\mu_i(\theta) = \frac{1}{t_i-t_{i-1}} \int_{t_{i-1}}^{t_i} \mu(t|\theta)dt$$

This in fact we used. The definitions of $\bar{x}_i$, $\vec{x}_{ik}$ and $\overleftarrow{x}_{ik}$ become more complicated, but do not give additional insight. So we have not given them here.

A set of linear restrictions on β as mentioned in section 2.3 can be represented by

$$C\beta = 0 \tag{3.7}$$

where C is a $c \times [3+2\bar{k}]$ matrix of rank c, with c the number of independent constraints. The number of free parameters is thus $\tilde{k}=3+2\bar{k}-c$. If $c>0$, we proceed like Johnston (1984, p.266): we split c into an arbitrary $c\times c$ matrix C_I of rank c and a $c\times\tilde{k}$ matrix C_{II}, so that (3.7) can be replaced by

$$C_I\beta_I + C_{II}\beta_{II} = 0 \quad \text{or} \quad \beta_I = -C_I^{-1}C_{II}\beta_{II} \tag{3.8}$$

which reduces (3.6) to

$$y = [-X_I C_I^{-1} C_{II} + X_{II}]\beta_{II} + u = \tilde{X}\beta_{II} + u \tag{3.9}$$

This implicitely defines the $n \times \tilde{k}$ matrix $\tilde{X}$. Expression (3.9) can also be adopted for the unconstrained case c=0, where $X_{II}=\tilde{X}=X$, $\beta_{II}=\beta$, and X_I and β_I do not exist.

Estimation of (3.9) is done by two step GLS. First, we compute $b_{II}^{OLS}=(\tilde{X}'\tilde{X})^{-1}\tilde{X}'y$, $\hat{u}=y-\tilde{X}b_{II}^{OLS}$, and $r=\Sigma_{i=2}^{n}\hat{u}_i\hat{u}_{i-1}/\Sigma_{i=2}^{n}(\hat{u}_{i-1})^2$, and then we apply OLS on the transformed variables (Prais-Winston):

$$y_1^* = \sqrt{1-r^2}\, y_1 \;, \quad \tilde{x}_1^* = \sqrt{1-r^2}\, \tilde{x}_1 \tag{3.10a}$$

$$y_i^* = y_i - r y_{i-1} \;, \quad \tilde{x}_i^* = \tilde{x}_i - r\, \tilde{x}_{i-1} \qquad i=2,\ldots,n \tag{3.10b}$$

This leads to the two-step GLS estimators:

$$b_{II} = [(\tilde{X}^*)'(\tilde{X}^*)]^{-1}\,(\tilde{X}^*)'y^* \qquad b_I = -C_I^{-1}C_{II}b_{II} \tag{3.11a}$$

$$s^2 = \frac{1}{n-\tilde{k}}\,(y^* - \tilde{X}^* b_{II})'(y^* - \tilde{X}^* b_{II}) \tag{3.11b}$$

$$\hat{V}\begin{Bmatrix} b_I \\ b_{II} \end{Bmatrix} = s^2 \begin{pmatrix} -C_I^{-1} C_{II} \\ I_c \end{pmatrix}\left[(\tilde{X}^*)'(\tilde{X}^*)\right]^{-1}\begin{pmatrix} -C_I^{-1} C_{II} \\ I_c \end{pmatrix}' \tag{3.11c}$$

The following measure $\tilde{R}^2$ may help to judge the descriptive performance of an estimated model

$$\tilde{R}^2 = 1 - \frac{s_u^2}{s_0^2} \quad \text{where } s_u^2 = \frac{s^2}{1-r^2} \text{ and } s_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i)^2 \tag{3.12}$$

It gives the relative variance reduction. Note, however, our models have no constant term.

3.2 ESTIMATION RESULTS AND ADMISSIBLE MODELS

For each of the three projects we can estimate various models and try to select the best one. Our procedure is to set an upperbound of 6 for $\bar{k}$, so that we get a feasible number of models to be estimated and compared. This choice is rather arbitrary, but it seems to guarantee enough flexibility. We now have 36 models. (see Figure 3).

From the estimation of these 36 models we find a conspicious difference with respect to the equality of shape. The models with equality of shape produce rather stable estimates of $\bar{\beta}$ (e.g. for the small project between 88 and 122), whereas the unrestricted ones show extremely high and low values for $\bar{\beta}$ (for the small project from −272475 to 2651). Clearly, in the latter cases we do not have a robust and reliable estimate of $\bar{\beta}$. This is due to multicollinearity, which could be expected from the content of our model. To clarify this we look at the partitioned regression matrix $X=[\bar{x}\ \vec{X}\ \overleftarrow{X}]$, disregarding the equality of shape restrictions. In the limit case where $\tau_1=\tau_2$ and $\tau_3=\tau_4$ (inflow and outflow coincide) we have $\bar{x}=0$ and $\vec{X}=-\overleftarrow{X}$, so that the rank of X becomes $r(X)=r(\vec{X})=1+\bar{k}<3+2\bar{k}$, which means we have perfect multicollinearity. In the other extreme where $\tau_2>\tau_3$ (inflow and outflow completely distinct) we have $\vec{X}'\overleftarrow{X}=0$, thus perfect orthogonality of inflow and outflow variables. If we move from the first to the latter extreme by shrinking the overlap period $[\tau_2,\tau_3]$, the degree of multicollinearity between inflow and outflow is gradually lowered, because $\vec{X}$ and $-\overleftarrow{X}$ become gradually less resembling and more orthogonal, as can be seen from (3.5). In other words the degree of multicollinearity between inflow and outflow depends on the length of the interval $[\tau_2,\tau_3]$ of overlapping inflow and outflow relative to the total construction period $[\tau_1,\tau_4]$, or formally it depends on the value of

$$M(\tau) = \begin{cases} 0 & \text{if } \tau_2 \geq \tau_3 \\ \dfrac{\tau_3-\tau_2}{\tau_4-\tau_1} & \text{if } \tau_2 < \tau_3 \end{cases}$$

which is 1 if $\tau_1=\tau_2$ and $\tau_3=\tau_4$, whereas it is 0 in the other limiting case.

Our data of τ result in a rather high value of $M(\tau)$ viz. 0.88, 0.75 and 0.46 leading to a high degree of multicollinearity. In the models with equality of shape this multicollinearity vanishes due to restrictions. Therefore, to avoid instability of the estimates, we drop the models without equality of shape (partial equality of shape is beyond the scope of this study). The remaining 18 models are suitable only if estimated outflow does not exceed estimated inflow and both are non-decreasing. For each model in Figure 3 we have indicated whether these requirements are satisfied (+) or not (-) in the small, medium and large project respectively. The figure shows for example that for the medium-size project the requirements are violated by the models $P_1E_RS_N$ and $P_2E_RS_N$ or 1RN and 2RN for short. Such models are dropped too, so that we are now left with 10, 16 and 11 admissible models for the small, medium and large project respectively.

3.3 MODEL SELECTION

We may apply a model selection procedure to determine our favourite choice among the set of admissible models. A recent survey of such procedures is given in Maddala (1988,Ch.12). Each procedure compares simpler models with more extensive models, so implicitely there is a trade-off between simplicity and accurate description of the past. Simplicity is connected with greater empirical content of the theory (Popper). One should weigh this against a more accurate description of the data at hand. Different procedures amount to different ways to make this trade-off. Our procedure is to apply model selection for each possible balance between simplicity and data compliance. Thus in selecting a range of models we cover a whole set of procedures.

Often selection methods are judged on their ability to select the "true model" among a set of alternatives. But as a model is always an abstraction of reality, a true model does not exist. The question remains: To what extent do we simplify reality? Because reality cannot dictate the answer, the final choice within the range of selected models is in the end arbitrary or depends on the purpose of the model.

In our approach to model selection we characterize each model by both its parsimony and its data compliance. We quantify the lack of parsimony

as $\tilde{k}$, the number of free parameters, and the lack of data compliance as -ln M, where M is the maximum likelihood. Each possible model thus corresponds to a point in the $\tilde{k}\times(-\ln M)$ plane, so we get a discrete and finite set of admissible points (for examples, see Figure 4).

In order to choose among this set we need a preference ordering over the $\tilde{k}\times(-\ln M)$ plane. This ordering should exclude inefficient models. For example we should not prefer a model, if there is another admissible model with equal parsimony and better data compliance. Although this still allows all kinds of preference orderings, we favour minimizing the Additive Disutility Criterion (ADC):

$$ADC = -2 \ln M + 2\lambda\tilde{k} \tag{3.13}$$

where $\lambda \geq 0$ is a fixed parameter, and may be chosen exogenously to obtain the desired balance between parsimony and data compliance. If we choose $\lambda=1$ we have ADC=AIC (see Akaike, 1974), and if we choose $\lambda=\frac{1}{2}\ln n$ we have ADC=BIC (see Schwartz, 1978), so in fact we specified a generalization of both information criteria by allowing λ to take any nonnegative value. The advantage of our preference ordering is therefore that model selection cannot only be performed for some particular choice of the balance parameter λ, such as AIC or BIC, but for a whole range of λ's.

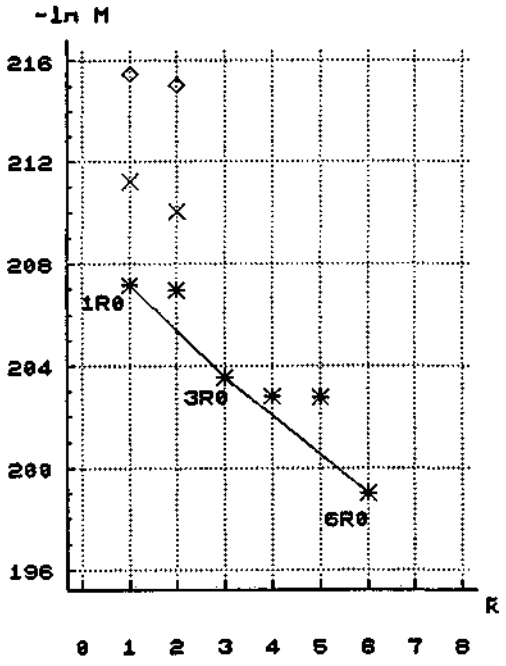
In order to find ADC for a certain model we need to specify the likelihood function L. Assuming normality of v_i in (3.3) we have:

$$\ln L(\beta,\sigma,\rho) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 + \tfrac{1}{2}\ln(1-\rho^2) - \frac{1}{2\sigma^2}(1-\rho^2)(y_1-x_1'\beta) - \frac{1}{2\sigma^2}\sum_{i=2}^{n}[(y_i-x_i'\beta)-\rho(y_{i-1}-x_{i-1}'\beta)]^2 \tag{3.14}$$

Because in our applications the number of workdays n is always large (≥ 96), we can approximate ln M by substituting b, $\tilde{s}=\sqrt{[(n-\hat{k})n^{-1}s^2]}$ and r for β, σ and ρ:

Figure 4. The ADC-efficiency frontier. $\bar{K}$ is the number of free parameters, M is the maximum likelihood. □, ✳, ×, ◇ *indicate admissible models with smoothness of degree N, 0, 1, 2 respectively.*

Small project



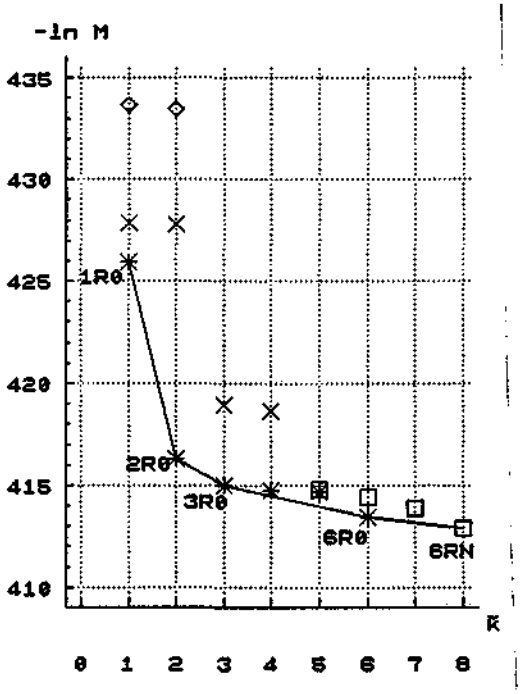
Frontier models:

1R0 for $1.80 \leq \lambda$

3R0 for $1.52 \leq \lambda \leq 1.80$

6R0 for $\lambda \leq 1.52$

Medium project



Frontier models:

1R0 for $9.64 \leq \lambda$

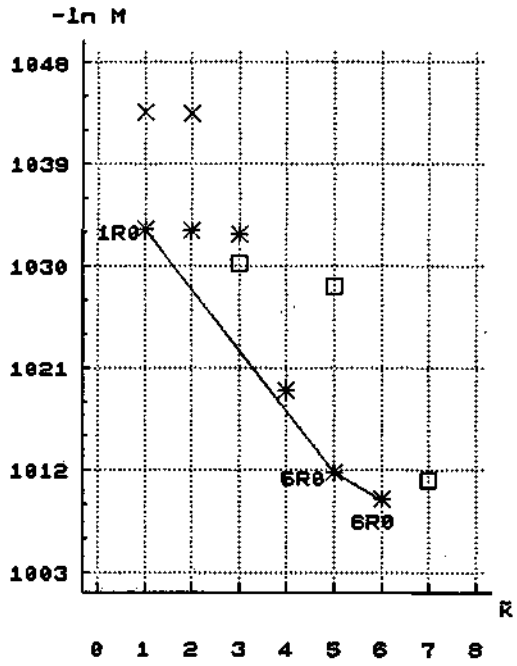
2R0 for $1.26 \leq \lambda \leq 9.64$

3R0 for $0.52 \leq \lambda \leq 1.26$

6R0 for $0.28 \leq \lambda \leq 0.52$

6RN for $\lambda \leq 0.28$

Large project



Frontier models:

1R0 for $5.36 \leq \lambda$

5R0 for $2.43 \leq \lambda \leq 5.36$

6R0 for $\lambda \leq 2.43$

$$\ln M \approx \ln L(b,\tilde{s},r) = -\frac{n}{2} 2\pi - \frac{n}{2}\tilde{s}^2 + \tfrac{1}{2}\ln(1-r^2) - \frac{n}{2} \qquad (3.15)$$

This approximation is asymptotically equivalent and economizes on computer time. Note that r is a consistent and asymptotically efficient estimator of ρ, while conditional on ρ=r the maximum likelihood estimators of β and σ are given by b and $\tilde{s}$. This explains why we use $\tilde{s}$ rather than s. For details see Harvey (1981, Ch.6).

We applied ADC for all values of $\lambda \geq 0$ on our three construction projects. Figure 4 shows the geometry of each optimization procedure in the $\tilde{k}\times(-\ln M)$ plane. Each admissible model corresponds to one point in this plane and each iso-disutility curve corresponds to a straight line with nonpositive slope ($\lambda \geq 0$). It follows that the efficient models lie on a piece-wise linear line, as shown in Figure 4. This line is the efficiency frontier of the set of admissible models. Which of the frontier models is chosen depends on λ. Note that AIC (λ=1) selects model 6RO for the small, 3RO for the medium and 6RO for the large project. Similarly BIC selects respectively 1RO (λ=2.28), 2RO (λ=2.52) and 5RO (λ=2.87).

Comparing the frontiers of the three projects in Figure 4, we conclude they do not coincide. However, the predominance of the models with smoothness of degree 0 is unquestionable. If we take the union of the three frontiers we obtain the complete class of zero smoothness models, and in addition model 6RN. Naturally the all-encompassing model 6RN is selected if λ is close enough to zero, provided that it meets the theoretical requirements. For the medium-size project this occurs if $\lambda \leq 0.278$, which is far below the AIC and BIC values. We conclude that our pilot projects are efficiently described by the models with smoothness of degree 0. The optimal model within this class -in other words the optimal degree of the polynomials- depends upon the particular choice of λ and thus upon the preferred balance between parsimony and data compliance.

As an example we present in the Appendix the estimates of the models, selected for λ=1.5. This is 6RO for the small and large projects and 2RO for the medium project. Also, we plot the observed and estimated employment patterns as well as estimated inflow and outflow.

4. EMPIRICAL SIGNIFICANCE OF THE MODEL

4.1 COMPARING WITH A DIRECT POLYNOMIAL MODEL

On theoretical grounds we have modelled inflow and outflow separately rather than modelling employment directly. It may be questioned whether the increased complexity is compensated by greater empirical significance. To answer this we compare the inflow-outflow model (INOUT) with a model that specifies employment by a direct polynomial (POLY). Using the same symbols as before, POLY assumes the expected employment at time t is given by

$$\mu(t|\theta) = \begin{cases} \sum_{h=0}^{\bar{h}} \gamma_h \left[\frac{2t-\tau_4-\tau_1}{\tau_4-\tau_1}\right]^h & t \in [\tau_1,\tau_4] \\ 0 & \text{otherwise} \end{cases} \qquad (4.1)$$

where $\gamma_h \in \mathbb{R}$ for $h=0,\ldots,\bar{h}$ and $\theta=(\tau_1,\tau_4,\gamma')' = (\tau_1,\tau_4,\gamma_0,\ldots,\gamma_{\bar{h}})'$. This model must satisfy the nonnegativity requirement: $\mu(t|\theta) \geq 0$ for all t. Different models are obtained by varying the degree of smoothness and the degree of the polynomial. The relevant constraints are similar to those in section 2.3. We indicate the POLY model with polynomial degree $h = 0,1,2\ldots$ and smoothness degree $g = N,0,1,2,\ldots$ by the code P_hS_g, so we can construct Figure 5 of POLY model specifications in a similar fashion to Figure 3.

Specification (4.1) is made applicable to the data by adopting expressions (3.1) to (3.3). This leads to a linear regression model in γ with AR(1) disturbances, which can be estimated similar to INOUT.

In the application we choose an upperbound of 7 for $\bar{h}$ to obtain the same maximum number of free parameters as in the inflow-outflow model. Figure 5 shows we have 20 relevant models. For some cases where smoothness is absent the nonnegativity requirement is violated. This brings the number of admissible POLY models down to 17, 20, and 17 for the small, the medium and the large project. These models were estimated and ADC was applied, which generated an efficiency frontier for each project. Figure 6 compares these frontiers (POLY) with those of Figure 4 (INOUT). The short-dash lines, indicated as INOUTe are explained below.

Figure 5. POLY-model specifications with number of free parameters and fulfilment of the nonnegativity requirement +/– (small, medium, large project)

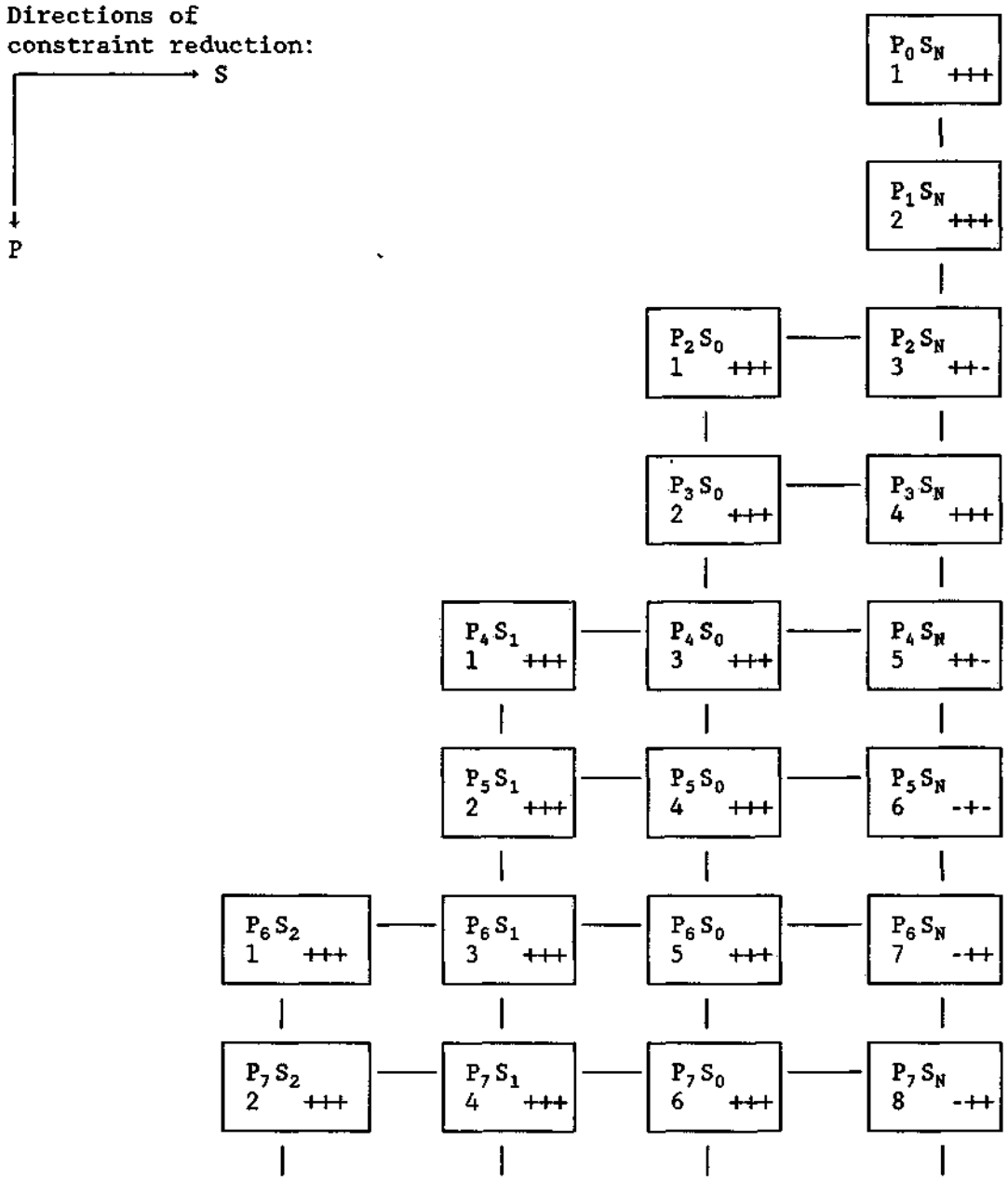
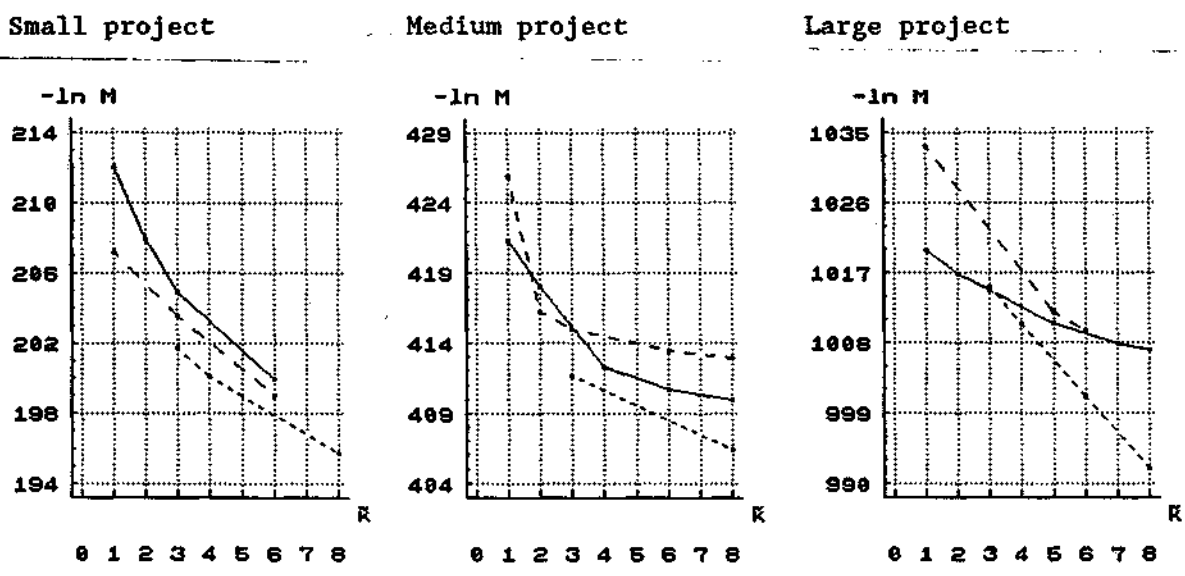


Figure 6. Comparison of ADC-efficiency frontiers. $\bar{k}$ is the number of free parameters, M is the maximum likelihood. The frontiers are: —— direct polynomial models (POLY), --- inflow-outflow models (INOUT), ····· inflow-outflow models with endogenous marking points (INOUTe).



For the small project the INOUT frontier lies completely below the POLY frontier. This indicates that the selected INOUT models always perform better than the selected POLY models, irrespective of the preferred value of λ. For the large project the reverse holds. For the medium project both frontiers intersect, so here the empiricial significance of INOUT depends on the desired value of λ.

4.2 ENDOGENOUS MARKING POINTS

The performance of INOUT relies heavily on the marking points τ_2 and τ_3. These do not measure properly what is required. We measured the ending time of only one operation (τ_2) and the starting time of only one operation (τ_3), but in view of the assumed linear operation times (see

Figure 2) it would have been better to work with averages. We now try to improve our models by using unobservable "structural" marking points.

One way of doing so is to determine the values of τ_2 and τ_3 endogenously. We have done so for models 1RO,...,6RO by varying the values of τ_2 and τ_3 in the direction that increases ln M given in (3.15). This was carried out by a search procedure, starting with the observed values of τ_2 and τ_3. Consequently the number of free parameters of each model is raised by 2. Minimizing ADC generates the INOUTe frontiers in Figure 6. For the small and medium project the INOUTe frontier lies completely below the POLY frontier, which demonstrates the empirical significance of the inflow-outflow setup in both cases. In the large project case a similar conclusion holds as long as we do not assign a heavy weight to parsimony. To make the latter conclusion more precise, we applied ADC on the combined set of all admissible POLY, INOUT and INOUTe models for the large project. The efficient models are POLY model P_2S_0 for $\lambda \geq 4.00$ and INOUTe model 6RO for $\lambda \leq 4.00$. The latter domain includes AIC (λ=1) and BIC (λ=2.87), so model 6RO with endogenous τ_2 and τ_3 is optimal as long as λ is not extremely large.

Unfortunately the endogenously determined values of τ_2 amd τ_3 vary with the degree of the polynomial (see Table 3). For the small and the medium project τ_3 is rather stable, but τ_2 is only stable if we restrict to models 2RO to 5RO. This suggests that model 1RO is too restrictive, and model 6RO is too flexible to get robust marking points. For the large project there is no stability at all. The probable cause for this is the summer-dip in the level of employment in the middle of the construction period, which is easy to observe in Figure A3 of the Appendix (both other projects did not include a summer). Therefore further research should be done to improve the determination of τ_2 and τ_3.

Table 3. Endogenously determined values of τ_2 and τ_3

Model	Small project		Medium project		Large project	
	τ_2	τ_3	τ_2	τ_3	τ_2	τ_3
1RO	24	663	126	811	671	1685
2RO	13	668	97	815	693	1726
3RO	16	672	98	820	605	1819
4RO	14	667	94	816	31	1438
5RO	18	658	104	811	202	1427
6RO	6	667	10	807	259	1430
Observed	24	699	40	892	248	1300

5. SUMMARY AND CONCLUSIONS

In this study we have introduced a model for the employment on a building site as a function of time. A typical characteristic of this model is that it gives employment as the difference between (accumulated) inflow and outflow, which are both polynomial functions of time on the relevant intervals. We have proposed several theoretical restrictions to be imposed on both functions, concerning the degree of the polynomials, equality of shape and smoothness. Restricting ourselves to polynomials up to degree 6 we thus derived a three dimensional framework of 36 possible models. In order to select among these models we described an Additive Disutility Criterion (ADC), which is a generalization of both AIC and BIC by allowing a variable balance λ between parsimony and data compliance. Application of ADC for all possible λ's generates a frontier of efficient models.

The relevant models were studied by using estimations of a small, a medium-size and a large project. We summarize some conclusions that hold for these pilot projects.

First, the models without equality of shape restrictions were found to be inappropriate, because of a high degree of multicollinearity.

Second, among the models with equality of shape those without any smoothness restriction often gave negative estimates of the jumps at the beginning and end of inflow and outflow, thus violating the theoretical requirements.

Third, ADC applied on the set of admissible models (i.e. models with equality of shape that meet the theorectical requirements) yields models with smoothness of degree 0 (simply continuous) provided that λ is not below 0.278, and therefore also if AIC (λ=1) or BIC (λ=2.52) is used. Hence, we can restrict our choice to models with equality of shape that are smooth of degree 0.

Fourth, for the set of models with equality of shape and with smoothness of degree 0, the optimal degree of the polynomials increases for a given project, if a lower value of λ is chosen. It implies that higher degree polynomials must be selected if incorporating significant information (better description) is more important than eliminating redundant information.

Fifth, the empirical significance of the model was investigated by comparing it with a direct polynomial model for employment. The inflow-outflow model performs better than the polynomial model for the small project. For the medium project the inflow-outflow model is only superior for some values of λ, and for the large project the inflow-outflow model performs less well compared to the polynomial model.

Sixth, the assumption that the average time of the various operations in a construction project changes linearly as time proceeds may not be incompatible with the observations, but the observed marking points are not very indicative in discribing this average.

Seventh, an inflow-outflow model that is both superior to the polynomial model and the inflow-outflow model with observed τ_2 and τ_3 can be constructed by endogenizing the marking points. The implied estimates of the marking points are not robust, however, so further research on the proper determination of the marking points is required.

It is useful to expand this study to a larger sample of projects. We intend to do so.

REFERENCES

Akaike, H. (1974): *A new look at the Statistical Model Identification*, IEEE Transactions on Automatic Control AC-19, pp. 716-723.

Harvey, A.C. (1981): *The Econometric Analysis of Time Series*, Philip Alan, Oxford.

Johnston, J. (1984): *Econometric Methods*, McGraw-Hill Book Company, Singapore.

Merkies, A.H.Q.M. and J.A. Bikker (1981): *Aggregation of lag patterns with an application in the construction industry*, European Economic Review 15, pp. 385-405.

Maddala, G.S. (1988): Introduction to Econometrics, Macmillan Publishing Company, New York.

Poirier, D.J. (1976): *The econometrics of stuctural change*, North Holland Publishing Company, Amsterdam.

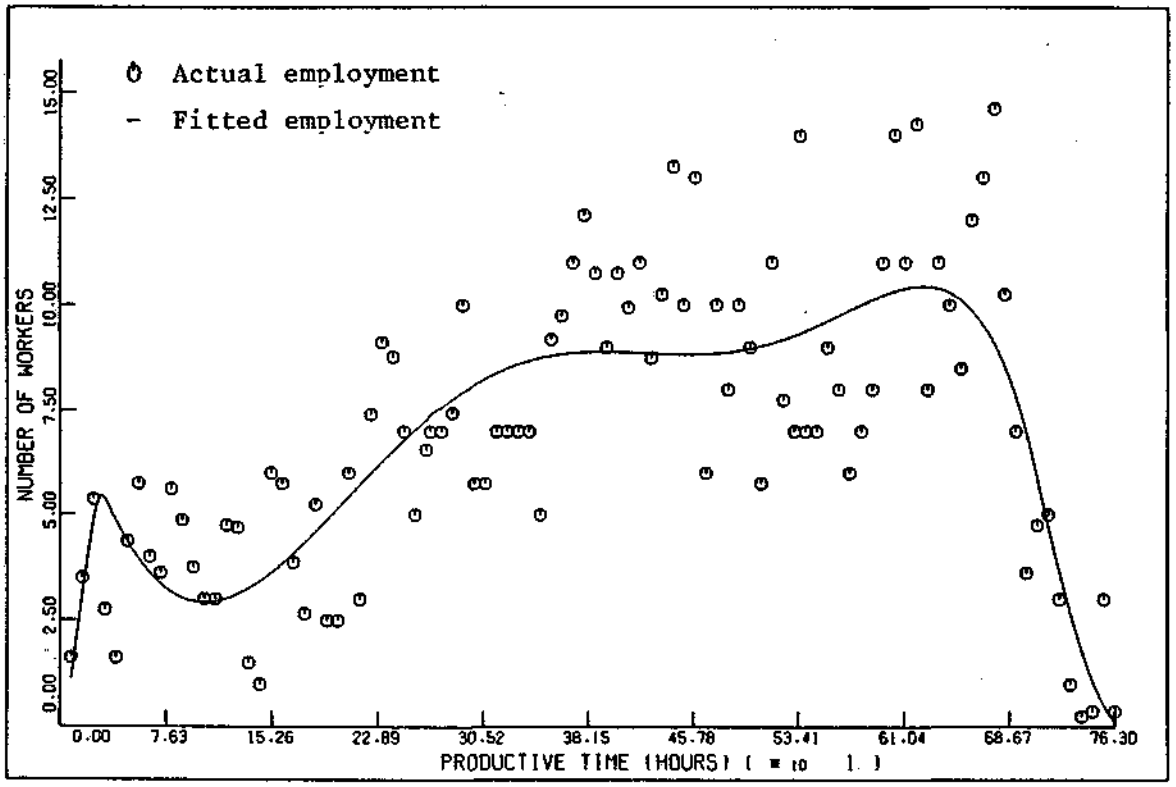
Sawa, T. (1978): *Information criteria for discriminating among alternative regression models*, Econometrica 46, pp. 1273-1291.

Schwartz, G (1978): *Estimating the dimension of a model*, Annals of Statistics 6, pp. 461-464.

Appendix. Estimation results for selected inflow-outflow models (λ=1.5)

	Small Project (4 dwellings) Model $P_6 E_R S_0$	Medium-size project (30 dwellings) Model $P_2 E_R S_0$	Large project (256 dwellings) Model $P_6 E_R S_0$
Marking points in worked days (hours)			
τ_1	0 (0)	0 (0)	0 (0)
τ_2	3 (24)	5 (40)	37 (248)
τ_3	88 (699)	124 (892)	185 (1300)
τ_4	96 (763)	154 (1132)	309 (2272)
Free parameters $\tilde{k}=3+2k-c$	6	2	6
Observations n	96	154	309
Estimations (stand. dev.)			
$\bar{b}$	119.4 (5.7)	140.0 (5.0)	188.0 (3.5)
$\vec{b}_0$	-0.7 = $-\vec{b}_2-\vec{b}_4-\vec{b}_6$	13.0 = $-\vec{b}_2$	4.2 = $-\vec{b}_2-\vec{b}_4-\vec{b}_6$
$\vec{b}_1$	12.9 = $-\vec{b}_3-\vec{b}_5$	-	8.7 = $-\vec{b}_3-\vec{b}_5$
$\vec{b}_2$	-13.5 (10.2)	-13.0 (2.3)	-83.4 (15.3)
$\vec{b}_3$	-23.4 (11.6)	-	-54.7 (15.3)
$\vec{b}_4$	52.5 (21.4)	-	138.8 (40.0)
$\vec{b}_5$	10.5 (8.3)	-	46.0 (11.7)
$\vec{b}_6$	-38.3 (13.3)	-	-59.6 (26.5)
r	0.328	0.470	0.565
s	1.99	3.63	6.40
s_u	2.11	4.11	7.76
$\Sigma y_i/n$	6.91	16.02	51.34
s_0	3.46	6.83	23.40
$\tilde{R}^2$	0.63	0.64	0.89
D.W.	2.032	1.932	2.150

Figure A1. Small project, model $P_6E_RS_0$



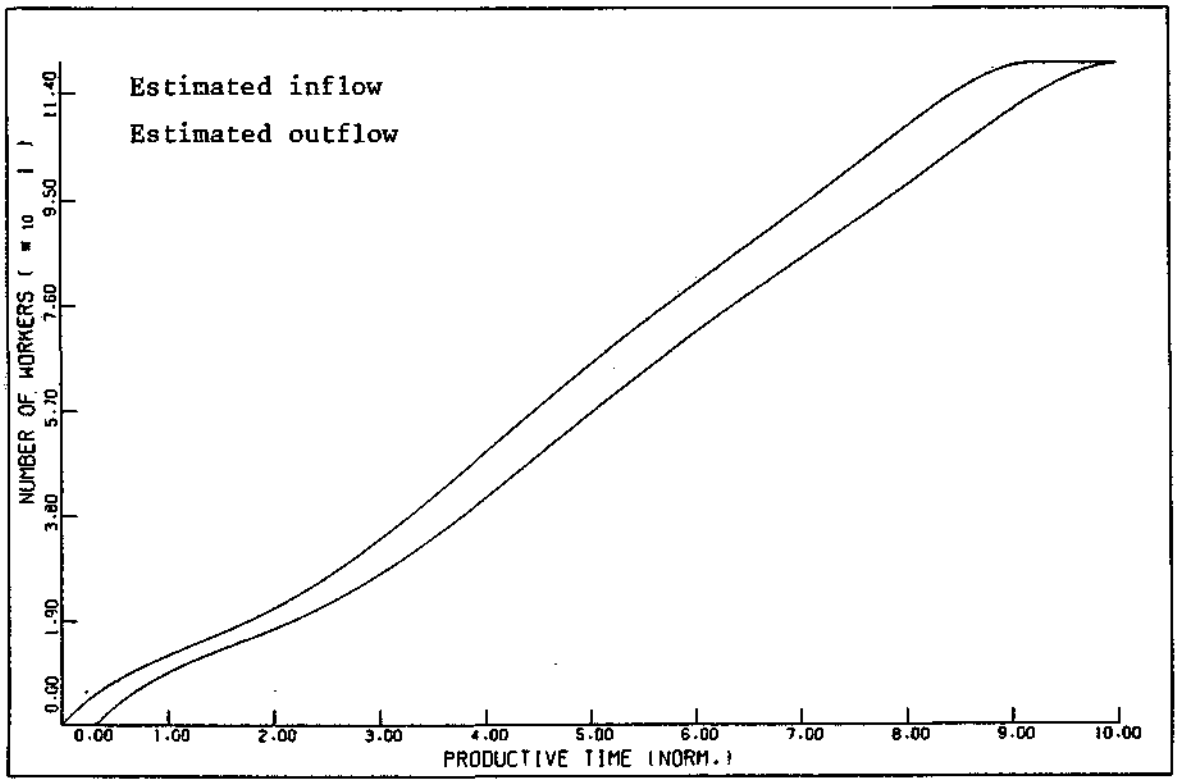


Figure A2. Medium-size project, model $P_2E_RS_0$

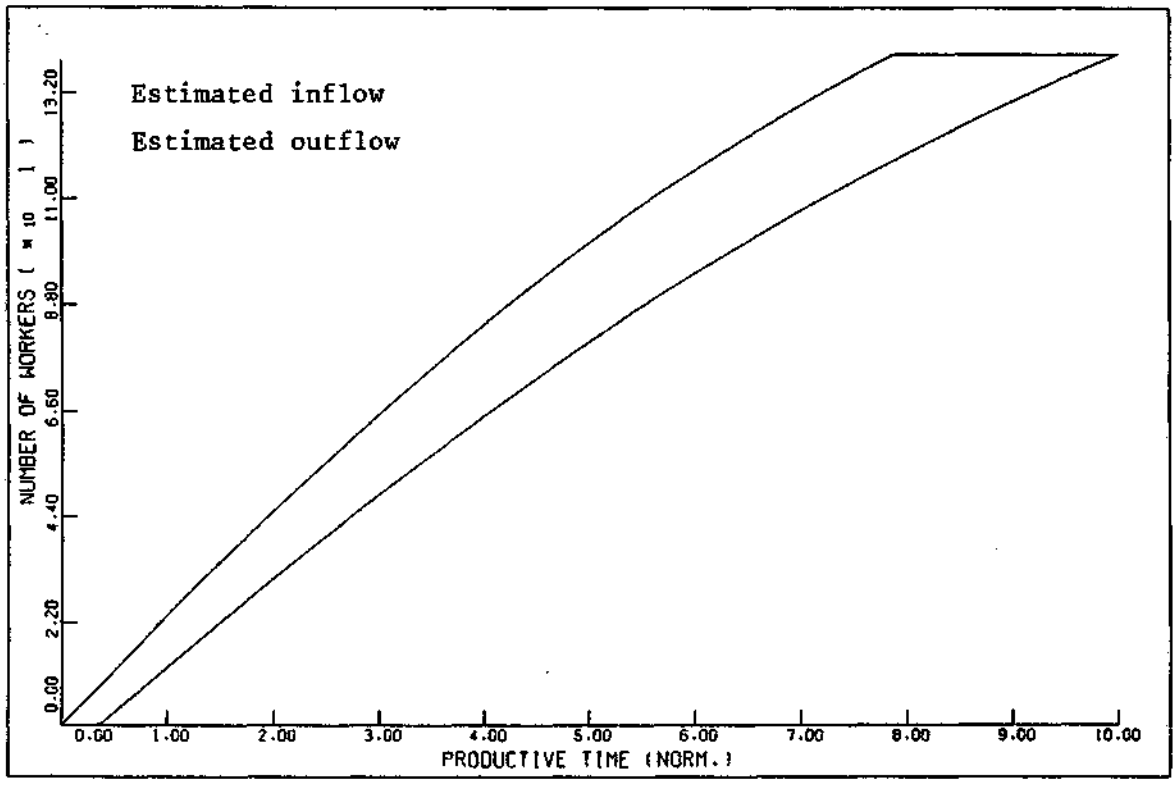


Figure A3. Large project, model $P_6E_RS_0$