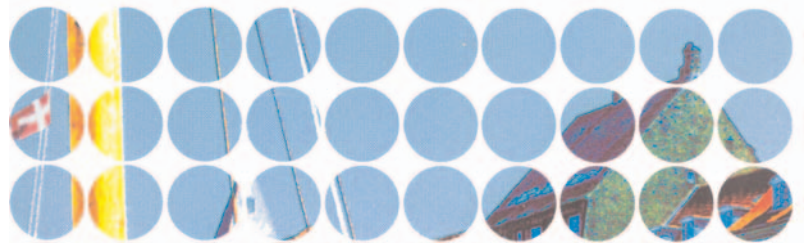




Department for Development and Cohesion Policies
Ministry for Economy and Finance



SAS Forum International
Copenhagen 2004

Building Knowledge From Past Data: an Estimate of Public Expenditure for the "Completamenti" Infrastructure Programme

Francisco V. Barbaro, Carlo Amati

INDEX

Abstract	2
Introduction	2
Programme features and project selection	3
Project features	4
The project's life-cycle	6
The monitoring system	6
Model framework	7
Model specification	8
The model in action	9
What the model does not do	24
Conclusions	25
References	25
Contact information	25

Paper presented at SAS Forum International, Copenhagen, 15-17 June 2004

BUILDING KNOWLEDGE FROM PAST DATA: AN ESTIMATE OF PUBLIC EXPENDITURE FOR THE “COMPLETAMENTI” INFRASTRUCTURE PROGRAMME

Carlo Amati, Ministry of Economy and Finance, Italy
Francisco V. Barbaro, Ministry of Economy and Finance, Italy

ABSTRACT

It appears that public administrations often make biased estimates of the distribution of expenditure over time when they plan the development of infrastructure projects. Most investment programmes are supported by monitoring systems that collect qualitative and quantitative data at single-project level on a regular basis. These data are used mainly for controls but in fact they can be further exploited to build models that, learning from the past, give insight into future behaviour and allow to make more accurate estimates.

We present the case of a relatively small public investments programme, the “Completamenti”, launched in 1999, for which extensive data collection began in mid-2002.

With a mixed model it is possible to combine either fixed or random quantitative variables (i.e. time, cost) and descriptive features (i.e. region, class of duration) and to build a unitary framework for all the projects involved in the programme, which can then be used to estimate the distribution of their expenditure over time.

When the results of the model are aggregated at programme-level they give an estimate of the whole expenditure over time. This, in turn, can support policy-makers’ decisions on future resource allocation programmes.

INTRODUCTION

The Italian panorama of public investments for infrastructures is associated with several monitoring systems, corresponding to different programmes and/or institutions, and managed by central and local administrations. Recent regulations have defined a new global monitoring system of public investments (MIP), that is currently under construction and that will provide a general framework linking all the existing monitoring systems.

In particular, the Public Investment Verification Unit (UVER) of the Ministry of Economy and Finance, to which the authors belong, is working on a project which currently focuses on merging and comparing the quality of data extracted from different databases managed by various government and local/regional public agencies. The goals of this project are twofold. First, to create a consistent dataset on public investment projects. Second, to build an infrastructure development model in order to exploit the aforementioned dataset for policy analysis.

This paper focuses on the work performed on the database, built and managed directly by UVER, containing data on the “Completamenti”, a public infrastructure programme comprising 302 projects worth about 1.5 billion euros, which have been assigned through Government funds to depressed or “under-utilised” (according to the official terminology adopted by the Ministry) areas in 1999. The Interdepartmental Committee for Economic Planning (CIPE), in charge of choosing the projects, appointed UVER for all the reporting activities on the programme.

PROGRAMME FEATURES AND PROJECT SELECTION

The programme’s denomination stems from the specific prerequisites of the projects that could be admitted to funding: priority would be given to projects leading to the conclusion of uncompleted works.

The prompt response of the administrations led to more than 1000 funding requests to CIPE. All requests underwent a selection procedure that assigned a score to each project according to various

criteria, among which the design stage (explained further below), the amount of residual costs and/or co-financing, the consistency with the region’s economic policy objectives, etc.. The financing procedure was meant to combine warranties and awards: 70 per cent of resources would finance the best projects according to regional rankings, thus ensuring that each region would be accredited with a predetermined amount of funds; the remaining 30 per cent would be assigned to the best remaining projects, put together into a single national ranking, thus rewarding the regions with the best projects in absolute.

In summary, CIPE funded 302 projects with 1.5 billion euros of Government funds and co-financed by the sponsor administrations with further 1.1 billion euros; the projects aimed at completing larger uncompleted works worth nearly 7.9 billion euros.

PROJECT FEATURES

The funded projects vary a lot on different dimensions: amounts, durations and locations are widely distributed among most infrastructure sectors, ranging from small cultural heritage projects worth a few thousand euros to a major one in the transportation sector worth nearly 700 million euros, that alone absorbs over 30 per cent of the total funds.

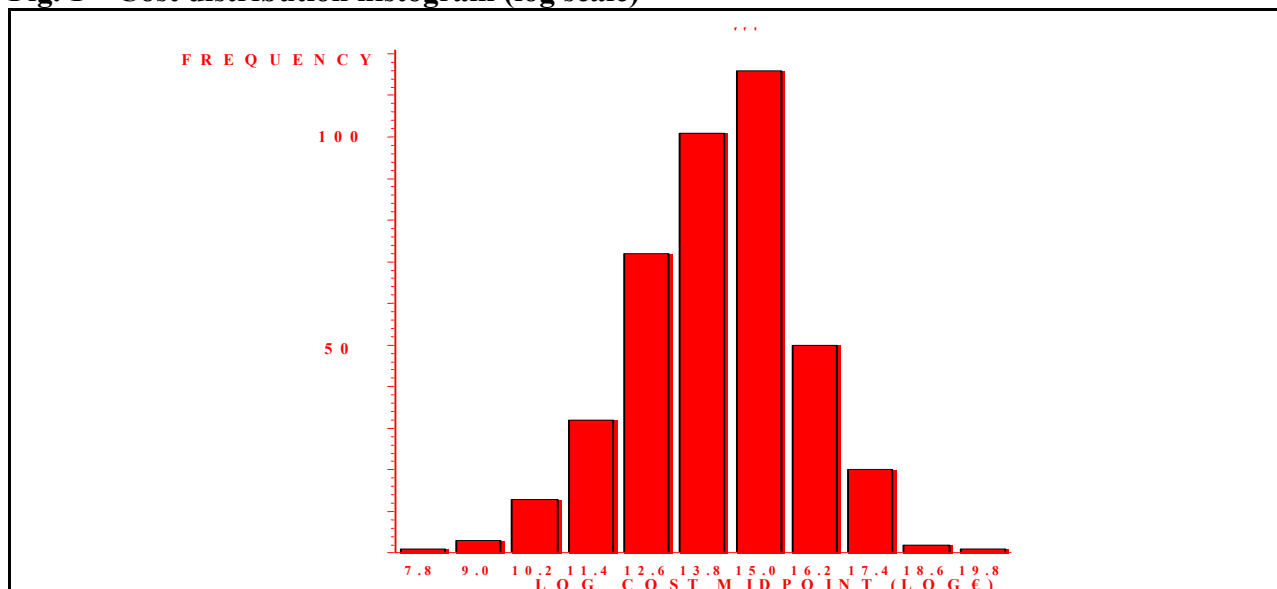
We now briefly review some more details on the features of the projects in the “Completamenti” programme.

A project’s total expected cost is allocated in a very detailed way over several design stages, described in the project’s technical report. The expected cost can be covered with funds from different sources (private and public). Related to the project’s expected cost is the concept of realised cost, which is the amount of money actually spent on the project once completed. Hence, the true value of the cost is only known upon completion. During the project’s development the cost is only provisional. In this paper we always refer to provisional realised cost.

The most common forms of procurement usually produce a significant rebate. For this reason the project’s realised cost will be smaller than the amount specified in the design, unless some cost-increasing factors come about during the project’s development. In some cases the extra cost can be covered by the rebate, whereas in other cases further funding may be needed (indeed, lack of funds is one of the most frequent reasons for uncompleted works).

The cost distribution of the projects is plotted in Fig. 1.

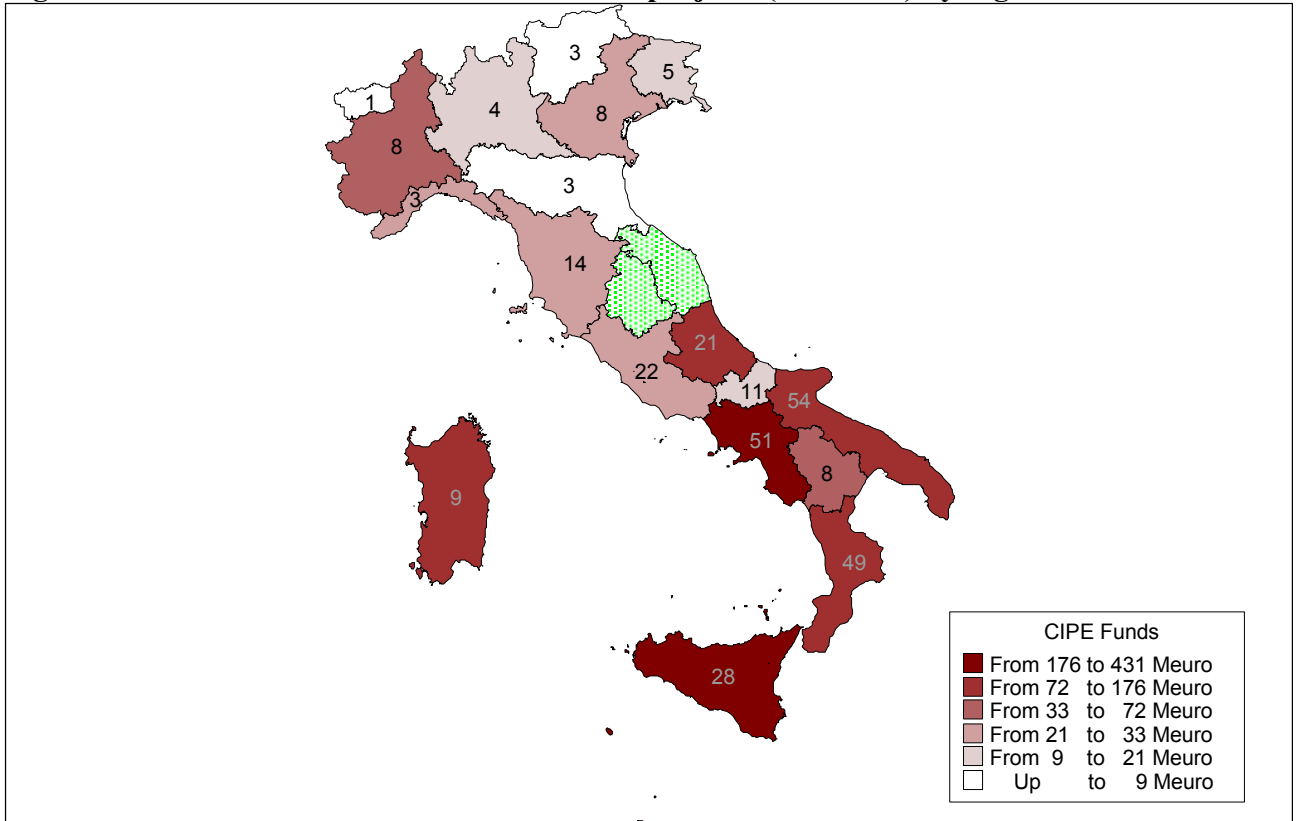
Fig. 1 – Cost distribution histogram (log scale)



The costs are best represented on logarithmic scale as they vary over several orders of magnitude.

The location of a project can be specified down to municipality and may involve more of them (i.e. networks, motorways, etc.); however, for our purposes it suffices to consider regions. The “Completamenti” projects fall in the under-utilised areas of the country, which correspond to all the southern regions and to sparse areas in the Centre-North.

Fig. 2 – Amount of CIPE funds and number of projects (total=302) by region

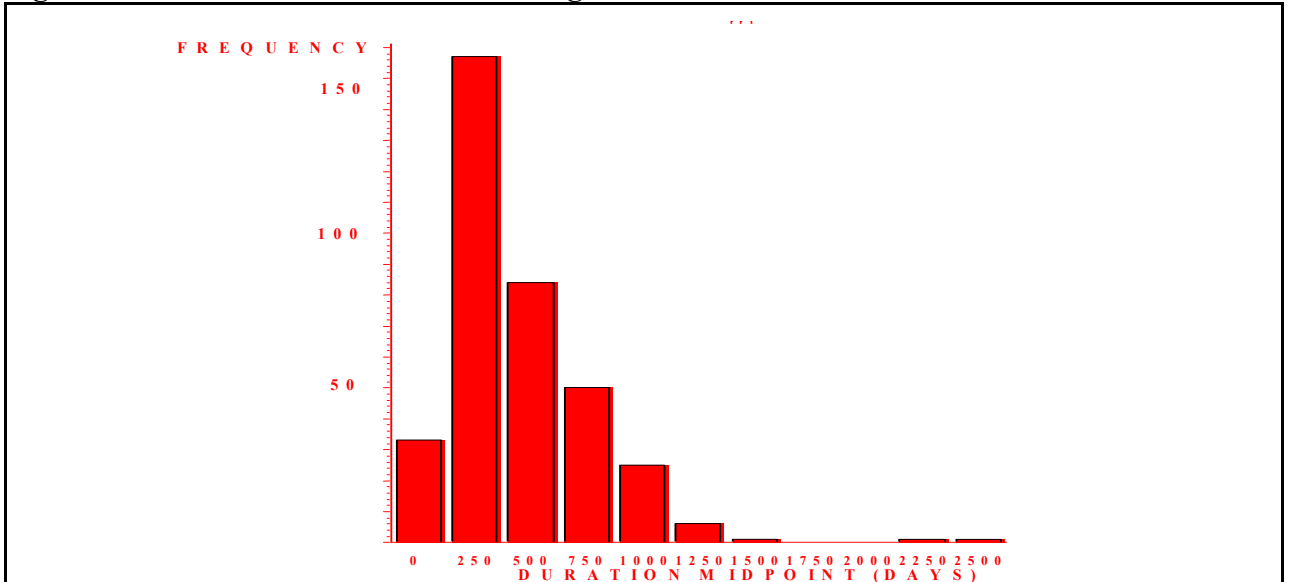


The distribution of CIPE funds is represented in Fig. 2, where the number on each region indicates the number of projects in that region. The projects are located in 19 regions¹ and are more concentrated in the South, where nearly 90 per cent of funds are absorbed.

The detailed design stages described in each project’s technical report define the initial contract duration. Like cost, duration is subject to increase: in fact, as we will later see in more detail, it is frequent that the initial duration is largely over-run and the completed project reaches the operational stage much later than initially planned. This can happen for a number of reasons, among which design changes, unexpected conditions, poor project management, inappropriate contractors, are the most common causes. The distribution of initial durations is shown in Fig. 3.

¹ Two autonomous provinces are assimilated to regions.

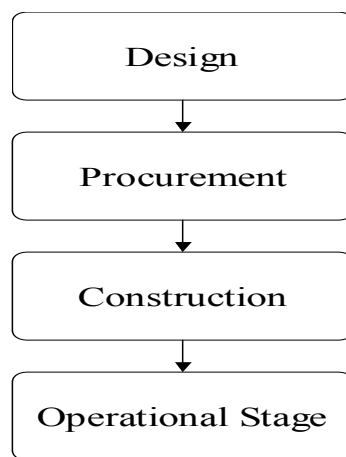
Fig. 3 – Initial duration distribution histogram



THE PROJECT’S LIFE-CYCLE

A project goes through a series of typical stages, namely design, procurement, construction and operational stage.

Fig. 4 – Project’s life-cycle



During the design stage, the project is defined and planned, through a series of progressively more detailed documents. This stage can be very time-consuming, as it is at this stage that all the necessary authorisations must be obtained.

When the design has reached the required form, the project enters procurement stage, that involves the selection process of the firm that will undertake the construction works of the project.

Once the firm is chosen, construction works begin. It is during this stage that the major part of expenditure takes place.

After the works are completed, there is another stage during which tests and other activities are undertaken before the project is fully operational.

THE MONITORING SYSTEM

Even though the project selection procedure put the funded projects under the bull’s eye from the start of the programme, a fully structured monitoring system saw its birth only in mid-2002. This

marked a turning point in the analyses that could be performed: until then the most advanced reports consisted of mere lists of projects by status and by region.

The monitoring system collects quantitative and qualitative data throughout the entire life-cycle of a project. This data is very detailed, referring to every single procurement undertaken within the project. In fact a project can be carried out via different procurements, each with its own physical, procedural and financial identity; in our case the initial number of 302 projects grows up to 411 sub-projects, each with its own entry in the monitoring system. However, for sake of simplicity, from now on when we mention a project we will refer to one of the 411 sub-projects.

The data are collected every four months through forms exchanged either by fax or by email between UVER and the project managers on a voluntary basis. UVER pre-fills in the forms with all the data available from the latest monitoring date and sends them out, addressing them to project managers personally. The returned data are first validated manually by experts and then automatically by the database system. In case of errors or inconsistencies the project managers are contacted for further controls. On average over the different monitoring phases nearly 90 per cent of the projects have responded with a mean response time that has varied between 10 and 20 days.

The comprehensive monitoring system has determined major quality improvements for potential data analyses. The first major improvement relates to the fact that the “generic status information” (not started, started, suspended, completed project) has been enriched with a number indicating the percentage of advancement for each project. Secondly, the availability of cross-sectional observations² on the projects’ advancement has suggested the development of a model for forecasting future expenditure.

MODEL FRAMEWORK

The observed values of the expenditure curve that we intend to model are best approximated by the ratio between the amount of works reported in the latest SAL (a document that lists in detail type and cost of all the works carried out until a certain date) and the total amount of contracted works at each monitoring date.

Although spending on projects can already begin at design phase, especially when the design is contracted-out, in our case most of the financed projects were either designed in-house or provided with a detailed design. We can therefore safely assume that the expenditure curve describes the distribution of expenditures during the construction stage.

Accordingly, we can consider only contractual time, which starts by default at the beginning of construction works.

The monitoring phases that have currently been completed are 5, so this is the maximum number of repeated observations for a project. Each observation point is represented by a couple of coordinates, namely time and percentage of expenditure. In fact we can increase the number of observations by adding the point (0,0) to all projects, that is, by including the information that the expenditure is always zero at start-time.

² Various observations over time

The frequency of the number of observations for the projects are listed below.

The FREQ Procedure					
num_obs	Frequency	Percent	Cumulative Frequency	Cumulative Percent	
1	56	13.63	56	13.63	
2	165	40.15	221	53.77	
3	50	12.17	271	65.94	
4	32	7.79	303	73.72	
5	55	13.38	358	87.10	
6	53	12.90	411	100.00	

Although more than 53 per cent of the projects has one or two observations, by means of considering also the origin at (0,0) for each project, we have 6 points for nearly 13 per cent of the projects³.

The low number of observations for each project confirms the need of a model that jointly exploits all the projects as it would be unfeasible to estimate separate sets of parameters for each project.

MODEL SPECIFICATION

From the data of observed expenditure over time, adding a number of structural features of the projects, it is possible to model the evolution of the percentage of expenditure over time.

The model represents the percentage of expenditure at a given time as a function of the project's structural and contextual features.

Time is considered as a continuous variable, relative to the date of handover to the contractor, i.e. for every project the works are started at $t = 0$. All the other variables (covariates) are static categorical ones, with different numbers of levels.

The objective function that we want to model is the cumulative function of expenditure over time $C_i(t)$, which is expressed as the percentage of expenditure of the i -th project at time t and then varies in $[0,1]$.

From the data in the monitoring system we can observe that, in general, $C_i(t)$ has a slow start, then grows more steadily and slows down again towards the end of the construction stage. In other words it has a sigmoid shape, suggesting that it can be well approximated by a logistic function:

$$C_i(t) = \frac{\exp(\alpha_i + \beta_i t)}{1 + \exp(\alpha_i + \beta_i t)} \quad (1)$$

where α_i and β_i are combinations of effects and parameters for the i -th project.

If we perform a logit transformation on the C function:

$$LC_i(t) = \text{Logit}[C_i(t)] = \ln\left(\frac{C_i(t)}{1 - C_i(t)}\right) = \alpha_i + \beta_i t \quad (2)$$

we obtain a linear function of time that can be estimated easily.

We said that α_i and β_i depend on the i -th project, that is, once the i -th project is identified and the model parameters are known, then we have an $\{\alpha_i, \beta_i\}$ couple for each project. In fact, we might fit a linear model for each of the P projects, obtaining P couples of parameters $\{\alpha_i, \beta_i\}$. This would probably fit well the observed expenditure curves, but would leave no space for any interpretation

³ We will not use the data of two 6-point projects to estimate the model parameters as they are strongly biased.

of the effects. In addition, the estimates of the projects with 2 or less observations would be either highly biased or undetermined.

Instead we are interested in a global model in which the projects are jointly considered for the estimate of a single set of parameters maximising the likelihood of the observed data.

The logit transformation is convenient mathematically in order to have a linearised function whose parameters can be estimated with standard methods. One drawback of this transformation is that it is not defined in 0 and 1, exactly where a lot of our observations are concentrated. A workaround is to shift the extreme values in 0 and 1 of a small quantity on the y-axis to the extreme values of a smaller interval, say, 0.05 and 0.95, shifting accordingly the corresponding time values on the x-axis. The new extreme values have been chosen under the general criterion that they should not be too far from where the next non-extreme observed values are.

All the available features of the projects are considered as fixed effects. In particular we consider both qualitative effects, such as the region where the project is located and the type of sponsor administration, and quantitative effects, such as the initial duration of works and the project's cost. Some quantitative variables are then converted into classes and inserted into the model both as main effects and along with their interaction with time.

The information on the particular shape of the observed curve of advancement over time, instead, enters the model as a random effect. This shape is described by the intercept and slope of the logit regression of the percentage of expenditure of the single project over time. The intercept-slope couples are then clustered into groups that are used as the subject of the random effects, assuming that the parameters used for clustering are randomly distributed.

The model allows us to make predictions of the expenditure for the years ahead and to perform some analyses on the parameters.

An assessment of the goodness of predictions can be achieved using the estimated parameters. Since all the covariates are static, from the (2) we have an $\{\alpha_i, \beta_i\}$ couple for each project that does not change in time.

Now, let's consider their meaning in relation to the curve $C_i(t)$ represented in (1). The derivative of $C_i(t)$, that can be interpreted as the intensity of expenditure⁴, has the form:

$$C'_i(t) = \beta_i C_i(t)(1 - C_i(t)) \quad (3)$$

so β_i is the parameter that governs the growth rate of the expenditure. The larger β_i , the faster the resources are spent.

It can be shown that the intensity of expenditure $C'_i(t)$ reaches its maximum when $t = -\alpha_i/\beta_i$ and is symmetric with respect to it. By definition, the duration of a project starts at $t = 0$, so the symmetry of the intensity curve allows us to identify an expected duration of each project, $T_i = -2\alpha_i/\beta_i$. It is then clear that, given the growth rate β_i , α_i is the parameter that governs the duration of the project,.

THE MODEL IN ACTION

The first runs of the model are made with fixed effects only, starting with all the available covariates. The class variables are initially spread into many levels and the model is progressively refined according to the significance of the effects and of the difference between levels.

⁴ A more accurate definition for the derivative of the cumulative curve of expenditure would be density of expenditure.

So we start by considering all the following variables both as main effects and in interaction with time:

Name	Description	Type	Values
logc	Logarithm of cost (€)	Continuous	7.35-20.35
dur	Initial duration of works	Class	1: $dur \leq 1$ yr; 2: $1 \text{ yr} < dur \leq 2$ yrs; 3: $2 \text{ yrs} < dur \leq 3$ yrs; 4: $3 \text{ yrs} < dur \leq 5$ yrs; 5 : $dur > 5$ yrs
reg	Region	Class	Abruzzo; Basilicata; Calabria; Campania; Emilia Romagna; Friuli Venezia Giulia; Lazio; Liguria; Lombardia; Molise; Piemonte; P.A. Bolzano; P.A. Trento; Puglia; Sardegna; Sicilia; Toscana; Valle D'Aosta; Veneto
end	Observed end-year of works	Class	1: 2000; 2: 2001; 3: 2002; 4: 2003; 5: 2004; 99: Not ended
adm	Sponsor administration	Class	1: Central; 2: Regional

Time itself is a continuous variable expressed in units of 1000 days (i.e. $time=0.5$ represents a calendar time of 500 days). Although duration and end-year are natively expressed as either number of years or calendar dates, i.e. continuous variables, it is better to convert them into class variables, because some projects have missing data. For example, we do not have an end-year in case of projects that are still running, and we might not have an initial duration for projects still at design stage. In the first case (end-year) it is straightforward to put all the running projects into the same class level (99: Not ended); in the latter (initial duration), the same procedure would not be meaningful, because we would be grouping together projects with potentially different durations instead of assigning them the correct class level. Then we can use a different criterion, following the assumption that projects with a small duration will have reasonably low costs and viceversa. In that case, we group the projects into cost classes and we assign a project the `dur` class level corresponding to its cost class level.

The SAS procedure that we use is `proc mixed` because, even if we are now considering only fixed effects, we will later see how additional random effects boost the goodness of the model.

The SAS code is the following:

```
proc mixed data=completamenti ic noclprint covtest;
  class dur reg end adm
  model logit= logc dur reg end adm
           time time*logc time*dur time*reg time*end time*adm
           / ddfm=satterth;
run;
```

The `ddfmsatterth` option is intended to produce an accurate F-approximation.

The following table shows some general information on the model. In particular, we can see that the number of data points is 1245.

The Mixed Procedure	
Model Information	
Data Set	WORK.COMPLETAMENTI
Dependent Variable	logit
Covariance Structure	Diagonal
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Residual
Dimensions	
Covariance Parameters	1
Columns in X	34
Columns in Z	0
Subjects	1
Max Obs Per Subject	1245
Observations Used	1245
Observations Not Used	0
Total Observations	1245

From the results shown below, we see that `logc`, `reg`, `adm` and `time*adm` are definitely not significant and can be discarded. On the contrary, `logc*time` is only slightly above the significance threshold and for now is kept in the model.

The Mixed Procedure				
Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
logc	1	1185	0.74	0.3892
dur	4	1185	3.88	0.0039
reg	18	1185	1.24	0.2233
end	5	1185	4.03	0.0013
adm	1	1185	1.72	0.1896
time	1	1185	30.25	<.0001
logc*time	1	1185	3.61	0.0577
time*dur	4	1185	2.88	0.0218
time*reg	18	1185	3.76	<.0001
time*end	5	1185	15.70	<.0001
time*adm	1	1185	1.31	0.2527

The model is re-run using the following code:

```
proc mixed data=completamenti ic noclprint covtest;
  class dur reg end adm
  model logit= dur end
           time time*logc time*dur time*reg time*end
           / ddfm=satterth;
run;
```

This time all the effects are significant, including `time*logc`, that was not significant in the previous run. If we add the next not significant effect from the previous run (`time*dur`), using a forward selection method, the result is that the effect remains not significant.

The Mixed Procedure				
Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
dur	4	1206	6.53	<.0001
end	5	1206	4.34	0.0006
time	1	1206	57.17	<.0001
time*logc	1	1206	15.09	0.0001
time*dur	4	1206	4.13	0.0025
time*reg	18	1206	4.60	<.0001
time*end	5	1206	16.90	<.0001

Now, let's consider the significance of the difference between the levels of the class variables. From the solution for the fixed effect `dur` shown below, we see that the parameters for some of the levels are close to each other, especially if we consider the size of the error.

The Mixed Procedure								
Solution for Fixed Effects								
Effect	dur	adm	end	Estimate	Error	DF	Standard t Value	Pr > t
dur	1			1.1551	2.0352	1205	0.57	0.5704
dur	2			0.4185	2.0333	1205	0.21	0.8370
dur	3			0.1483	2.0396	1205	0.07	0.9420
dur	4			0.5081	2.0925	1205	0.24	0.8082
dur	5			0

In this case we can use `proc mixed` to estimate whether the difference between the parameters of two levels is significant. If the difference is not significant then we can group the levels. This can be easily done for `dur` and `end`, as they are ordered variables, whereas for the regions it makes more sense to group them into macro-areas (Centre-North and South). All the desired estimates are obtained with the following SAS code:

```
proc mixed data=completamenti ic noclprint covtest;
  class dur ma end adm
  model logit= logc dur reg end adm
           time time*logc time*dur time*reg time*end time*adm
  / ddfm=satterth;
  estimate 'dur 1-2' dur 1 -1 0 0 0;
  estimate 'dur 2-3' dur 0 1 -1 0 0;
  estimate 'dur 3-4' dur 0 0 1 -1 0;
  estimate 'dur 4-5' dur 0 0 0 1 -1;
  estimate 'end 1-2' end 1 -1 0 0 0 0;
  estimate 'end 2-3' end 0 1 -1 0 0 0;
  estimate 'end 3-4' end 0 0 1 -1 0 0;
  estimate 'end 4-5' end 0 0 0 1 -1 0;
```

```

estimate 'time*end 1-2' time*end 1 -1 0 0 0 0;
estimate 'time*end 2-3' time*end 0 1 -1 0 0 0;
estimate 'time*end 3-4' time*end 0 0 1 -1 0 0;
estimate 'time*end 4-5' time*end 0 0 0 1 -1 0;

```

run;

The 99 level for the end effect is left out of the estimates as it corresponds to running projects without an observed end date: it would not make sense to group these projects with some others, only on the basis of similar values of the parameters. However, it can be shown that an analysis would confirm a significant difference between the 99 level and the others.

The estimates of the differences are shown below. A significant difference for dur appears only between levels 1-2. This means that we can group together all levels from 2 to 5. In other words, given the other factors, there is a statistical difference between projects with initial duration up to one year and the other ones.

An analogous reasoning for the end effect leads us to group levels 1-2 and 4-5. From the estimates of the differences between levels of the main effect we might as well group level 3 with 4-5, but when the effect is considered in interaction with time the difference 3-4 becomes significant, so we keep level 3 separated.

The results for the macro-area show that is neither significant as a main effect nor in interaction with time so for the further analyses we will keep the reg disaggregated effect.

The Mixed Procedure					
Estimates					
Label	Estimate	Standard Error	DF	t Value	Pr > t
dur 1-2	0.7369	0.1643	1206	4.49	<.0001
dur 2-3	0.2713	0.2416	1206	1.12	0.2616
dur 3-4	-0.3650	0.5530	1206	-0.66	0.5094
dur 4-5	0.5121	2.0915	1206	0.24	0.8066
end 1-2	0.2081	0.3785	1206	0.55	0.5826
end 2-3	0.6895	0.2184	1206	3.16	0.0016
end 3-4	-0.2523	0.2166	1206	-1.16	0.2443
end 4-5	1.8982	1.4297	1206	1.33	0.1845
time*end 1-2	0.9743	1.0644	1206	0.92	0.3602
time*end 2-3	-0.9414	0.4558	1206	-2.07	0.0391
time*end 3-4	1.9153	0.3573	1206	5.36	<.0001
time*end 4-5	-1.3004	1.8906	1206	-0.69	0.4917

The final version of the fixed-effects model contains the following variables:

Name	Description	Type	Values
logc	Logarithm of cost (€)	Continuous	7.35-20.35
dur2	Initial duration of works	Class	1: dur ≤ 1 yr; 2: dur > 1 yr
reg	Region	Class	Abruzzo; Basilicata; Calabria; Campania; Emilia Romagna; Friuli Venezia Giulia; Lazio; Liguria; Lombardia; Molise; Piemonte; P.A. Bolzano; P.A. Trento; Puglia; Sardegna; Sicilia; Toscana; Valle D'Aosta; Veneto
end2	Observed end-year of works	Class	1: 2000 - 2001; 2: 2002; 3: 2003 – 2004; 99: Not ended

Let's see some fit statistics. They will be most useful when compared with the improved model.

The Mixed Procedure				
Covariance Parameter Estimates				
Cov Parm	Estimate	Standard Error	Z Value	Pr > Z
Residual	2.0724	0.08405	24.66	<.0001
Fit Statistics				
	-2 Res Log Likelihood		4428.5	
	AIC (smaller is better)		4430.5	
	AICC (smaller is better)		4430.5	
	BIC (smaller is better)		4435.6	

Before interpreting the results of the model above, we can perform some more explicit measures of its goodness.

It is important to remember that our aim is to predict the evolution of expenditure, so we need a model that fits well the existing data and is general at the same time (the latter reason makes us keep only significant effects).

A measure of the goodness of fit can be given by the correlation between the observed data and the corresponding model outcome; the correlation, though, is a measure of co-linearity between two data sets, whereas the best fit is obtained where the model outcome is identical to the observed data. Therefore a measure of identity is given by the slope of the regression line of the model outcome over observed data. A unity slope and a zero intercept correspond to identity.

Since we are interested in the overall expenditure evolution, the following measures are made on the true data, back-transforming the model outcome into the interval (0,1) and pondering the projects' data by their costs.

For the fixed effects model we have:

The CORR Procedure					
Pearson Correlation Coefficients, N = 1245					
Prob > r under H0: Rho=0					
		observed	estimated		
observed		1.00000	0.89082	<.0001	
estimated		0.89082	1.00000	<.0001	

The REG Procedure					
Parameter Estimates					
Dependent Variable: estimated					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
observed	1	0.90065	0.00847	106.31	<.0001

Both correlation and slope are quite high, but the 0.90 slope shows a systematic underestimate of the expenditure.

A major improvement can be obtained with the addition of a random component both to the intercept and the slope of the (2) that now reads:

$$LC_i(t) = (\alpha_i + a_k) + (\beta_i + b_k)t \quad (4)$$

for the the k -th value of the variable that governs the random components, still to be identified.

The most natural candidate for the subject of the random effect is the project itself: the projects in the “Completamenti” programme are chosen at random from the project universe. Having 411 projects this choice would lead to the estimate of 822 additional parameters; it is clear that although with 1245 data points the problem is still well-defined, the number of parameters grows too much.

So, in order to reduce the number of parameters, we try to group the projects according to some criterion and use the group as the subject of the random components.

The effect of the random components is to vary the expenditure profile of the projects: therefore we just try to use the expenditure profile as a grouping criterion.

This means that the relation between the i index and the k index in (3) is n to 1 and the formula can still be expressed in the synthetic form:

$$LC_i(t) = A_i + B_i t \quad (5)$$

Let’s say something on the clustering procedure. We have already seen that for each project we have a number of observations ranging from 1 to 6. All 355 projects with 2 or more observations have already produced part of an expenditure curve. A logit regression performed singularly on each of these projects yields 355 intercept-slope couples of parameters that are used for bidimensional clustering.

As for the projects with 1 observation, corresponding to zero expenditure, they are all put into the same cluster.

The above procedure yields 58 clusters, identified by the new `clus` variable, that is used as subject for the random effects.

The model selection procedure carried out for the fixed effect model starts again for the random effects model.

The variables are the same used for the first run of fixed effects model, plus the following:

Name	Description	Type	Values
<code>clus</code>	Identifier of projects cluster	Class	CL01-CL58

The SAS code for the mixed model with fixed and random effects is:

```
proc mixed data=completamenti ic noclprint covtest;
  class dur reg end adm clus;
  model logit= logc dur reg end adm
         time time*logc time*dur time*reg time*end time*adm
         / ddfm=satterth;
  random int time / subject=clus type=un;
run;
```

The option `type=un` allows the estimate of an additional random intercept-slope covariance. Nevertheless it ends up to be not significant, as shown below and is then discarded keeping the default option `type=vc`.

The Mixed Procedure					
Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr > Z
UN(1,1)	clus	8.0551	3.3211	2.43	0.0076
UN(2,1)	clus	-1.9413	7.7723	-0.25	0.8028
UN(2,2)	clus	253.01	61.0653	4.14	<.0001
Residual		0.5179	0.02374	21.81	<.0001

In the new run with 2 random effects, the fixed effects have the following significance levels.

The Mixed Procedure				
Type 3 Tests of Fixed Effects				
Effect	DF	Num DF	Den F Value	Pr > F
logc	1	1023	0.00	0.9817
dur	4	59.5	1.39	0.2484
reg	18	992	0.46	0.9751
end	5	1001	1.85	0.1009
adm	1	993	1.10	0.2944
time	1	43	19.55	<.0001
logc*time	1	1009	9.20	0.0025
time*dur	4	128	1.02	0.3988
time*reg	18	979	1.19	0.2626
time*end	5	986	3.31	0.0057
time*adm	1	982	1.33	0.2487

It appears that only time, logc*time and time*end are significant. This time we use a backward elimination method, manually removing one effects at a time, according to their level of significance.

This methods ends up with the following significant fixed effects:

The Mixed Procedure				
Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
dur	4	60.4	2.99	0.0254
end	5	1024	2.46	0.0318
time	1	45.7	65.17	<.0001
time*logc	1	1005	17.25	<.0001
time*reg	18	996	2.21	0.0026
time*end	5	1009	2.92	0.0127

The intercept is not significant either. Now, as before, we try to reduce the number of levels of the fixed effects. An attempt to substitute the region with the macro-area leads to non significant macro-area. As for the other effects, we use the `estimate` option as before. Here are the results:

The Mixed Procedure					
Estimates					
Label	Estimate	Standard Error	DF	t Value	Pr > t
dur 1-2	0.1001	0.06096	1005	1.64	0.1010
dur 2-3	0.04700	0.07439	998	0.63	0.5277
dur 3-4	-0.4542	0.1514	993	-3.00	0.0028
dur 4-5	-1.0866	2.9939	17.3	-0.36	0.7210
end 1-2	-0.6610	0.2343	1040	-2.82	0.0049
end 2-3	0.1810	0.2185	1011	0.83	0.4076
end 3-4	0.1654	0.1381	1024	1.20	0.2315
end 4-5	0.2647	0.1489	1057	1.78	0.0758
time*end 1-2	1.4685	0.5810	1007	2.53	0.0116
time*end 2-3	-2.0244	0.5872	998	-3.45	0.0006
time*end 3-4	0.4965	0.3375	1009	1.47	0.1416
time*end 4-5	-0.1813	0.2403	1036	-0.75	0.4509

The levels of the above effects can be aggregated as follows:

Name	Description	Type	Values
dur2	Initial duration of works	Class	1: dur ≤ 3 yrs; 2: dur > 3 yrs
end2	Observed end-year of works	Class	1: 2000 2: 2001; 3: 2002 – 2004; 99: Not ended

The final version of the full model yields the following results:

The Mixed Procedure					
Covariance Parameter Estimates					
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Intercept	clus	8.0004	2.9729	2.69	0.0036
time	clus	244.38	57.6985	4.24	<.0001
Residual		0.5149	0.02306	22.33	<.0001
Fit Statistics					
		-2 Res Log Likelihood		3215.8	
		AIC (smaller is better)		3221.8	
		AICC (smaller is better)		3221.9	
		BIC (smaller is better)		3228.0	

If we compare them with those of the fixed effects model, previously reported, we can notice that the residual variance is about one quarter of the previous one and also `-2ResLogLikelihood` is

significantly lower. Let's see how these improvements are confirmed by the measures of the goodness of fit.

The CORR Procedure					
Pearson Correlation Coefficients, N = 1245					
Prob > r under H0: Rho=0					
		observed		estimated	
	observed	1.00000		0.97057	<.0001
	estimated	0.97057		1.00000	<.0001
The REG Procedure					
Parameter Estimates					
Dependent Variable: estimated					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
observed	1	0.96839	0.00465	208.31	<.0001

The correlation has increased to 0.97 and also the slope of the regression of the predictions over the observed values is closer to 1, even if the difference is still significant.

We can conclude that the mixed model is better than the one with fixed effects only.

Now, before coming to the main goal of this work, that is, an estimate of the expenditure for the years ahead, let's focus on the meaning of the model results.

Let's start with an analysis of the parameters of the fixed effects, keeping in mind that the interpretation of the parameters taken singularly might not always be straightforward as the structure of the data is strongly unbalanced.

The estimates for the parameters of the model are the following:

The Mixed Procedure							
Solution for Fixed Effects							
Effect	dur2	end2	Estimate	Error	DF	Standard t Value	Pr > t
Intercept			-4.4656	0.4497	24	-9.93	<.0001
dur2	1		-0.4205	0.1423	1014	-2.95	0.0032
dur2	2		0
end2		1	0.5737	0.2262	1037	2.54	0.0114
end2		2	0.3938	0.1571	1051	2.51	0.0124
end2		3	0.1700	0.1033	1041	1.64	0.1003
end2		99	0
time			19.1732	2.2390	45.9	8.56	<.0001
time*logc			-0.1792	0.03241	1009	-0.53	<.0001
time*reg		
time*end2		1	-1.4707	0.5599	1012	-2.63	0.0088
time*end2		2	0.6088	0.3498	1026	1.74	0.0821
time*end2		3	0.1380	0.1529	1025	0.90	0.3671
time*end2		99	0

We have intentionally dropped the values of the `time*reg` effect from the above table, because they do not give particular information on the projects' development, other than highlighting some differences between the expenditure skills of the regions (not always significant due to the restricted number of projects in some regions).

The parameter that can be more readily interpreted is `time*logc`. Being negative, it implies that greater costs are associated with smaller growth rates of the expenditure expressed as a percentage. In other words, it is easier to produce, say a 10 per cent expenditure, for a project that costs 100 than for one that costs 1000, because the same percentage corresponds, in absolute terms, to an expenditure of 10 for the first and 100 for the latter. In the same amount of time, it is generally easier to spend smaller amounts.

From the parameters associated to `dur2` we see that `dur2=1` is negative: the relative increase of predicted duration is greater for projects with initial duration less than 3 years. This is easier to understand with an example: an increase of 100 days weighs proportionally more on shorter projects.

From the parameters associated to `end2` we see that, with respect to running projects (level 99), those ended in or after 2001 (levels 2-3) spend faster and have smaller predicted durations. A similar interpretation of level 1 for `time*end2`, which is negative, would lead us to think that projects ended in 2000 had a smaller growth rate. This is only partly true, as the negative effect is well compensated by the random effect on slope.

Let's remember how the effects are related to expenditure. The development of the percentage of expenditure is represented by a sigmoid function which is ruled by an $A + Bt$ factor. B represents the sum of the parameters of the significant interaction between the main effects and time, plus the parameter of `time`, and is related to the growth rate of the percentage of expenditure in a time interval: greater parameters imply that in the same time interval this percentage grows faster. All B_i 's must be positive.

The parameters of the other main effects, instead, are summed into A and are related to the projects' predicted duration $-2A/B$. Given B , the duration is proportional to the absolute value of A . All A_i 's must be negative.

An analysis of expected duration and growth rate either with respect to observed data or split into some class levels is much interesting.

First of all, a check is made that the sign constraints on the A_i 's and B_i 's are satisfied for all projects. This is confirmed by the simples statistics on their distributions, reported below.

The MEANS Procedure					
Variable	N	Mean	Std Dev	Minimum	Maximum
A	411	-3.4717493	1.1676945	-11.0777749	-0.1699082
B	411	10.1113712	10.3313835	1.7946607	65.2473542

Let's see the model results in terms of duration of the projects. We want to test if there is a significant difference between the observed duration of the finished projects and the corresponding estimate of the model. The task is performed with the following code, that performs a one sample t-test for the hypothesis of zero means for the selected variables:

```
proc ttest data=results;
  where end=1;
  var dur_obs dur_est dur_diff;
  weight cost;
run;
```

The projects are weighted by cost, as it important to express duration in terms of time necessary to spend a certain amount of money. The results are as follows:

The TTEST Procedure					
Weight: cost					
Statistics					
Variable	N	Lower CL Mean	Mean	Upper CL Mean	Std Err
dur_obs	239	1098.3	1231.6	1365	67.7
dur_est	239	1098.7	1180.9	1263.1	41.717
dur_diff	239	-127.2	-50.71	25.799	38.838
T-Tests					
Variable	DF	t Value	Pr > t		
dur_obs	238	18.19	<.0001		
dur_est	238	28.31	<.0001		
dur_diff	238	-1.31	0.1929		

The difference between the means of true and expected duration of finished projects is not significant.

The regression of estimated durations over the observed ones shows that the model, although good on average, overestimates the duration slightly.

The REG Procedure					
Dependent Variable: dur_est					
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
dur_obs	1	1.09342	0.02126	51.42	<.0001

A more effective way of considering duration is to examine the ratio between the final observed duration and the initial one, that is, the time-increase ratio (*t_{ir}*). With the following code we perform a two-sample t-test for the hypothesis that the two sample means of the selected variables are different.

```
proc ttest data=results;
  class end;
  var tir_obs tir_est tir_diff;
  weight cost;
run;
```

The TTEST Procedure						
Weight: cost						
Statistics						
Variable	end	N	Lower CL Mean	Mean	Upper CL Mean	Std Err
tir_obs	0
tir_obs	1	236	1.8155	1.9779	2.1402	0.0824
tir_obs	Diff (1-2)
tir_est	0	122	1.7719	1.9559	2.14	0.093
tir_est	1	236	1.9945	2.1955	2.3965	0.102
tir_est	Diff (1-2)	.	-0.521	-0.24	0.0417	0.143
tir_diff	0
tir_diff	1	236	-0.299	-0.218	-0.136	0.0413
tir_diff	Diff (1-2)
T-Tests						
Variable	Method	Variances	DF	t Value	Pr > t	
tir_obs	Pooled	Equal	0	.	.	
tir_est	Pooled	Equal	356	-1.67	0.0948	
tir_diff	Pooled	Equal	0	.	.	

In particular, the two-sample test can be carried on only for *tir_est*, as the other variables are defined only for finished projects (*end=1*).

The mean of the time-increase ratio for finished projects is 1.98: on average it has taken nearly twice as long to complete those projects. The corresponding estimate is a bit higher (2.20), and the t-test shows that the difference of the estimated ratio for finished and unfinished projects is not statistically significant. In other words, we can expect an average doubling in times for the entire programme.

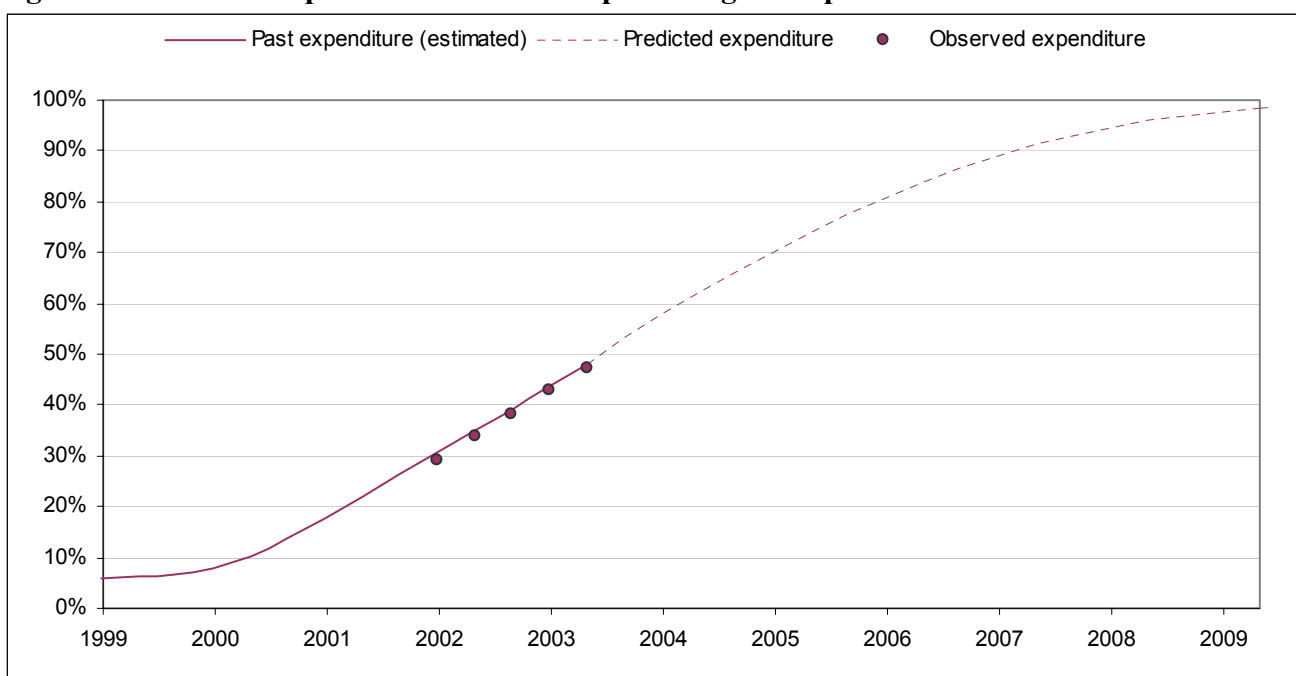
All the above interpretations give insight into the factors that influence the development of the projects, but the most valuable information for policy-makers is given by predictions.

Once we have a set of parameters identifying the objective function, the model can predict the percentage of expenditure for each project at any time. So the model is applied until 2009 and the corresponding project estimates at a given time are all summed together and weighted by the projects' costs.

In Fig.5 the predictions of the model are compared with observed data. Before mid-2002 there is no observed expenditure as the monitoring system was activated in August 2002: the plotted values are those estimated by the model.

The predictions are in very close agreement with observed data. The level of 90 per cent of expenditure will be reached in 2007 and the programme will be ended approximately in 2009.

Fig. 5 – Observed and predicted cumulative percentage of expenditure



The same data can be plotted as a yearly distribution, as shown in Fig. 6. This time, along with the projects' costs, also the quota of the funds allocated by CIPE is specified. This can still be obtained by summing the percentages of expenditure for each project, but weighing by the correspondent CIPE amount.

The most interesting information for policy-makers is the forecast of the yearly amount of CIPE funds that will be spent in the next years. The model so represents an additional tool that supports the allocation of public cash for an investment programme in each year.

Fig. 6 – Observed and predicted cumulative percentage of expenditure



From the above plot it is easy to see that in 2003 the expenditure still lies in the rising half of the bell-shaped profile. The expenditure peaks in 2004 and then starts decreasing quite regularly until 2009.

The agreement of the model with the observed data is stronger on the total expenditure than on the CIPE funds. Nevertheless we must notice that this plot is more subject to uncertainty than the previous one, as the yearly amounts are obtained as differences of the cumulative expenditure profile, that is, the estimated variable.

WHAT THE MODEL DOES NOT DO

It is important to stress that this model is applied only to construction stage and not to the entire development cycle of a project. This is convenient for this particular programme as most projects have already entered this stage. On the contrary, for other investments programmes, the number of started project may be a small percentage of the total. In this case it is essential to use a model that deals with different data, more typical of the earliest development stages, where the expenditure is usually negligible.

Another stage that is kept out of modelling is the operational one, that is, time needed after completion before the project is actually put in operation. Again, there is no major expenditure occurring at this stage and the main variable of interest is time. Moreover, the amount of observed data regarding the operational stage in the existing monitoring systems is small, because that stage has been reached by a relatively small number of projects. Nevertheless it is under study and will be considered in future versions of the model.

A more general approach to modelling the development of infrastructures should include models dealing not only with time increase but also with cost increase. In fact, the model presented here handles cost as a pure project's constant. Nevertheless it must be said that the "Completamenti" projects were mostly funded at an advanced design stage, when cost is generally under control, apart from unexpected events occurring during construction works. If the cost is modified, then the corresponding constant is readjusted accordingly.

CONCLUSIONS

The main results of this model are an insight into the factors that influence the development of infrastructures and a forecast of expenditure for the years ahead.

The proposed model yields estimates of the cumulative percentage of expenditure of a single project over time. The estimates can be transformed into a time distribution, aggregated by year and then yield a forecast of yearly amounts of expenditure.

Hence policy-makers responsible for the assessment of the expenditure for public investments are provided with additional valuable information, implicitly available in existing monitoring systems, but in fact inaccessible until the development of the model.

For the “Completamenti” infrastructure programme, considered in this paper, the model shows that the most important factors that influence the expenditure are project’s cost, region and initial duration. The model confirms that time needed for the completion of expenditure is nearly doubled with respect to initial duration of works, and produces a forecast for the end of the programme in 2009.

The results are obtained under some conditions. The first one is that the start dates of construction works (i.e. when expenditure starts) are known and the second is that costs remain constant throughout the development of the project. The above conditions are generally satisfied for the programme considered in this paper, but in general the situation is much more varied.

Different models are currently under study that will be able to deal with broader programmes.

REFERENCES

Coppi R. (1998), *Lezioni di Analisi Statistica Multivariata*, Roma: Università degli Studi di Roma “La Sapienza”

Der, G. and Everitt, B.S. (2002), *A Handbook of statistical Analyses using SAS, Second Edition*, Boca Raton FL: Chapman & Hall/CRC

Littell, R.C., Milliken, G.A., Stroup W.W. and Wolfinger, R.D. (1996), *SAS[®] System for Mixed Models*, Cary NC: SAS Institute Inc.

SAS Institute Inc. (2000), *SAS/STAT[®] User’s Guide, Version 8*, Cary NC: SAS Institute Inc.


UVER - Unità di Verifica degli Investimenti Pubblici (2004) “Nota informativa per il CIPE sui completamenti – Aggiornamento al 31.12.2003” (Unpublished:

<http://www.dps.tesoro.it/documentazione/docs/UVER/compnota0404.pdf>)

CONTACT INFORMATION

Carlo Amati
Ministero dell’Economia e delle Finanze
Dipartimento per le politiche di sviluppo e
coesione
Unità di verifica degli investimenti pubblici
Via Sicilia 162 - 00187 Roma - Italia
Phone: +39 06 47619801
Fax: +39 06 47619872-3
Email: carlo.amati@tesoro.it

Francisco V. Barbaro
Ministero dell’Economia e delle Finanze
Dipartimento per le politiche di sviluppo e
coesione
Unità di verifica degli investimenti pubblici
Via Sicilia 162 - 00187 Roma - Italia
Phone: +39 06 47619803
Fax: +39 06 47619872-3
Email: francisco.barbaro@tesoro.it



**Dipartimento per le Politiche di Sviluppo
Ministero dell'Economia e delle Finanze
Via Sicilia, 162/c
00187 Roma**

web: www.dps.tesoro.it

mail: comunicazione.dps@tesoro.it