Secretaria de Planejamento e Coordenação da Presidência da República
Fundação Instituto Brasileiro de Geografia e Estatística — IBGE

RELATÓRIOS TÉCNICOS

ENCE

ESCOLA
NACIONAL
DE CIÊNCIAS
ESTATÍSTICAS

FORECASTING THE NUMBER OF AIDS CASES IN BRAZIL

Dani Gamerman
Universidade Federal do Rio Janeiro
and
Hélio S. Migon
Escola Nacional de Ciências Estatísticas
Universidade Federal do Rio Janeiro

As matérias publicadas nos **RELATÓRIOS TÉCNICOS** são preprints, com tiragens limitadas, de trabalhos elaborados por professores da **ENCE**, em complementação a suas atividades de ensino, com ênfase para as pesquisas realizadas no Laboratório de Estatística da Escola.

FORECASTING THE NUMBER OF AIDS CASES IN BRAZIL

Dani Gamerman
Universidade Federal do Rio Janeiro

and

Hélio S. Migon
Escola Nacional de Ciências Estatísticas
e Universidade Federal do Rio Janeiro

Summary

This paper presents a method to describe and forecast the incidence of AIDS in Brazil using time series models. The method is based on the class of generalized exponential growth models and uses the ideas of non-linear dynamic modelling. The aim is to provide good predictions and to inform sequentially on the asymptotic or explosive behaviour of the data series. Intervention to model unexpected changes in the data, on-line variance estimation and variance dependence on the mean are used to adequately model the data. The data is analysed with some particular models from the above class and the resulting inferences compared in terms of short-term and long-term predictive perfomance and model fit.

Keywords: Predictive performance, Generalized exponential growth models; Dynamic models; Intervention.

1. Introduction

The epidemics of AIDS is increasingly being recognised as a disease of prime importance and a great deal of research effort is now focused in primarily understanding and hopefully controlling and erradicating it. In the past few years, statisticians have begun to contribute to this research and a recent issue of the Journal of the Royal Statistical Society, Series A is just an example of this increasing trend.

Among the main areas of research are the studies of the number of infected individuals in a given population, the time for the development of the disease and the number of individuals that have developed the disease. Malice and Kryscio (1988) tackled the problem by using an epidemic model to describe the number of infected and diagnosed individuals at any time and relate these via an infection rate which essentially describes the incubation period for AIDS. De Gruttola and Lagakos (1987) describe the distribution of the number of infected individuals via a convolution of the distribution of the incubation time and the number of AIDS cases. Our task is simpler because we are only dealing with the number of notified cases. The approach differs substantially, however, through the use of time series models. This seems the natural approach since the data is time indexed and proves to be adequate in the sequel by allowing many useful facilities to be implemented.

The data series considered is the number of monthly notifications of AIDS cases in Brazil from September/85 to January/88 plotted in Figure 1 and given in the Appendix. It shows a stable pattern, basically linear up to January/87 when its pattern drastically changes. The overall trend is very irregular and can only be adequately described with a non-linear model. This feature is probably due to the nature of the disease, still in its onset, and the implementation of policies to control it, but can also be related to undereporting and the aggregation of data from a country as large and uneven as Brazil. This is also probably the main cause of decrease in July/87. These points are briefly discussed in the final Section.

< Place Figure 1 about here >

The importance of the study of AIDS in Brazil cannot be overstressed since it is becoming the second largest country in number of cases. When one considers that the odds in favour of discovery of cure within the next years are slim and the main weapon against AIDS is control via massive educational campaigns, one realises the problems facing

Brazil and, to some extent, other underdeveloped countries.

From a statistical point of view, it is important to have an appropriate model description for two reasons. Firstly, because it allows one to produce reliable predictions to the future and decisions at government and health levels can be taken accordingly. Secondly, because construction of adequate models allows identification and monitoring of features of interest in the series. In the specific case the AIDS data series, the models used are able to identify the acceleration of spread of the desease via a single parameter. If this is smaller than one, the process is expected to eventually reach an asymptote that one hopes, to be as low as possible. Otherwise, the nature of the process is explosive and no such limit is attainable. This explosive behaviour makes constant monitoring of this specific parameter of vital importance. It should indicate whether the process is out of control and additional measures are required. Signals would then be triggered to government policy bodies and international organizations like WHO that would act accordingly.

It is shown that far from being constant, this parameter wanders around the value of 1, sometimes above it, reflecting the changing pattern of the process at hand. Dynamic models are recommended in such cases because they naturally allow for this variation. They also prove useful to model intenvention when more drastic variation takes place as is the case when $t = February/87$.

In the next Section, the models used are presented. The commonly used logistic curves are considered within a larger class of potentially useful models called generalized exponential growth models (GEGM). This class is obtained after a Box-Cox transformation on the mean trajectory of the series following a logistic curve. A simpler model representation giving essentially the same results is obtained via a reparametrisation and used in the sequel. This essentially reduces the non-linearity inherent to logistic curves and GEGM's. Section 3 summarises the main elements used in dynamic Bayesian inferences and discusses methods of comparing the performance of diferent models. These results are applied in Section 4 to analyse the AIDS data. A number of models are used and many features of dynamic modelling are described. In particular, it is shown that a correctly signalled intervention is vital for fast adaptation to changes. The problem of discrimination among possible alternatives within the class of GEGM's is also discussed and the tentative conclusion points at long-term forescating as the ultimate test. Section 5 draws on some final

comments with respect to the class of GEGM's and other problems related to AIDS data.

2. Generalized Exponential Growth Models

2.1. Definition

One of the key elements of a time-series model is its mean response function defined as

$$\mu_t = E[y_t|\underline{\beta}]$$

where y_t is the observation made at time t and $\underline{\beta}$ is the vector of parameters used to model the series. Standing at any time t, the future trajectory of the series can be obtained by looking at μ_{t+k} as a function of k and used to define a model. A commonly used model for the mean response function in epidemics (Bailey, 1975; Duong & MacNeill,1987) is the logistic curve given by

$$\mu_t^{-1} = \beta_1 + \beta_2 \beta_3^t \tag{1}$$

β_3 is the key parameter here. If $\beta_3 < 1, \mu_t$ will eventually converge to the asymptote β_1^{-1} and otherwise μ_t increases without limits provided $\beta_2 < 0$. The future trajectory of the series at time t is given by the inverse of

$$\beta_1 + \beta_2 \beta_3^{t+k} = \beta_1 + \beta_2^{\star} \beta_3^k \quad \text{for} \quad \beta_2^{\star} = \beta_2 \beta_3^t$$

and follows the same pattern of the mean response.

The logistic curve can be embedded in a more general family of curves by allowing transformations on the mean other than the inverse. A suitable class is generated by allowing simple polinomial transformations (Box & Cox, 1964) on the mean as

$$g(\mu) = \begin{cases} \mu^{\lambda}, & \text{if } \lambda \neq 0 \\ \log \mu, & \text{if } \lambda = 0 \end{cases} \tag{2}$$

The transformation g is also called a link because it relates the mean to the parameters used to describe it. The logistic curve is obtained for $\lambda = -1$. The Gompertz curve given by

$$\mu_t = exp(\beta_1 + \beta_2 \beta_3^t)$$

is qualitatively similar to the logistic curve and is obtained for $\lambda = 0$. It is important to make it clear that this transformation is performed on the mean rather than on the

observation. These are kept at their original measurements to retain interpretability. Heteroscedasticity problems implied by this approach are dealt with in Section 3.

Exponential curves are given, of course, by $\lambda = 1$. Since this represents the simplest form of link, the identity, the class is named generalized exponential growth models (GEGM). Throughout GEGM's, the influence of β_3 in the qualitative behaviour of the series is the one observed with the logistic curves depending basically on whether it lies below or above 1. Essentially, the transformation parameter λ is unknown and has to be estimated. This has proved to be a difficult task (see Mar-Molinero, 1976) and we do not attempt it here. We rather consider the above three main forms and compare them in many diferent ways.

Consider now a vector sequence $\theta_t = (\theta_{1,t}, \theta_{2,t}, \theta_{3,t})$ recursively defined via

$$\theta_{1,t} = \theta_{1,t-1} + \theta_{2,t-2} \tag{3.a}$$

$$\theta_{2,t} = \theta_{3,t-1}\theta_{2,t-1} \tag{3.b}$$

$$\theta_{3,t} = \theta_{3,t-1} \tag{3.c}$$

and related to the mean response function as

$$g(\mu_t) = \theta_{1,t} \tag{4}$$

Prior to (4), θ_1 represents the current level of the series, θ_2 is the current rate of change and θ_3 is a dampening or accelerating factor of this change depending on whether it lies below or above 1 respectively. Repeated use of (3.a-c) gives

$$\theta_{1,t+k} = \beta_1 + \beta_2\beta_3^k$$

for $\beta_1 = \theta_{1,t} + \theta_{2,t}/(1 - \theta_{3,t})$, $\beta_2 = -\theta_{2,t}/(1 - \theta_{3,t})$ and $\beta_3 = \theta_{3,t}$ as shown in Migon & Gamerman (1988). For the purposes of this paper, it suffices to recognise that one can work with form (1)-(2) or (3)-(4) since they describe the same trajectories. The latter is chosen here because it reduces the non-linearity of the model making it more stable in estimation. An alternative description is given by Meade (1985).

A simplification is obtained by taking $\theta_{3,t} = 1, \forall t$. In this case, is straighforword to obtain future trajectory at t as

$$\theta_{1,t+k} = \theta_{1,t} + k\,\theta_{2,t}$$

With an identity link ($\lambda = 1$), this specialisation gives the linear growth models (Harrison & Stevens, 1976). Although they are simpler and retain linearity, they are shown to be inadequate in modelling epidemics data like AIDS in section 4.

2.2. Dynamic modelling

As antecipated in the previous Section and suggested by Duong and MacNeill (1987), structural changes on the observed AIDS series are to be expected and have to be accommodated. These changes are generally adequately modelled with small disturbance terms that are added to the structure of the series. This can be implemented in model (3)-(4) by adding a three dimensional disturbance term $\underline{w}_t$ to (3). This can be concisely written as

$$\underline{\theta}_t = G\left(\underline{\theta}_{t-1}\right) + \underline{w}_t, \quad \underline{w}_t \sim [\underline{0}, W_t] \tag{5}$$

The above notation specifies that $\underline{w}_t$ have $\underline{0}$ mean and covariance matrix W_t increasing the uncertainty about $\underline{\theta}$ and G is the function describing the deterministic part of the time evolution of $\underline{\theta}$ as given by (3). Under this formulation, the dampening factor θ_3 is allowed to change in time thus making sense of (3.c).

Occasionally, major structural changes that cannot be catered for by small disturbances take place. In the absence of information on the nature of these changes, a parsimonious approach is to inflate the covariance matrix W_t increasing the uncertainty in the process and allowing it to be more adaptive to fresh information and less dependent on the past. This intervention proves very useful for the AIDS data as we show in Section 4.

3. Inference Procedures

The first step is to model the observational distribution of the series. It is assumed here that Y_t follows a normal distribution with mean μ_t. We have also assumed that $V[y_t|\mu_t, \phi_t] = \phi_t^{-1}\mu_t$ to accommodate for the large range of values assumed by the series. When $\phi_t = 1$, this is the variance law given by the Poisson model and comparisons with it are made in Migon & Gamerman (1988). Generally, ϕ_t is unknown and is estimated by the model. It is expected to be stable and to change only slightly with time.

The inferential procedure is sequential processing one observation at a time as follows:
(i) - at time $t-1$

$$\begin{aligned} \underline{\theta}_{t-1}|\phi_{t-1}, D_{t-1} &\sim [\underline{m}_{t-1}, \phi_{t-1}^{-1} C_{t-1}] \\ \phi_{t-1}|D_{t-1} &\sim G(n_{t-1}/2, S_{t-1}/2) \end{aligned}$$

where D_{t-1} is the information set at time $t-1$ containing all the information available at that time including the observed values $y_1, \cdots, y_{t-1}$ and G(a, b) is the Gama distribution with density proportional to

$$x^{a-1}e^{-bx}$$

(ii) - Evolution to time t can be performed through (5) to give the distributions for $\underline{\theta}_t|\phi_t, D_{t-1}$. The time evolution for ϕ is modelled via

$$\phi_t|D_{t-1} \quad \sim \quad G(\delta_v n_{t-1}/2, \delta_v S_{t-1}/2)$$

The factor δ_v is generally taken as slightly less than 1 increasing the variance of ϕ to reflect more uncertainty above its value. In this paper, δ_v is taken as 0.98. When $\delta_v = 1$, ϕ_t retains the distribution of ϕ_{t-1} and ϕ is essentially independenty of t.

(iii) - One-step ahead predictions are given by the implied t distribution for $y_t|D_{t-1}$. For more steps ahead, repeated use of **(ii)** and **(iii)** sequentially gives the distributions for

$$y_{t+k}|D_{t-1}, \quad k = 1, 2, \cdots$$

(iv) - the information obtained after observing y_t is used to update the distributions of $\underline{\theta}_t|\phi_t, D_t$ and $\phi_t|D_t$.

Full details of this cycle are given in Migon & Gamerman (1988) based on the inference for dynamic generalized linear model (West, Harrison and Migon, 1985). For non-linear models, the procedure follows Migon (1984).

A simple method to define the matrices W_t is given by the discount approach (Ameen & Harrison, 1984). The variance of each parameter is inflated to reflect the amount of information lost. If $100\delta\%$ of the information about θ_i passes through time, set $V[\theta_{i,t}|D_{t-1}] = V[\theta_{i,t-1}|D_{t-1}]/\delta_i, \quad i = 1, 2, 3$. The implied value of W_t is a diagonal matrix with entries $V[\theta_{i,t-1}|D_{t-1}](\delta_i^{-1} - 1), \quad i = 1, 2, 3$. The static model is attained at $\delta_1 = \delta_2 = \delta_3 = 1$. The discount concept is specially helpful in specifying intervention. As previously discussed, one wishes to substantially reduce the amount of information passing at these points. This can be achieved by considerably lowering the values of δ_i.

Inference for the AIDS series is primarily devoted to estimation of θ_3 and prediction of future observations. The estimation of θ_3 can be based on the on-line distributions $[\theta_{3,t}|D_t]$. Time trajectories for θ_3 can be drawn taking the mean of these distributions

as estimates. Particular attention is given to the relative position of these trajectories with respect to the value of 1. Similarly, time trajectories for ϕ^{-1} based on the inverse of the mean of $[\phi_t|D_t]$, $t = 1, 2, \cdots$ provide an indication of the dispersion of the data, values above 1 indicating overdispersion with respect to the Poisson law of variance. They provide a rough guide to the way the model copes with the observational uncertainty in the series.

The discussion of prediction is inherent to time series and a model can only be accepted if it leads to good predictions. Although of little dispute, this point has sometimes been overlooked by forecasters whose main optimality criteria is to obtain the best fit. This is also important but not essential. In the present formulation, the fit of the model is given by the sum of the squares of $y_t - \hat{y}_t$ where $\hat{y}_t$ is an estimate of μ_t taking the whole data (including y_t) into account. Predictive optimality is attained, for example, by minimization of a suitable function of the forecast errors given by

$$e_t = y_t - E[y_t|D_{t-1}]$$

since $y_t = \mu_t + e_t$ and $E[e_t] = 0$, this is $y_t - E[\mu_t|D_{t-1}]$ which does not use y_t or any subsequent observation to estimate μ_t.

4. Application

The data consists of 29 monthly notifications of AIDS cases as provided by the official Health Ministry agency in Brazil. It ranges from t = 1 (September/85) to t = 29 (January/88). The approach adopted here is to try out as many models as possible within the class of GEGM's and compare them. Unfortunately, there is no such thing as the ultimate test to discriminate between models. We are however essentially concerned with the ability of the model to predict well rather than fit the data well. See Migon & Gamerman (1988) for a discussion. The measures of predictive performance can be divided in short-term and long-term performances as a model can perform poorly in the long term but can be very effective in the near future. The main measures of short-term performance are the predictive likelihood (West and Harrison, 1986) and the sums of the absolute and squared one-step ahead forecast errors. In this paper, we concentrate on the last one taking the sum only from $t = 4$ to allow the model to learn about the three-dimensional parameter specifying the mean.

With all the models, analysis starts with prior $[\underline{\theta}_1|\phi_1, D_0] \sim [\underline{a}_1, \phi_1^{-1}R_1]$ where $\underline{a}_1 = (a_{11}, a_{12}, a_{13})^T$, $R_1 = diag(r_{11}, r_{12}, r_{13})$ and $\phi_1|D_0 \sim G(.05, .05)$ implying that:

i - $E[\phi_1|D_0] = 1$ and $V[\phi_1|D_0] = 20$ indicating large uncertainty about the initial guess $\phi_1 = 1$;

ii - The interval with endpoints $a_{1i} \pm 2\sqrt{r_{1i}}$ should contain θ_i with probability roughly .95; the values of $\underline{a}_1$ and r_{11}, r_{12} and r_{13} corresponding to the exponential ($\lambda = 0$), Gompertz ($\lambda = 1$) and logistic ($\lambda = -1$) models are given in Table 1 below

Table 1

Prior moments specification

Model	$\underline{a}_1$	r_{11}, r_{12} and r_{13}
$\lambda = -1$	$(.002, -.0001, .95)$	$.05, .05, .05$
$\lambda = 0$	$(6, 4, .95)$	$1, 1, .04$
$\lambda = 1$	$(400, 50, 1)$	$400, 100, .1$

It can be seen that the mean is positioned roughly where the data is (observe that $(.002)^{-1} = 500$ and $e^6 = 403.4$) but with a large variance representing very little knowledge of the process a priori. The mean of the θ_3 is positioned near 1 but slightly below for $\lambda = \pm 1$ but exactly at 1 for $\lambda = 0$. The reason being that exponential curves do not exhibit S-shaped forms and therefore the only curves consistent with the data have $\theta_3 > 1$. Small changes in either values of $\underline{a}_1$ or R_1 will not cause any significant difference due to the large initial uncertainty.

The evolution is set using the discount approach. This essentially subjective specification is done considering θ_3 to be more durable than θ_1 and θ_2 and therefore, its information is less discounted. Among a range of possible values typically above .9, it was chosen that $\delta_1 = \delta_2 = .9$ and $\delta_3 = .95$ for all three models except when $\lambda = -1$ in which case, δ_3 is set to .98. As previously stated, the variance discount δ_v is set to .98. A major feature of the series is a change of pattern at $t = 18$ (February/87). Any method not adjusting for this change is doomed to failure as we show below. Some form of intervention in the model is called for. As discussed in Section 3, this can be economically achieved by lowering the discount values. In this case, it was considered that the change was basically due to drastic change in the level and rate of change of the level of the series and not the acceleration of

that change. This was implemented by setting $\delta_1 = \delta_2 = 0.25$ when t = 18, thus reducing by a factor of 4 the information on θ_1 and θ_2 and preparing for change.

We started the analysis by using the simplest model on the class, namely the linear growth model (LGM). Table 2 gives the sum of squared errors of this and other models. The reduction in error due to intervention is evident. If one allows θ_3 to be estimated by the data, one gets further reduction in forecast errors no matter what transformation on the mean is used. Also, there is a substantial reduction by intervening in the series. Also listed in Table 2 are the SSE for the static models, i.e., those assuming $\underline{\theta}$ specified by (3).

Table 2

Sum of squared one-step ahead forecast errors (SSE)

Form of growth	Discounts	Intervention	SSE	Fit
LGM	.9, .9	No	612, 112	206, 293
	.9, .9	Yes	377, 131	94, 317
Exponential	1, 1, 1	No	989, 782	526, 853
	.9, .9, .95	No	367, 016	46, 357
	.9, .9, .95	Yes	256, 084	30, 779
Gompertz	1, 1, 1	No	723, 227	410, 920
	.9, .9, .95	No	375, 958	42, 533
	.9, .9, .95	Yes	268, 771	31, 441
Logistic	1, 1, 1	No	773, 161	316, 081
	.9, .9, .95	No	382, 015	21, 971
	.9, .9, .95	Yes	282, 748	19, 020

The improvement of the dynamic model over the static version is evident confirming the suggestion of frequent changes in the structure of the model. These are small specially when compared to the change taking place at t = 18. The final column in Table 2 gives the sum of the squared fitted errors. These are given by the difference between the observation and the level of the series as estimated using the whole data. It is of secondary importance here because it can only be obtained retrospectively and therefore, is of little use for

forecasting. It is striking, however, to see how small they are when compared to their respective SSE showing how misleading fitted figures can be in forecasting.

We consider in detail now the best model within each form of growth, namely, that with discounting and intervention. It is interesting to assess first the importance of intervention. It reduces the SSE by up to 34% and the main reason for it is given in Figure 2. It shows the trajectories of θ_3 with intervention (continuous line) and without intervention (dashed line). Although δ_3 has not changed, the intervention in θ_1 and θ_2 is enough to allow great changes in θ_3. If intenvention is not performed, however, the model, interprets the event as indicative of greater change in θ_3 than is really taking place increasing it far too much. The most drastic effect is observed for the Gompertz growth in Figure 2.b where the lack of intervention sustains the value of θ_3 above 1 whereas intervention brings it well below 1 after 6 months. A similar but less marked effect can be observed in Figure 2.c. Also, the trajectories of ϕ^{-1} exhibit similar patterns but the variance without intervention almost doubles that obtained with intervention as shown in Figure 3 for the identity link. This doubling occurs because the model without intervention treats the change almost entirely as extrarandomness in the observation process.

< Place Figure 2 about here >

< Place Figure 3 about here >

The forecasts obtained with the three models are shown in Figure 4. The dashed lines are the data series and the dotted lines are the 2 s.d. limits of the t-student forecast distributions of $y_t|D_{t-1}$. They show fast adaptation to the change at t = 18 and the only other weak point is at t = 23 (July/87), a possible outlier. They show great resemblance particularly between the exponential and Gompertz models. Their pattern can be examined more closely by looking at the forecast errors in Figure 5, the dotted lines being their 2 s.d. limits. The interval formed by these limits should include 0 if the model is adequately predicting into the future. The exponential and Gompertz models only fail to do so at t = 23 and the logistic model also fails at t = 20. These residuals however show no particular structure as autoregression and are considered adequate.

< Place Figure 4 about here >

< Place Figure 5 about here >

Figure 6 shows the trajectories of θ_3 for the above models along with respective 2 s.d. limits. The exponential model retains a trajectory above 1 most of the time although with enough uncertainty to make it inconclusive as to whether it is an explosive behaviour. The same can be said about the other models where three distinct periods can be identified: a stable one up to the end of 1986 with an asymptotic behaviour, the next semester where there is some indication of an explosive behavour. Following this period, the value of θ_3 drops substantially indicating once again an asymptotic behaviour but steadily increases back towards 1. These graphs serve as a controlling device signalling when the process is out of control or leading towards it as indicated at the end of the series. These could be useful indications to government policy bodies and international institutions like WHO that would act accordingly.

< Place Figure 6 about here >

During the process of analysing this series, four data points from t = 30 (February/1988) to t = 33 (May/1988) became available to us. They were then used in a medium-term forecast exercise. Namely, the forecasts for y_{29+k}, $k = 1, 2, 3, 4$ based on D_{29} where derived and compared to the actual values of the series. These are shown in Table 3 and plotted as small squares in Figure 4 along with their 2 s.d. limits. Apart from the value for February/88, another possible outlier, the forecasts based on the logistic model seem very good. They are better then those of the other models that do not increase that fast.

We do not attempt at this stage to provide any long-term forecasts. As it is clear from the above discussion, the process is changing substantially in different periods of time and there is no indication that this is not going to hold in the near future. Forecasts long into the future are to be avoided since they are likely to be based on information that is no longer correct. Also, the 2 s.d. limits become so far apart that a likely interval for the forecasts will be hopelessly large.

Table 3

Medium-term forecasts with the intervention models

Month	Data	Logistic	Gompertz	Exponential
Feb/88	2537	2762	2734	2745
Mar/88	2956	2925	2872	2890
Apr/88	3100	3105	3018	3043
May/88	3378	3305	3171	3205

5. Discussion

This paper presents an approach to model the evolution of AIDS in Brazil. Using the class of GEGM's, it was observed that the links that are more commonly used ($\lambda = \pm 1, 0$) give comparable results. This also occurred with Duong & Mac Neill (1987) when analysing AIDS data from Canada. One reason for it is the difficulty in discriminating between these models mentioned above. With AIDS data, there is an extra complication caused by the fact the we are still observing the initial part of the evolution. The identification of the particular model better suited to the data involves consideration of a considerable part of the evolution which requires extrapolation. We have aheady cautioned against extrapolating too for ahead and the data seem to be supporting it.

An improvement of the model could be obtained by use of explanatory variables. We have only used time series models in order to establish the pattern of evolution followed by the data but variables judged to be related to the number of cases could be considered in the model. There are many different ways that this could be done since they can affect only some of the parameters of the model or directly through the mean function.

A related point is data disaggregation. One possibility is to analyse the evolution of AIDS by each of a number of groups instead of using the grand monthly total. Data could be divided geographically by state or region, by risk group, by sex, age and so forth. This could perhaps lead to a more established pattern in some of the groups. Another data problem is underreporting. The number of AIDS cases not reported in Brazil is considered to be substantial and is a worrying feature from the control point of view. Theoretically, it could be incorporated into the analysis by modelling the mechanism

of underreporting. This is generally unknown and its knowledge could help solving the problem of underreporting itself.

Acknowledgement

The research of the first author was supported by CNPq-Brazil.

Appendix

Number of notified cases of AIDS in Brazil

Month	No. of cases	Month	No. of cases	Month	No. of cases
Sep/85	432	Sep/86	841	Sep/87	2102
Oct/85	483	Oct/86	875	Oct/87	2237
Nov/85	520	Nov/86	921	Nov/87	2325
Dec/86	540	Dec/86	982	Dec/87	2458
Jan/86	574	Jan/87	1012	Jan/88	2651
Feb/86	625	Feb/87	1263		
Mar/86	657	Mar/87	1542		
Apr/86	673	Apr/87	1696		
May/86	725	May/87	1835		
Jun/86	739	Jun/87	1981		
Jul/86	790	Jul/87	1906		
Ago/86	829	Ago/87	2013		

References

Ameen, J. R. M and P. J. Harrison, 1985, Normal discount Bayesian models, in: J. M. Bernardo et al., eds., Bayesian Statistics (University Press, Valencia).

Bailey, N. T. J., 1975, The mathematical theory of infectious diseases, 2nd. ed. (Hafner Press, New York).

Box, G. E. P., and D. R. Cox, 1964, An analysis of transformations, Journal of the Royal Statistical Society, Series B, 26, 211-252.

De Gruttola, V. and S. W. Lagakos, 1987, The value of AIDS incidence data in assessing the spread of HIV infection, technical report, Department of Biostatistics, Harvard School of Public Health.

Duong, Q. P. and J. B. Mac Neill, 1987, Selection and estimation of growth models with application to forecasting AIDS, technical report TR-87-09, Department of Statistical and Actuarial Sciences, University of Western Ontario.

Harrison, P. J. and C. F. Stevens, 1976, Bayesian forecasting. Journal of the Royal Statistical Society, Series B, 38, 205-247.

Malice, M. P. and R. J. Kryscio, 1987, A stochastic model for the incidence of AIDS, paper presented at the XIV International Biometric Conference, Namur, Belgium.

Mar-Molinero, C., 1980, Tractors in Spain: a logistic analysis, Journal of the Operational Research Society, 31, 141-152.

Meade, N., 1985, Forecasting using growth curves - an adaptive approach, Journal of the Operational Research Society, 36, 1103-1115.

Migon, H.S., 1984, An approach to non-linear Bayesian forecasting problems with applications, unpublished Ph. D. thesis, University of Warwick.

Migon, H.S. and Gamerman, D. (1988). Generalized exponential growth models: a Bayesian approach, technical report no. 41, Laboratório de Estatística, Federal University of Rio de Janeiro.

West, M. and P. J. Harrison, 1986, Monitoring and adaptation in Bayesian forecasting models, Journal of the American Statistical Association, 81, 741-750.

West, M., P. J. Harrison and H. S. Migon, 1985, Dynamic generalized linear models and Bayesian forecasting, Journal of the American Statistical Association, 80, 73-97.

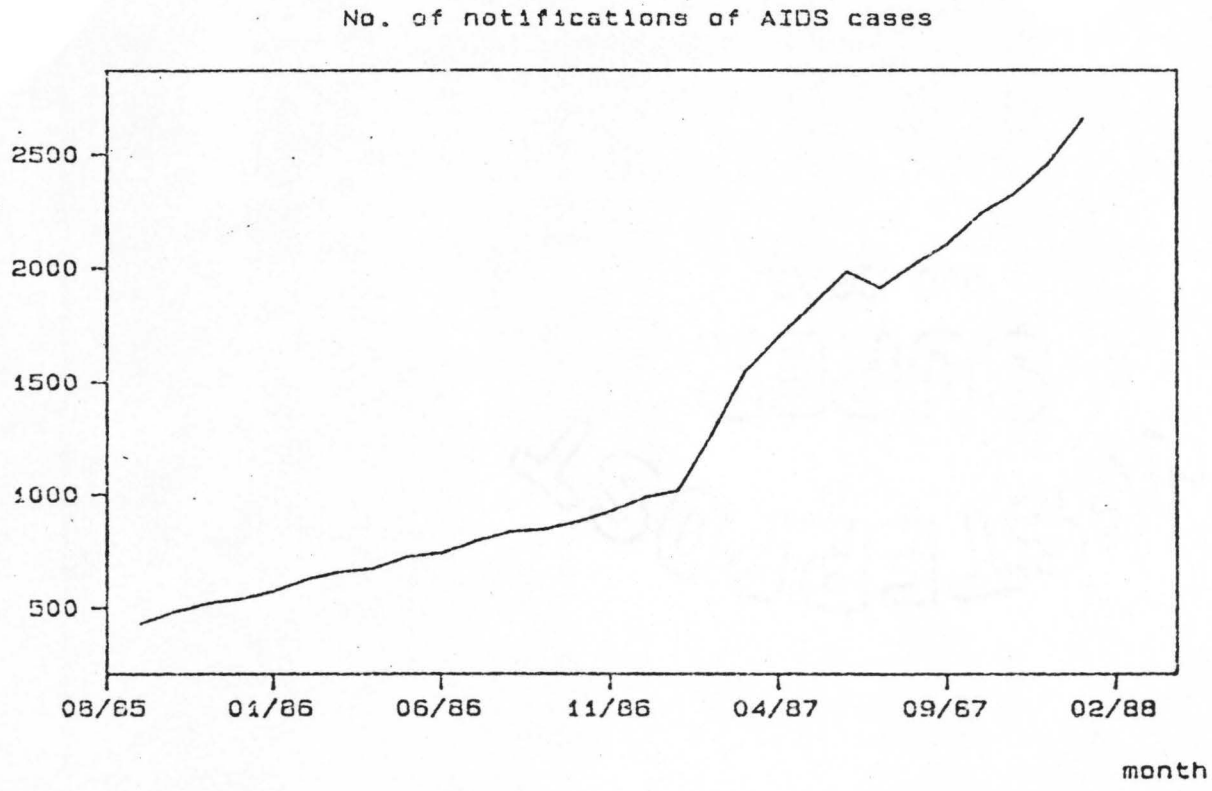
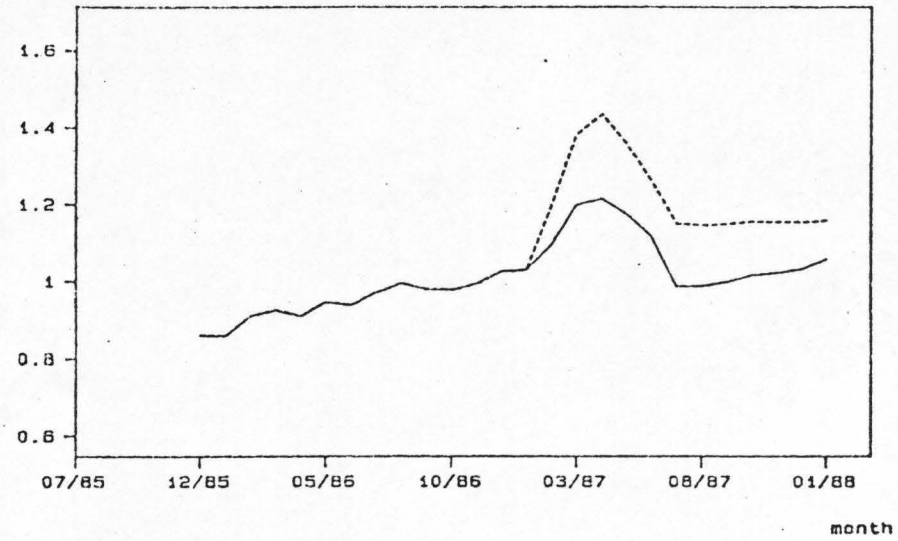
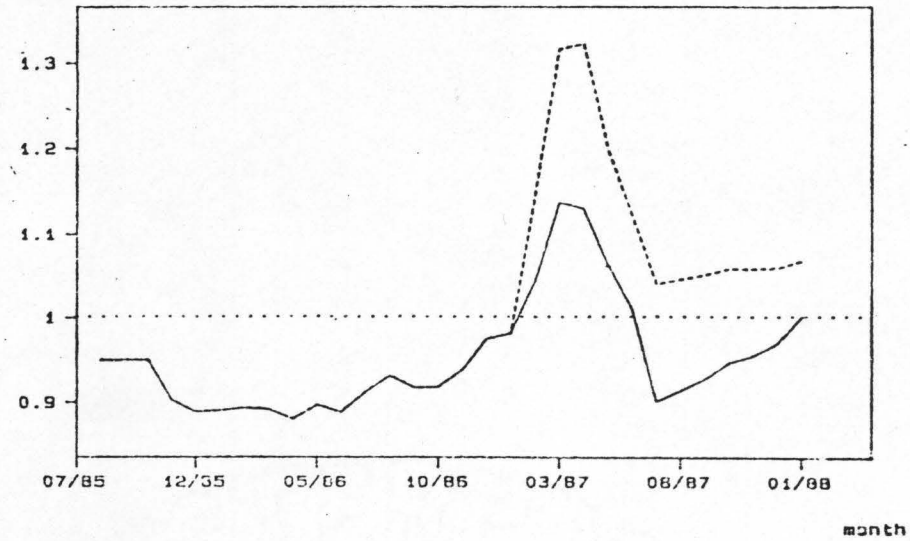

No. of notifications of AIDS cases
2500
2000
1500
1000
500
08/85
01/86
06/86
11/86
04/87
09/87
02/88
month

Figure 1

(a) THETA3 estimates with identity link and with/without intervention



(b) THETA3 estimates with log link and with/without intervention



(c) THETA3 estimates with inverse link and with/without intervention

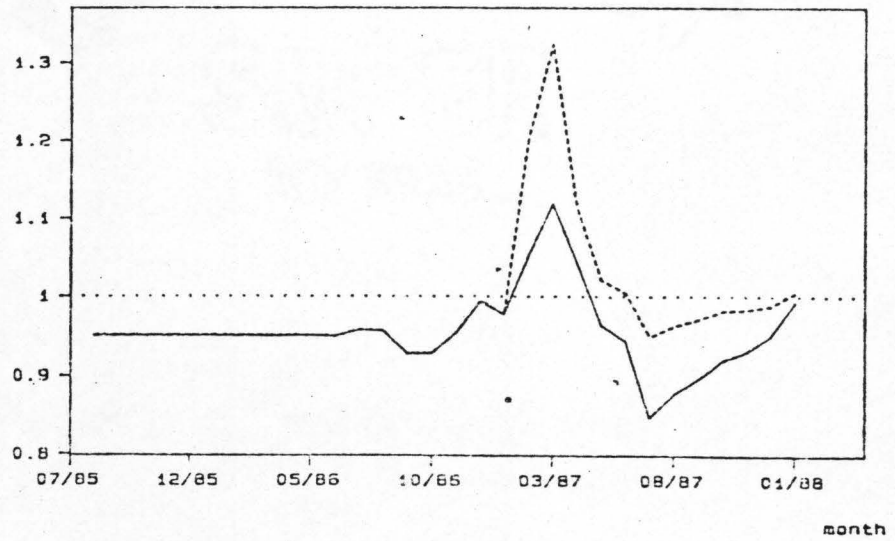


Figure 2

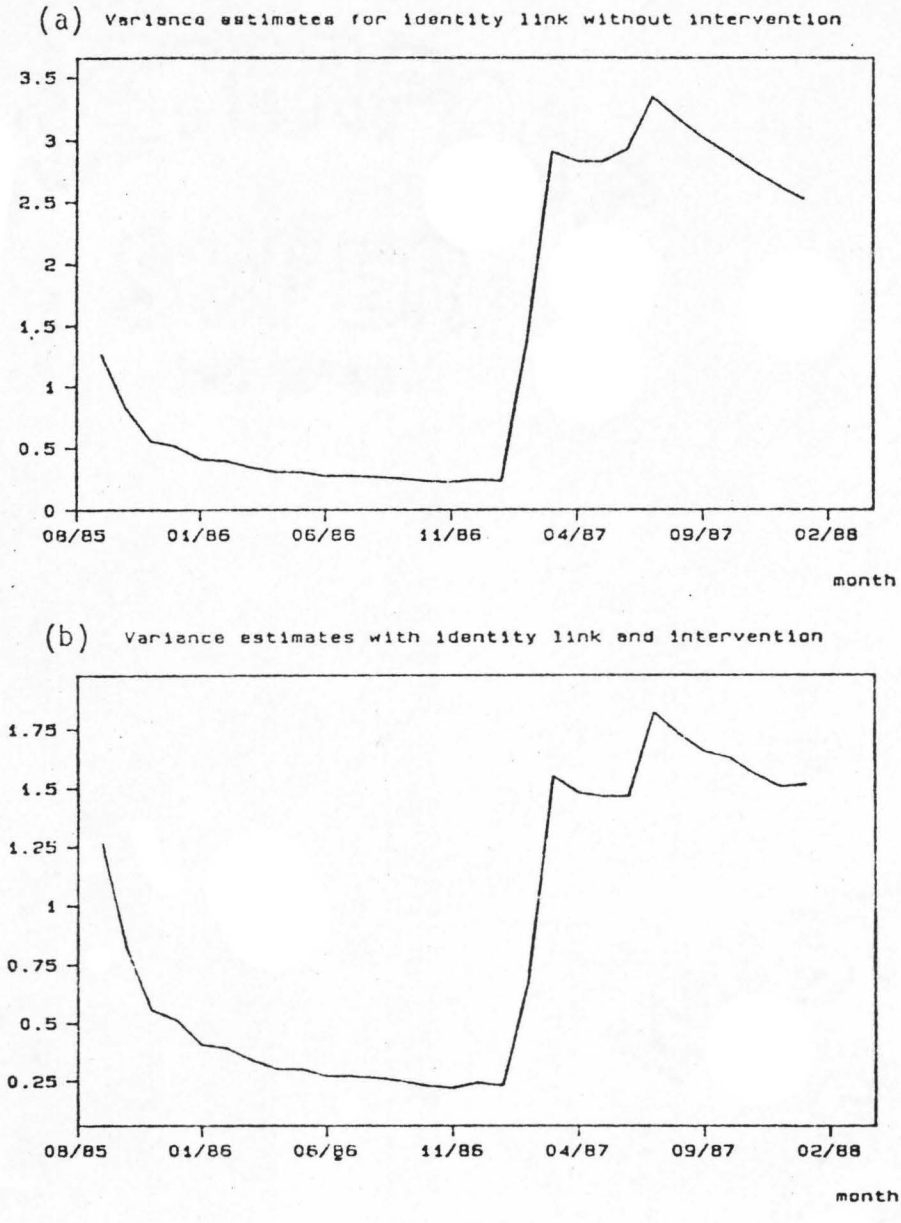
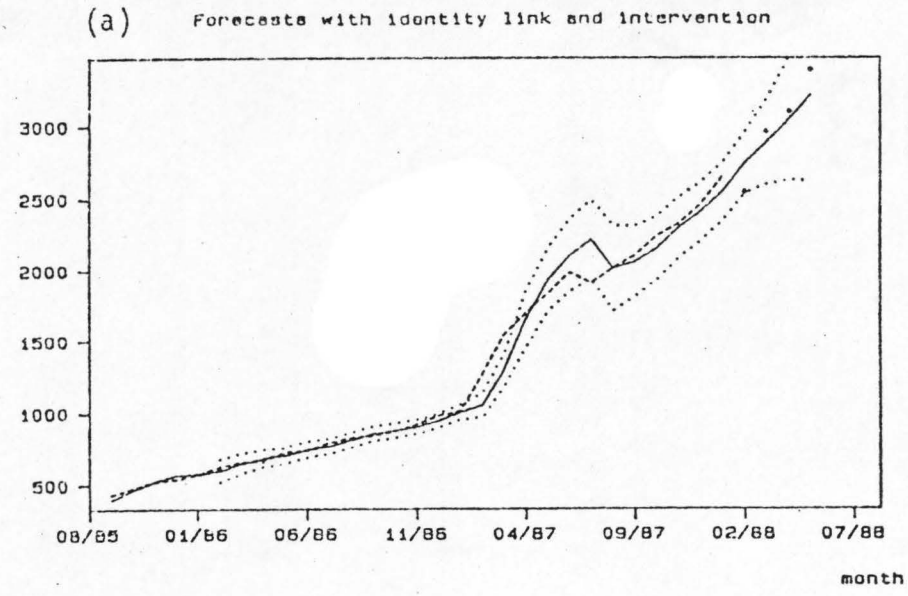
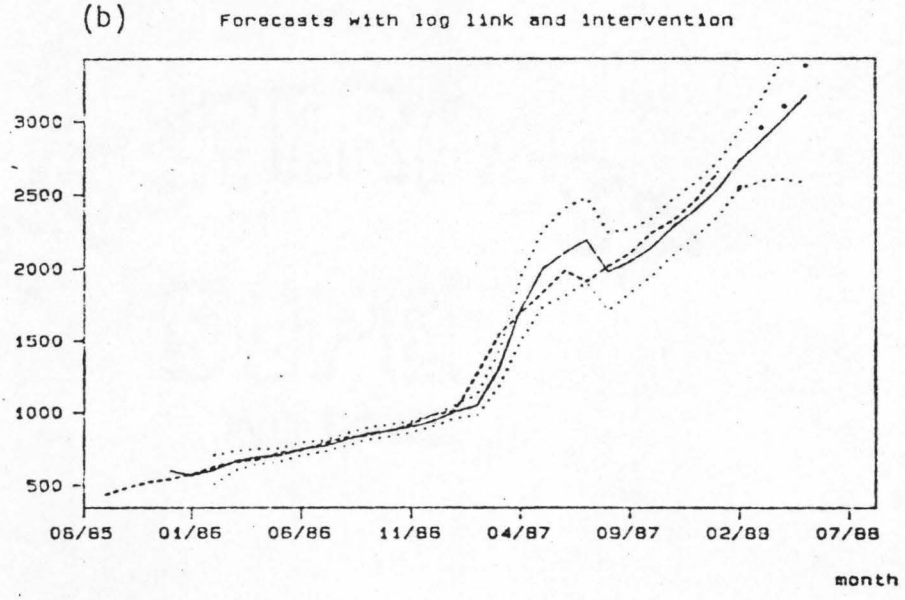

(a) Variance estimates for identity link without intervention
3.5
3
2.5
2
1.5
1
0.5
0
08/85
01/86
06/86
11/86
04/87
09/87
02/88
month
(b) Variance estimates with identity link and intervention
1.75
1.5
1.25
1
0.75
0.5
0.25
08/85
01/86
06/86
11/86
04/87
09/87
02/88
month

Figure 3


(a) Forecasts with identity link and intervention
3000
2500
2000
1500
1000
500
08/85
01/86
06/86
11/86
04/87
09/87
02/88
07/88
month


(b) Forecasts with log link and intervention
3000
2500
2000
1500
1000
500
08/85
01/86
06/86
11/86
04/87
09/87
02/88
07/88
month

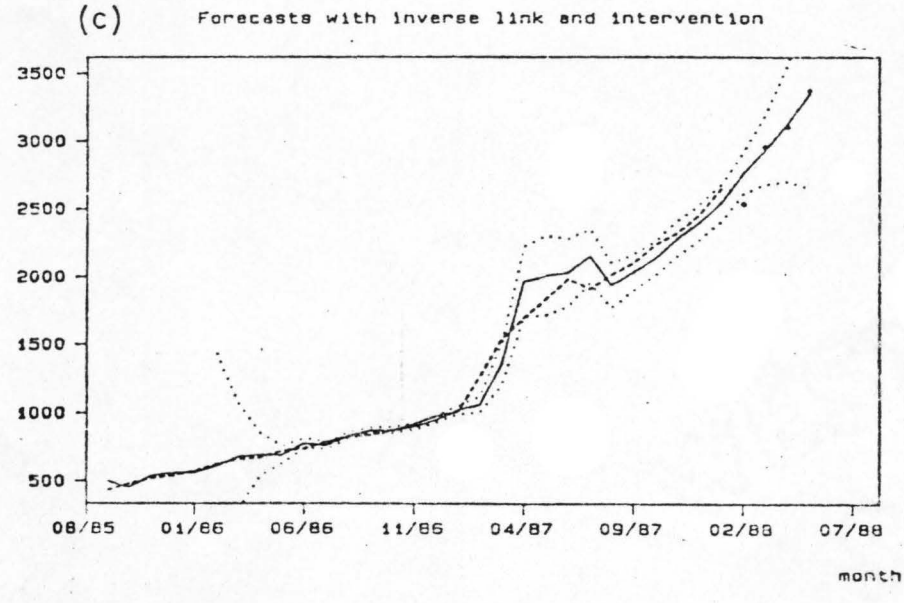
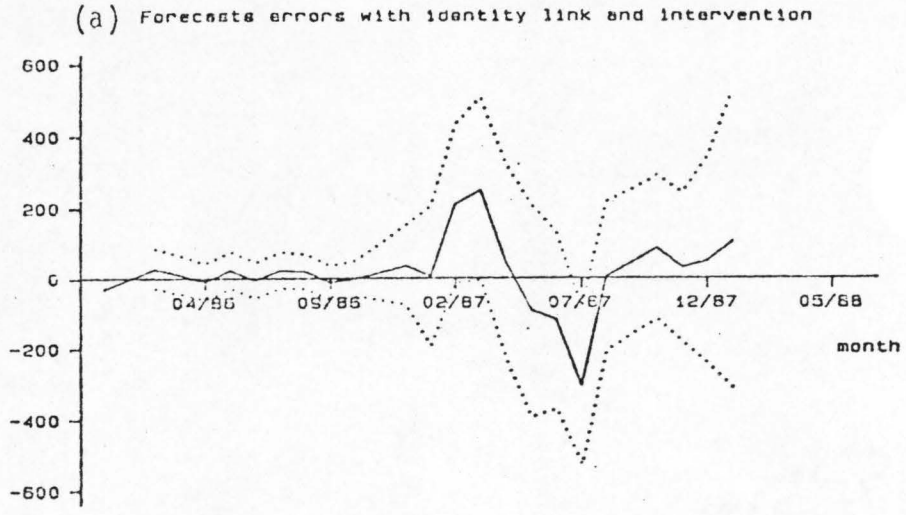
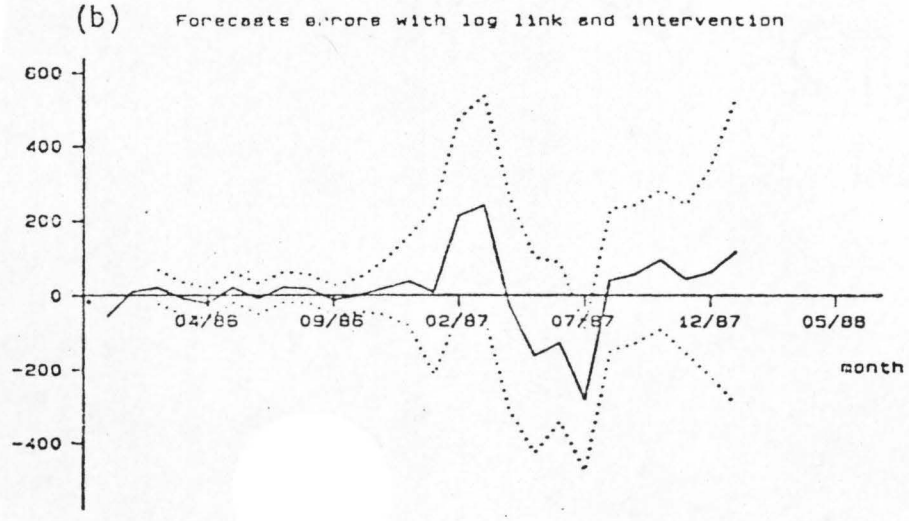

(c) Forecasts with inverse link and intervention
3500
3000
2500
2000
1500
1000
500
08/85
01/86
06/86
11/86
04/87
09/87
02/88
07/88
month

Figure 4


(a) Forecasts errors with identity link and intervention
600
400
200
0
-200
-400
-600
04/86
09/86
02/87
07/87
12/87
05/88
month


(b) Forecasts errors with log link and intervention
600
400
200
0
-200
-400
04/86
09/86
02/87
07/87
12/87
05/88
month

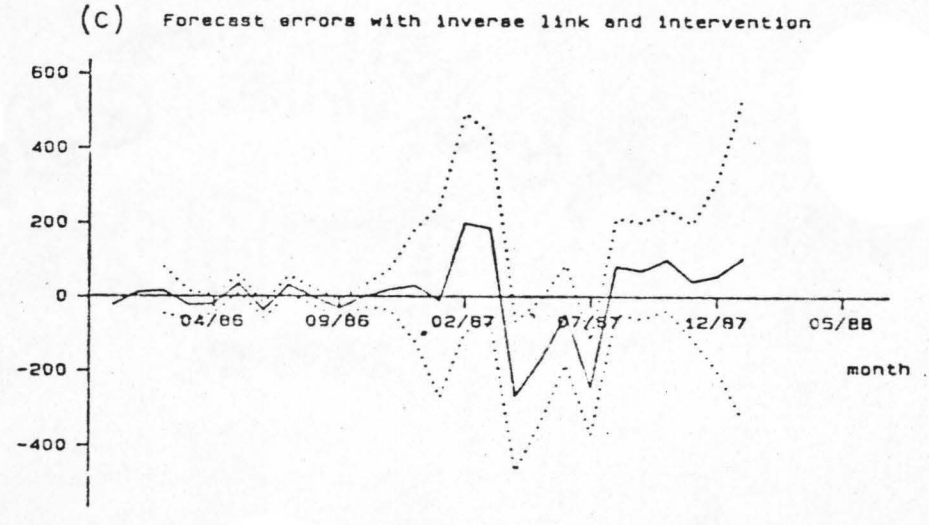

(c) Forecast errors with inverse link and intervention
600
400
200
0
-200
-400
04/86
09/86
02/87
07/87
12/87
05/88
month

Figure 5

NÚMEROS JÁ PUBLICADOS:

01/88 - **CRÍTICA DE RAZÕES NO CENSO ECONÔMICO**

Renato Martins Assunção (ENCE/IBGE)
Rosana de Freitas Castro(DEIND/IBGE)
José Carlos R.C.Pinheiro (ENCE/IBGE)

02/88 - **USO DE AMOSTRAGEM EM SIMULAÇÃO DE LEGISLAÇÃO TRIBUTÁRIA**

José Carlos da Rocha C. Pinheiro (ENCE/IBGE)
Manuel Martins Filho (DISUL/SERPRO)

03/88 - **FORECASTING THE NUMBER OF AIDS CASES IN BRAZIL**

Dani Gamerman (IME/UFRJ)
Hélio S. Migon (ENCE/IBGE & IME/UFRJ)