CORE
Metadata, citation and similar papers at core.ac.uk
Provided by PUblication MAnagement

Research Evaluation, volume 12, number 1, April 2003, pages 47–69, Beech Tree Publishing, 10 Watford Close, Guildford, Surrey GU1 2EP, England

Efficiency analysis

A robust nonparametric approach to the analysis of scientific productivity

Andrea Bonaccorsi and Cinzia Daraio

Data on scientific productivity at institutes of the French INSERM and at biomedical research institutes of the Italian CNR for 1997 were analysed. Available data on human capital input and geographical agglomeration allowed the estimation and comparison of efficiency measures. Nonparametric envelopment techniques were used, and robust nonparametric techniques were applied in this work for the first time for evaluating scientific productivity. They are shown to be useful tools to compute scientific productivity indicators and make institutional comparative analyses. Taking into account a large number of methodological problems, a meaningful and rigorous indirect comparison is made possible. Several possible explanations of the observed differences in productivity are commented on.

Andrea Bonaccorsi and Cinzia Daraio are at the Sant'Anna School of Advanced Studies, Piazza Martiri della Libertà, 33, 56127 Pisa, Italy; email: bonaccorsi@sssup.it; cinzia@sssup.it

THE NOTION OF EFFICIENCY is highly problematic in the analysis of scientific research. While policymakers and scientists are ready to accept that research activity should be organised in such a way as to avoid inefficiencies and waste of resources, the exact definition of what accounts for efficiency is far from being accepted. Several theoretical and methodological problems are still unsolved.

The object of this study is to give an overview of efficiency analysis applied to scientific research. Based on very recent results in econometrics, we propose a methodology for using nonparametric and robust nonparametric approaches in the evaluation of the productivity of scientific research.

To illustrate the practical implications of the new techniques, in this paper we analyse data on scientific productivity at (almost) all institutes of the French Institut National de la Santé et de la Recherche Médicale (INSERM) and at biomedical research institutes of the Italian Consiglio Nazionale delle Ricerche (National Research Council, CNR) for 1997. Available data on human capital input and geographical agglomeration allow the estimation of efficiency measures for the two institutions. Taking into account a large number of methodological problems, a meaningful and rigorous indirect comparison is made possible. Since the data open a certain number of interpretative problems, we comment on several possible explanations of the observed difference in productivity.

We discuss several general methodological problems of efficiency analysis in science, present the alternative techniques available and discuss their limitations. We introduce the new results in robust nonparametric analysis and examine the potential of these techniques for scientometrics. Following sections presents the data on INSERM and CNR, apply and compare several productivity analysis techniques, offer a preliminary interpretation of observed differences, and conclude.

Methodological problems

Any notion of efficiency relates a vector of inputs to a vector of outputs. Unfortunately, in scientific research all three definitional elements of efficiency — inputs, outputs and the functional relation between the two — are affected by severe conceptual and measurement problems.

Scientific production is a multi-input, multi-output relation, in which, differently from standard production activity, both inputs and outputs are qualitatively heterogeneous and sometimes incommensurable, the relation is dynamic and not deterministic and the output is lagged but with a non-fixed structure. These features create formidable conceptual and measurement problems.

Ideally, at the level of inputs, we should include:

- the number of researchers differentiated by age, level of qualification or seniority (e.g. the number of years they have been working as researchers), level of quality (e.g. the cumulative number of publications or citations received);
- physical capital;
- research funds;
- cumulative stock of knowledge (i.e. the number and quality of publications in the past); and
- agglomeration factors.

In practice, it is extremely difficult to collect data on all types of inputs. In most cases very crude data on the number of researchers and on research funds are the only available evidence. As a matter of fact, most analyses do not really include all inputs.

At the level of outputs, most analyses work with count data (i.e. number of publications), although it is clear that the quantity of papers does not have a necessary relation with their quality (as measured by normalised received citations) or importance.

Furthermore, it should be recognised that the outputs of a scientific institute are not limited to publications but include teaching, training, patents, applied research for industry and other parties, services for the public administration, consulting and the like. For this reason, efficiency analyses limited to publication data are still considered with skepticism. Even though bibliometrics data are widely accepted in the evaluation of research productivity (Daniel and Fisch, 1990; Ramsden 1994; Narin and Hamilton, 1996; Van Raan, 1993, 1997), they are viewed with suspicion by some of those being evaluated (Collins, 1991). It is desirable not only that they cover many different aspects of research outputs (Martin, 1996) but also that the evaluees have a place in helping to create appropriate methodology by identifying the relevant categories of output (Lewison, 1998).

Again, in practice the collection of data on all these outputs is extremely difficult, unless with field surveys on a limited scale (see e.g. Bordons and Zulueta, 1997; Lewison and Dawson, 1998). For most large-scale analyses, simple publication data are considered acceptable.

In the evaluation of productivity, the definition of what accounts for *inputs* or *outputs* of scientific research is one of the most crucial points. From a substantive perspective all factors can be considered both as input and as output. There are no definitive answers to this problem. They have to be defined case by case, so that any factor can be considered as input or as output, taking into account the purpose of the analysis. The methodology we apply in this work takes the definition of inputs and outputs as given.

At the level of the functional relation there are also several problems. One of the most fundamental problem is *endogeneity*: the level of inputs in terms of funds and number of researchers are a function of past level of output, so that any specification that does not take these effects into consideration is likely to produce misleading results.

Another important problem is that scientific production does not follow the assumptions needed to adopt the *production function* approach.[1] Consequently, the toolbox of partial and total factor productivity has limited value. As a matter of fact, most published studies adopt a production function approach[2] even though its conceptual foundation in science is extremely weak. In order to overcome these limits, a large literature on nonparametric efficiency analysis has been developed, but this suffers from other methodological problems, which we will discuss later in the paper.

Another problem is in the *dynamic* relation between inputs and outputs. While in most productive processes the time sequence that relates the use of productive resources to the outcome is fixed and predictable, in science the outcome of research follows from inputs with a time-lag structure that is both unknown and variable over time.

In addition to definitional and specification problems, there are also measurement problems. From a measurement point of view, there are difficult problems in the field of scientific research, particularly in *indirect comparisons*. Although standardised international procedures exist for the definition and measurement of research inputs (e.g. full time equivalent), very often scientific institutions do not follow these procedures strictly and differ in the meaning they attach to the collaboration of scientists to activities. Consequently in indirect comparisons much care should be placed in establishing comparisons, as we shall see later in this paper.

In this paper we develop a methodological exercise using recently developed techniques that overcome most limitations of nonparametric tools and might be extremely powerful should we have access to all desirable data on inputs and outputs. We also develop a simple numerical example by comparing two large public research institutions, for which we had access to comparable data on a (small) subset of input and output data.

We recognise that available data are poor, so that it is difficult to draw any substantive implication on productivity. In particular, we examine data on just one year (1997), have just one type of output (publication count) and few inputs (number of researchers, geographical agglomeration), so that we are in very far from an ideal situation. Nevertheless, the data allow for an interesting comparison of potential and limits of several alternative productivity measures, so can be used as a simulation exercise for the techniques. The data show interesting patterns, which open the way to further investigation. Further research is currently underway to build up a more comprehensive dataset that might support a substantive interpretation of observed differences in productivity.

Alternative approaches

There are several possible methods for comparing efficiency measurement in science. This section offers an overview of these methods and discusses their limitations, with a view to developing an integrated methodology.

Ratio measures

A very simple approach starts with a crude comparison of *simple measures of productivity* (i.e. output/input ratios). This approach takes one type of input and relates it to one type of output, ignoring all relations of complementarity and substitution between inputs, and all effects of joint production in outputs. To anticipate the numerical exercise presented later in the paper, this approach would take the number of international publications of a research institute and relate this number to the number of researchers in the institute. Alternatively, the number of international publications can be related to the total number of employees of the institute.

The exact definition of inputs and outputs is critical to this approach. As an example, consider the data from INSERM and CNR. Assuming that output data are taken with the same standardised international methodology, they can be considered strictly comparable. By contrast, input data create a host of problems. In particular, the definition of what accounts for a research input is theoretically clear (e.g. in OECD Frascati and Oslo manuals), but in practice very difficult to respect at the micro-level.

Taking into consideration only one pair of variables at a time, ratio measures give a partial picture. They serve mainly as a first order approximation. As an example, the comparison based on crude ratios shows a striking difference between the two systems. Considering only international publications, on average each CNR researcher produces 4.61 papers per year while researchers working at INSERM produce 1.36 papers. The only way to reconcile the two figures is by considering only INSERM researchers in the computation, excluding researchers from hospitals and universities and from other research organisations. In this case the average productivity index rises at 5.06. This equates, however, to claiming that all other researchers declared by INSERM to be part of their institutes have a productivity of zero. Thus we are left with the uneasy situation according to which each of the two public research systems can be claimed to be ‘more efficient’ depending on the particular definition of input adopted.

Nonparametric indicators

The simple measures of productivity computed as ratios of output-to-input are sometimes referred as *partial* productivity measures. This terminology distinguishes them from *total* factor productivity measures because the latter try to obtain an output-to-input ratio value that takes into account *all* outputs and inputs. Moving from partial to total factor productivity measures by combining all inputs and all outputs to obtain a single ratio helps to avoid imputing gains to one factor (or one output) that should be attributed to some other input (or output). However, total factor productivity measures encounter difficulties, such as choosing the inputs and outputs to be considered and the weights to be used in order to obtain a single-output-to-single-input ratio.

Efficiency measures are generated by comparing each institute to the most efficient ones in its own comparison set. The most efficient institutes are those that minimise the use of inputs given a level of observable outputs (input-oriented), or maximise outputs given a level of observable inputs (output-oriented). Data envelopment analysis (DEA) does not require the user to prescribe weights to be attached to each input and output, as in the usual index number approaches, nor does it require prescribing the functional forms that are needed in regression approaches.

In the exercise presented later in the paper, in order to estimate a frontier (piece-wise-surface) over the data we calculated an input-oriented DEA with the assumption of variable returns to scale (VRS). In this formulation, which follows the model by Banker, Charnes and Cooper (1984), we can compute both the technical efficiency (TE) and the scale efficiency (SE).[3] The TE is a measure of the radial distance of an institute to the estimated efficient frontier. If TE is equal to 1 then the research institute is located on the efficient frontier. If TE is less than 1, its value represents the proportionate reduction of inputs (given the value of outputs) the institute should put in place, in order to be fully efficient. The SE can be roughly interpreted as the ratio of the average product of a research unit to the average

Efficiency measures are relative, implying that they cannot be compared directly in terms of absolute values. This means that each institute is compared to the most efficient ones in its group

product of a research unit operating at a point of technically and optimal scale. Again, if it is 1 the research institute is scale efficient, if it is less than 1 the institute is scale inefficient.

Robust nonparametric scientific indicators

The use of DEA for comparative analysis is subject to a fundamental limitation. Efficiency measures are relative, implying that they cannot be compared directly in terms of absolute values. This means that each institute is compared to the most efficient ones in its group.

Suppose there is a distribution of efficiency values in all possible groups; that is, it is possible to observe the universe of units. Then the probability that a comparison is made with the best units in the population increases with the size of the sample. This means that comparing samples of unequal size may be misleading; it may happen that the larger sample includes better units, so that the comparison becomes unfavourable to the rest of the sample. In other terms, data envelopment analysis is sensitive to extreme values and outliers; it is non-independent of the observed distribution of values.

This problem is solved by using the recently developed robust nonparametric estimation technique. The basic idea is that the benchmark is not made with the most efficient units in the group, but with an appropriate measure drawn from a large number of random samples of size *m* within the group. In this way size-dependent effects are eliminated.

The robust nonparametric approach on which we based the computation of the scientific indicators was introduced by Cazals, Florens and Simar (2002). The methodology to introduce environmental variables in this robust approach was developed by Daraio (2002), in whose work the computational aspects and methodological steps of the analysis have been implemented.

Since in the analysis of scientific production the underlying distributions of efficiency are highly asymmetric and size effects are very important, these techniques should find large application. Because of the interest generated by these newly developed techniques for efficiency analysis, we now undertake a more detailed review of their methodological foundations.

New techniques in efficiency analysis

Origins, and the nonparametric approach

The purpose of efficiency analysis is to make a relative benchmark or comparison among decision-making units (DMUs). Let us assume for the rest of this paper that a DMU represents a research institute. Each DMU is compared to the best performer included in the analysis. The comparison is therefore made on the basis of the real or observed performance of units, and not the theoretical maximum.

Efficiency analysis has been developed from the first empirical work of M J Farrell (1957) who built upon the work of Debreu (1951) and Koopmans (1951) to define a simple measure of firm efficiency that could account for multiple inputs and multiple outputs: ‘When one talks about the efficiency of a firm one usually means its success in producing as large as possible an output from a given set of inputs’ (Farrell, 1957, page 254). Farrell proposed that the efficiency of a firm consists of two components: *technical efficiency*, which reflects its ability to obtain maximal output from a given set of inputs; and *price* (or *allocative*) *efficiency*, which reflects the ability of a firm to use the inputs in optimal proportions, given their respective prices and the production technology.

He then suggested the use of:

- a nonparametric piece-wise-linear convex isoquant, constructed such that no observed point lies to the left or below it; and
- a parametric function fitted through the data, such that no observed point lies to the left or below it.

The first suggestion was taken up by Charnes *et al* (1978), resulting in the development of the *data envelopment analysis* (DEA) approach. DEA involves the use of linear programming methods to construct a nonparametric piece-wise surface (or frontier) over the data. Efficiency measures are then calculated relative to this surface.[4] From this original formulation an impressive literature developed, with a number of extensions and refinements. DEA encompasses a variety of models for evaluating performance.[5]

A large literature has applied data envelopment analysis to problems of productivity in a large number of manufacturing and service settings. Several studies have used DEA-type approaches in assessing the technical efficiency of academic research; for example, Coelli (1996); Korhonen, Tainio and Wallenius (2001); Thursby and Kemp (2002). Studies applying DEA to education include Bessent and Bessent (1980); Bessent *et al* (1982); Charnes *et al* (1978); Fare, Grosskopf and Weber (1989); Grosskopf *et al* (1999); Grosskopf and Moutray (2001). Rousseau and Rousseau (1997, 1998) applied DEA to construct scientometrics indicators and assess research productivity across countries.

The parametric approach

As we have seen, Farrell (1957) suggested two alternative approaches, nonparametric and parametric. The parametric approach for Farrell's efficiency measures was taken up by Aigner and Chu (1968), who developed the deterministic frontier model approach based on the estimation of a parametric frontier production function of Cobb Douglas form. Models in this family are called deterministic because in the frontier model, the observed output, is bounded above by a nonstochastic — deterministic — quantity.

One of the main criticism of the deterministic frontier model is that no account is taken of the possible influence of measurement errors and other noise upon the frontier. All deviations from the frontier are assumed to be the result of technical inefficiency. Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977) independently proposed the stochastic frontier production function, in which an additional random error, is added to the nonnegative random variable which represents inefficiency.[6]

A comparison of methods[7]

The great virtue of stochastic production frontier models is that the impact on output of exogenous shocks can at least in principle be separated from the contribution of variation in technical efficiency. The stochastic frontier model is not, however, without problems.

The main criticisms are the following:

- the need to specify a functional form for the production function,
- the need to specify a distributional form for the inefficiency term;
- there is generally no *a priori* justification for the selection of any particular distributional form for the inefficiency;[8]
- it is more difficult to include multiple outputs.

Some of the advantages of the parametric approach over DEA are that it accounts for noise, and it can be used to conduct conventional tests of hypotheses. On the other hand, the choice of the nonparametric approach may be made because of the very few assumptions required and mainly because we do not have to specify the functional form of the relation inputs/outputs.

Most data envelopment analysis models are invariant with respect to the units of measurement and they may focus on either input reduction or output augmentation to achieve efficiency (input or output orientation). For a particular model, it is possible to incorporate categorical variables and/or nondiscretionary inputs or outputs and it is also possible further to constrain the multipliers. Both techniques can be applied on cross-section or panel data.

The main theoretical and practical problems of the DEA conducted in the traditional perspective are as follows:

- results can be biased by the exclusion of an important input or output;
- there is an influence of noise and measurement error on the shape and position of the frontier;
- the treatment of the inputs/outputs as homogeneous commodities (when they are heterogeneous) may distort the results;
- the test of hypothesis is more difficult in this context;
- there is an influence of outliers on the results;
- not allowing for environmental differences may give misleading indications of relative managerial competence.

Recent theoretical developments in efficiency analysis give the opportunity to overcome the traditional theoretical problems of data envelopment analysis.

Many have claimed that the main drawback of the DEA technique[9] is its deterministic nature, related to the mathematical programming on which computations are based. Another disadvantage of the DEA method is the lack of statistical tests procedures, available for the parametric frontier models.

On the first point, statistical inference based on nonparametric estimators is now possible.[10] This is done with an integrated approach in which the statistical model allows the determination of the statistical properties of the nonparametric estimators in the multi-output and multi-input case. Sampling distributions may be approximated by bootstrap distributions in very general situations.[11] These techniques allow correction for the bias of the efficiency estimators and estimation of confidence intervals for the efficiency measures. An application of these methods has been done in the illustrative exercise, and some comments on their usefulness are provided in the following. Finally, in the treatment of input/output heterogeneity, *normalization* methods can be useful.[12]

On the side of test of hypotheses, tests for whether inputs or outputs are irrelevant, as well as tests of whether inputs or outputs may be aggregated, have been formulated in the context of nonparametric models of technical inefficiency.[13] Because of data constraints we have not applied these tests, that in any case could be useful for researchers who have a lot of data and want to choose the relevant input/output to be introduced in the analysis. On the overcoming of the influence of outliers, see next section.

The robust nonparametric approach

In doing efficiency analysis, the interest lies in estimating the efficient level of DMUs considering the frontier of the possibility set, which is the set of attainable combinations of inputs (x) with outputs (y).

For a specific level of outputs, the efficient level of inputs delimits the frontier enveloping all the possible values. The efficiency of DMUs will then be defined as the *distance* between the observed value of DMU variables and the frontier.

The set Ψ of possibilities is the set of attainable points (x, y). It is defined:

$$\Psi=\{(x,y)\in R_+^{p+q} | (x,y) are\, attainable\} \tag{1}$$

where $x\in R_+^p$ is the vector of inputs and $y\in R_+^q$ is the vector of outputs. For all possible values of *y*, the section of possible value of *x* is the set defined as:

$$X(y)=\{x\in R_+^p | (x,y)\in\Psi\} \tag{2}$$

and its efficient boundary is the subset of $X(y)$ defined by:

$$\partial X(y)=\{x \mid x\in X(y), \theta x\notin \ni X(y), for\, all\ 0<\theta<1\}. \tag{3}$$

A measure of the efficiency of a particular DMU $(x_k y_k)$ can then be expressed by

$$\theta_k=\min\{\theta: \theta x_k\in X(y_k)\}. \tag{4}$$

It is the radial distance from x_k to the efficient boundary $\partial X(y_k)$. If $\theta_k=1$, the research institute (x_k, y_k) is considered as being 'efficient' in the sense that it achieves the minimal attainable level of x_k given y_k. The efficiency score $\theta_k<1$ represents the feasible proportionate reduction of x_k given the value of y_k in order to be considered as being efficient.

In efficiency analysis, the nonparametric approach is based on envelopment techniques, whose main estimators are DEA and free disposal hull (FDH).[14] These estimators rely on the idea that the attainable set is defined by the set of minimum volume containing all the observations. The DEA estimator relies on convexity of the set Ψ, whereas the FDH estimator does not impose such a restriction.

The DEA estimator of Ψ based on a sample of *n* observations (x_i, y_i), noted as $\hat{\Psi}_{DEA}$, is defined as follows:

$$\hat{\Psi}_{DEA}=\{(x,y)\in R_+^{p+q} \mid y\leq\sum_{i=1}^{n}\gamma_i x_i; \quad x\geq\sum_{i=1}^{n}\gamma_i x_i,\ for\ (\gamma_1,\ldots,\gamma_n) \quad s.t.\sum_{i=1}^{n}\gamma_i=1; \gamma_i\geq 0, i=1,\ldots,n\}. \tag{5}$$

The FDH estimator of Ψ based on a sample of *n* observations (x_i, y_i), noted as $\hat{\Psi}_{FDH}$, is the free disposal closure of the reference set $\{(x_i,y_i) | i=1,\ldots,n\}$. It can be defined as:

$$\hat{\Psi}_{FDH}=\{(x,y) | x\geq x_i, y\leq y_i, i=1,\ldots,n\} \tag{6}$$

The estimated DEA efficiency score of a particular research institute (x_k, y_k), noted as $\hat{\theta}_{DEA}$, is given by:

$$\hat{\theta}_{DEA}=\min_k\{\theta | (x_k,y_k)\in\hat{\Psi}_{DEA}\} \tag{7}$$

The estimated FDH efficiency score of a particular research institute (x_k,y_k), noted as $\hat{\theta}_{FDH}$, is given by:

$$\hat{\theta}_{FDH}=\min\{\theta | (\theta x_k,y_k)\in\Psi_{FDH}\} \tag{8}$$

By construction, both $\theta_{DEA}(x_i,y_i)$ and $\hat{\theta}_{FDH}(x_i,y_i)$ are ≤ 1 for all observed research institutes (*x*, *y*); a research institute is efficient when the efficiency score is equal to one.

As said above, one of the main drawbacks of nonparametric estimators (DEA/FDH) is their sensitivity to extreme values and outliers. In this framework, Cazals, Florens and Simar (2002, from now on CFS), propose a nonparametric estimator of the frontier, more robust to extreme values and outliers. It is based on the concept of the *expected minimum input* function of order-*m*. The efficient frontier of Ψ, described above, can be expressed in another way using the distribution function and the survival function of (x, y) defined as follows. In the input space, the lower level of input *X* attainable for a DMU producing at least a given level of output *y* can be characterized through the *conditional survivor function* of *X*, given that $Y\geq y$, defined as follows:

$$S_c(x|y)=\text{Prob}(X\geq x|Y\geq y)=\frac{S(x,y)}{S_Y(y)} \tag{9}$$

where $S(x, y)$ is the *conjoint survivor function,* defined as $S(x,y)=\text{Prob}(X\geq x, Y\geq y)$, and $S_Y(y)$ is the *marginal survival function* of *Y*, defined as $S_Y(y)=\text{Prob}(Y\geq y)$.

The lower boundary of the conditional survivor function can be defined for any value of *y* as:

$$\varphi(y)=\inf\{x | S_c(x|y)<1\}. \tag{10}$$

If the possibility set Ψ is *free disposal*,[15] then $\varphi(y)=\partial X(y)$ and we have just reformulated the estimation of the frontier of Ψ in the input space.

A natural nonparametric estimator of $\varphi(y)$ is given by plugging in; that is, by substituting in equation (10), the empirical version of $S_c(x|y)$ estimated over the sample. As pointed out in CFS, the obtained estimator is the FDH estimator, the boundary of the free disposal hull of the observed data points.

Now we can introduce the *order-m frontier*, a more flexible concept of frontier, which by construction does not envelop all the observed points and so is more robust to extreme points.

Consider $X^1, X^2, ..., X^m$ as the inputs drawn from a sample of size *m* of research institutes having a level of output $Y \geq y$. The order-*m* frontier is defined as:

$$\varphi_m(y) = E\left[\min\left(X^1, X^2, ..., X^m\right) \middle| Y \geq y\right] = \int_0^\infty \left[S_c(x|y)\right]^m dx \qquad (11)$$

It represents the expected minimal value of the inputs among a fixed number of *m* research institutes drawn from the population of research units with a level of output $Y \geq y$. For all the proofs and more details see CFS.

A nonparametric estimator of $\varphi_m(y)$ is given by plugging in (i.e. substituting) the empirical conditional survival function of *X*, computed on the sample, in equation (11). An algorithm is proposed in CFS to compute it in practice.

In Daraio (2002) the introduction of environmental variables (variables that affect the efficiency of research institutes, but that cannot be considered as inputs or as outputs) in this context is extensively treated, both from an analytical point of view and from the Monte-Carlo approximation algorithm side. The basic idea is that of introducing environmental variables, *z*, in the algorithm proposed by CFS using a nearest neighbours (NN) method (for a description of the method see e.g. Silverman, 1986).[16]

In this way, we can compute for each research institute the robust level of efficiency that has a clear economic interpretability: if it is smaller (greater) than 1, then the research institute is more inefficient (more efficient) with respect to the expected value of the minimum input of *m* research units drawn from the population of institutes with a level of output greater or equal to its value of output. If the value of robust efficiency is equal to 1 then the research institute is efficient with respect to the expected value of the minimum input of *m* research units drawn from the population of institutes with a level of output greater or equal to its value of output.

As pointed out in Daraio (2002), the parameter *m* plays a central role in the robust indicator computation: it has a *dual* nature. It is defined as a 'trimming' parameter for the robust nonparametric estimation. It defines also the level of *benchmark* we want to carry out over the population of units. The parameter *m* can be used in its dual meaning to provide both robust estimation and a *potential scenarios* analysis.

The first task can be accomplished by plotting the percentage of points outside order-*m* frontier as a function of *m*. By inspecting this graph we may choose the value of *m* that corresponds to the target (or desired) degree of robustness. The second application of *m* concerns the evaluation of a *potential scenario*, in which for each institute the efficiency score is computed for different values of potential *competitors* (*m*), and the evolution of such score is analysed by progressively increasing *m*. Furthermore, we can compute the robust level of efficiency for each research institute, *conditioned* to its level of environmental variables and compare it with the unconditioned robust nonparametric level of efficiency.[17]

Ideally, the input/output relation should include all relevant production factors; it is desirable to include in the estimation proxies for physical capital and for intermediate inputs

Based on this approach, we introduce two new measures, labelled s*cientific productivity index of order m* SPI(m), and *scientific productivity index of order m conditioned to the influence of environmental variables* SPI(m,g). The construction of the two indices is based on a procedure similar to that of DEA method. We have to define several inputs and outputs and then we have to benchmark each DMU with the frontier constructed, starting from the observations available. Here, the frontier is more realistic and does not envelop all data points (i.e. is more robust to extreme values and outliers). Moreover, we can evaluate the effects of environmental variables on the productivity comparison.

Ideally, the input/output relation should include all relevant production factors. Although scientific research is fundamentally generated by human capital factors, it is desirable to include in the estimation proxies for physical capital (scientific instrumentation, capital equipment) and for intermediate inputs (materials). Data on research funds are a reasonable approximation if we know something about the share of funds allocated to investment. The introduction of physical capital and intermediate inputs would allow the estimation of complementarity effects.

In this paper we also apply this methodology to assess the conditional effects of geographical agglomeration on research productivity in a comparative analysis, and we find interesting different pattern between the French INSERM and the Italian CNR. To the best of our knowledge, this paper is the first empirical application of the robust nonparametric approach by CFS in the evaluation of scientific productivity.

Data sources and descriptive statistics

Limitations of data

We provide data on input and output of biomedical scientific research for 213 INSERM institutes in France and 27 CNR institutes in Italy for 1997. Before describing data, a warning is needed. There are strong limitations in data, which have to be taken into account.

First of all, data refer to just one year. In the case of INSERM, available data refer to just one year and there is no way to improve the information, unless a specific research project is undergone using individual publication data. In the case of CNR, we use data for 1997 for reasons of comparability. In the literature on bibliometrics and the economics of science it is well known that data on scientific publications should be averaged over some years, in order to take into account the inherent variability of the phenomenon over time.

Second, we take as a definition of scientific production the *number* of total and international publications. We have no data on individual publication nor we can control for quotations of publications. In further research we will build *individual* career patterns, using bibliometrics indicators, for both CNR and INSERM scientists, but this will require lengthy work.

Third, we assume as valid the *self-declaration* of both institutions in terms of total and international publications. We checked against official documents and controlled for the definitions adopted, but could not have access to original files that gave origin to the self-declaration.

In terms of comparability the following problems emerge. On the French side we have the whole INSERM system, which is to say, a large part of the biomedical research system. On the Italian side we have all CNR biomedical institutes, which however are only a small fraction of CNR institutes, which themselves are a small part of the entire research system. This means that our exercise is *not* in any meaningful sense a comparison between two national systems. On the French part we should include the prestigious research activity of CNRS and several universities, while on the Italian side the role of universities should be included. Let us stress again that our analysis does not have any general implication on the analysis of national systems. It is rather a methodological exercise, telling us what kind of analyses we could carry out should we have access to more complete and strictly comparable data.

In the French case, the definition of INSERM personnel includes not only direct employees, but also researchers from university and hospitals that are allocated temporarily to institutes. It is possible that the actual work time of these researchers is not fully allocated to INSERM, resulting in a slight deterioration of productivity indexes. At the same time, qualitative observation of the French system has also highlighted the opposite phenomenon; that is, in some cases extremely productive university professors collaborate with INSERM institutes and increase significantly its publication score.

Also, while for CNR we can distinguish between total publications and international publications, this is not possible for INSERM. We have to rely on the self-declared definition of the institute, that total publications corresponds to international standards. By comparing only international publications for CNR and total publications for INSERM we believe we avoid large errors. However, should the INSERM counting include some non-international publications, this would result in an overestimation of its productivity.

These measurement problems will be discussed at length later in the paper.

The INSERM database

The French Institut National de la Santé et de la Recherche Médicale (INSERM) is a very large public research organisation, having 256 units, 61 teams and nine common facilities. More than 10,000 people work in INSERM facilities.

The INSERM database collects data on the number of researchers and publications of the INSERM institutes in 1997. The sample is based on 213 observations, which is a large part of the universe of institutes. We were able to access data on institutes by visiting websites systematically and by addressing a mail survey to directors. Although data refer to one year only, they offer a comprehensive view of the activity of a large part of the French biomedical research system.

The number of researchers is divided in three categories (INSERM researchers, researchers from hospital and university, other researchers), in addition post-doctoral students (*boursiers*) and technical-administrative personnel are included. For all institutes we define a geographical classification. For a subsample of 65 institutes we have information on the number and size of research teams. We also classify institutes by research area (see Table 1). On the basis of available data we can construct the simple descriptive statistics on INSERM institutes shown in Table 2.

The average institute is formed by 36 units of personnel, with a typical composition of 17 researchers, 11 technical and administrative staff, and eight *boursiers*. The largest institute has 147 units, the smallest one 13. The analysis of the composition of personnel shows several peculiar elements:

Table 1. Classification of INSERM institutes by research area

Research area	Number of institutes
Cat. 1: Other	5
Cat. 2: Mol. biology/Genetics	42
Cat. 3: Endocrinology	13
Cat. 4: Epidemiology	13
Cat. 5: Pharmacol./Biochemical	49
Cat. 6: Physio./Pathology	36
Cat. 7: Immunology/Cancer	55
Total	213

Table 2. Descriptive statistics of INSERM institutes

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Mean	5.16	7.11	6.37	10.38	7.90	36.32	22.12	17.14	0.61	5.06	5.68	1.36	2.35	4.73	77.68
Median	4.50	6.00	6.00	9.00	7.00	33.00	19.00	16.00	0.55	3.67	3.60	1.18	1.96	2.60	121.00
Max	24.00	25.00	31.0	54.00	49.00	147.0	91.00	45.00	2.17	36.00	36.00	6.00	14.50	81.00	141.00
Min	1.00	1.00	1.00	2.00	1.00	13.00	1.00	5.00	0.03	0.33	0.10	0.06	0.11	0.14	1.00
Std. Dev.	2.79	4.66	4.77	6.04	5.78	16.41	15.14	8.05	0.36	4.31	6.17	0.90	1.74	8.48	59.31
Skewness	2.32	0.90	1.74	2.74	2.78	2.32	1.41	0.82	1.46	2.66	2.73	2.13	2.92	6.52	-0.09
Kurtosis	10.40	0.67	5.04	13.40	14.27	10.27	2.75	0.27	3.03	12.84	8.64	7.21	14.33	50.27	-1.92
No. Obs.	212	184	196	213	205	213	213	213	213	212	196	213	213	205	213

Key: A: INSERM_RES: INSERM researchers
B: OTHER_RES: other researchers
C: HU_RES: hospital/university researchers
D: ITA: technical and administrative personnel
E: BORS: doc. and post-doc. students or scholarship holders (boursiers)
F: T_PERS: total number of personnel
G: INTPUB: total number of publications in year 1997
H: T_RES: total number of researchers
I: INTPUB/T_PERS: publication per capita
J: INTPUB /INSERM_RES: publication per INSERM researcher
K: INTPUB/HU_RES: publication per university and hospital researcher
L: INTPUB /T_RES: publication per researcher
M: INTPUB /ITA: publication per technical and administrative unit of personnel
N: INTPUB /BORS: publication per boursier
O: GAI: Geographic Agglomeration Index

- a relatively large number of doctoral and post-doctoral students or *boursiers*;
- a composition of researchers in which INSERM employees are complemented by researchers from hospitals and universities, and from other research organisations;
- a relatively small number of administrative and technical staff.

Coming to productivity indicators, each researcher (of all types) publishes slightly more than one paper per year (1.36), on average. There is a large variance around this value, with some institutes exhibiting a remarkably high average value (6.0 publications per researcher). If productivity is computed taking into account only INSERM researchers, the value is much higher (5.06 papers per unit). Finally, in terms of publications per unit of personnel, the average institute exhibits an average value of 0.61.

The index GAI is a measure of agglomeration. To each institute we assigned one point for each other CNR or INSERM institute located in the same city that is not of the same research aggregation; and two points for each other institute located in the same city that is also of the same research aggregation of the institute considered. The average institute has a value of 77.68, meaning that it is located in a region in which there are other 78 INSERM institutes across all research areas or other 39 institutes of the same research areas, on the average, or any combination between the two. This is an extremely high value. As a matter of fact, 155 institutes out of 213 are located in three regions (namely, Ile de France, Provence and Rhones Alpes). The French system is highly concentrated from a geographic point of view.

The CNR database

Founded in 1923, the Italian National Research Council (Consiglio Nazionale delle Ricerche, CNR) is the most important Italian national research institution, spanning many scientific and technological areas. The analysis in this paper is based on a larger study of the Italian CNR (Bonaccorsi and Daraio, 2002). CNR covers almost all scientific areas and does not have a specialised organisation for biomedical research. In this analysis we include 27 institutes in the so-called MA3 class, covering biomedical molecular biology, medicine and biology.

We built up an original database by manually integrating data on publications for the year 1997 (from an official report) with data on personnel drawn from administrative files. By combining available data and taking into account comparability with INSERM we are able to define the list of variables shown in Table 3.[18]

A representative (average) institute has 30 units, of which 14 are researchers, 14 are technicians and two are administrative staff. The largest institute has 112 units of personnel, the smallest is as small as a couple of people. The size of CNR institutes is slightly smaller than the size of INSERM institutes.

The geographical agglomeration index makes clear a much more scattered situation. The average GAI is 16.33, implying that the average institute has a small number of similar institutes in the same region. Clearly the index does reflect the size of sample, so it is not directly comparable to the index for INSERM. Nevertheless, it is clear that Italian institutes operate in a situation of more pronounced geographical dispersion.

Table 3. Descriptive statistics of CNR institutes in the biomedical area

	A	B	C	D	E	F	G	H	I	J
Mean	2.19	13.78	9.00	29.89	13.93	77.41	53.00	4.61	2.35	16.33
Max	12.00	65.00	28.00	112.00	43.00	382.00	209.42	0.67	0.46	39.00
Min	0.00	1.00	1.00	2.00	1.00	4.00	2.00	18.98	9.49	1.00
Std dev.	2.84	15.66	6.79	28.09	11.27	83.27	47.48	3.48	1.87	12.54
Skewness	2.21	2.23	1.19	1.71	1.08	2.56	1.90	2.81	2.42	0.83
Kurtosis	5.32	5.22	1.03	2.79	0.36	7.25	4.14	11.01	7.60	-.29
No. obs.	27	27	27	27	27	27	27	27	27	27

Key:
A: ADM: number of administrative staff
B: TECH: number of technicians
C: ORD_RES: number of researchers
D: T_PERS: total number of personnel
E: T_RES: total number of researchers
F: T_PUB: total number of publications
G: INTPUB: number of international publications
H: IPURES: number of international publications per researcher
I: IPUPERS: number of international publications per capita
J: GAI: geographic agglomeration index

Possible sources of distortion

Before entering into a detailed comparison between the two organisations, it is useful to eliminate possible sources of distortion. In particular, INSERM institutes cover all areas of biomedical research and are subject to higher heterogeneity. We are interested in examining whether some of the average values observed in the aggregate are mere composition effects, deriving from particular classes of biomedical research. We therefore subdivided the entire INSERM database into separate classes by research area and tabulated all variables accordingly.

Inspection of INSERM data shows that differences across research areas (see Table 1) do exist, but have a small magnitude. In particular, it seems that productivity indicators do not vary systematically across areas. In order to test this effect more rigorously, we performed a Kruskall-Wallis test, taking into consideration non-normality in the distribution of variables. From the results of this test,[19] we are led to accept the assumption of no difference between average values across categories for all productivity indicators, although the assumption must be rejected for size indicators (total number of researchers and total number of publications in 1997). In other words, institutes are significantly different across areas in their average *size*, but not in their average *productivity*. Therefore the variability across institutes in productivity is not explained by heterogeneity in research areas, but must have some other explanation.

An illustration based on biomedical research

Data envelopment analysis

We ran a DEA for each biomedical system: the Italian CNR and the French INSERM. For each institution we obtained the ranking of technical efficiency (TE) and scale efficiency (SE) computed for all institutes. On the basis of the ranking of institutes, descriptive statistics were obtained for the two institutions.

We used the following variables: as inputs we considered the total number of researchers (T_RES), and the geographical agglomeration index (GAI); as output we considered the number of international publications (INTPUB).[20] We also controlled results with a different specification of inputs, using the total number of researchers (T_RES) and technical and administrative personnel (ITA).[21] For the INSERM case, due to a possible measurement error in inputs we also run the estimation of the frontier using as input the variable INSERM_RES instead of T_RES.

The use of GAI as an input requires some comment. A large literature on the geographic dimension of knowledge spillover (see e.g. Zucker, Darby and Armstrong, 1998; Katz, 1994; Audretsch and Feldman, 1996) has stressed the impact of proximity and social interaction on knowledge flows across individuals and organisations.[22] From a policymaking point of view, decisions on location of research activities are often made on an assumption of external or agglomeration economies; that is, institutes geographically close to each other are more produc-

From a policymaking point of view, decisions on location of research activities are often made on an assumption of external or agglomeration economies

Table 4. DEA efficiency scores

Institution	Indicator	Mean	Std. dev.	Min	Max
CNR	TE	0.75	0.25	0.22	1
	SE	0.78	0.23	0.10	1
INSERM (1)	TE	0.47	0.23	0.15	1
	SE	0.63	0.27	0.03	1
INSERM (2)	TE	0.49	0.22	0.10	1
	SE	0.54	0.26	0.03	1

Notes: DEA (input-oriented) VRS, BCC model
CNR inputs: T_RES, GAI; output: INTPUB
INSERM(1) inputs: T_RES, GAI; output: INTPUB
INSERM(2) inputs: INSERM_RES, GAI; output: INTPUB no. obs. 212

tive because they exploit better opportunities generated by personal exchanges. By using the index GAI as an input we want to test the relevance of these effects.

Let us reiterate the comment that these measures are *relative* measures so that they *cannot* be compared directly. Having a higher average value of TE means that, on average, more institutes are located close to the efficiency level of the best performer, whatever their level of 'absolute' efficiency.

Table 4 shows the average value of technical efficiency (TE) and scale efficiency (SE) for the two institutions. Data clearly show that the relative efficiency of CNR is higher than the relative efficiency of INSERM, whatever the definition of input at INSERM is assumed. In other words, we can consider either all researchers declared by institutes or just internal employees as real inputs to the research process, but the final result in terms of relative efficiency is unchanged. Clearly, this is a result that goes much beyond the crude comparison of productivity ratios.

Table 5 translates the information into a number of efficiency indicators that describe the distribution of TE and SE values. According to these measures, the distribution of efficiency measures of CNR institutes is more favourable than the distribution at INSERM; that is, a larger proportion of CNR institutes are located close to the best performer. The I_TE indicator of CNR is higher than that of INSERM; that is, the percentage of CNR institutes with TE greater than 0.9 is 37% against 9% for INSERM Institutes. The same result applies for the I_SE indicator: the percentage of CNR institutes with SE greater than 0.9 is 48% against 22% for INSERM Institutes.

Table 5. Efficiency indicators

	percentage of institutes with TE > 0.9	percentage of institutes with SE > 0.9	1-min TE	1-min SE
Institution	I_TE	I_SE	TE Range	SE Range
CNR	37.04	48.15	0.78	0.90
INSERM (1)	9.39	22.07	0.85	0.97
INSERM (2)	8.02	12.74	0.90	0.97

Table 6. DEA efficiency scores corrected for bias and confidence interval

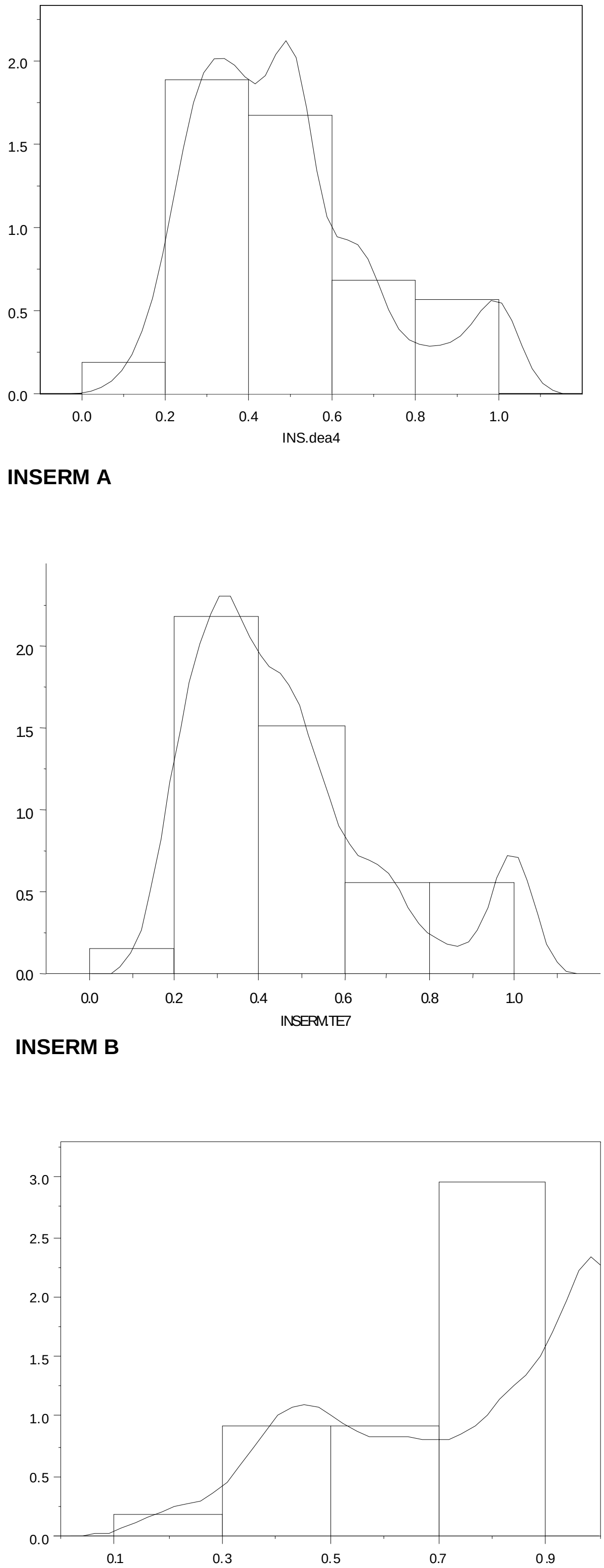
Institution	Mean	Std deviation	Mean length conf. interval
CNR	0.81 (0.06)	0.34	0.30
INSERM (1)	0.58 (0.11)	0.35	0.15
INSERM (2)	0.63 (0.14)	0.31	0.18

Notes: DEA input-oriented VRS model, Simar and Wilson (1998, 2000b) bootstrap procedure.
CNR inputs: T_RES, GAI; output: INTPUB
INSERM(1) inputs: T_RES, GAI; output: INTPUB
INSERM(2) inputs: INSERM_RES, GAI; output: INTPUB
The average bias is reported in brackets under the mean value.
The mean length of confidence interval at 95% has been computed applying the basic bootstrap procedure by Simar and Wilson (2000a).

Since DEA is a deterministic technique, it is not possible to draw statistical inferences from its scores. This problem is overcome by using bootstrap techniques in order to estimate the level of bias and the confidence interval of efficiency scores. Data in Table 6 show that CNR has a higher score, but also a larger confidence interval on average. Inspection of the relation between the two confirms, beyond any doubt, that the efficiency score is larger than at INSERM, at least within the definition of the particular inputs and outputs selected.

Let us explore the distribution of efficiency measures *within* each institution. Figure 1 compares the relative frequency of the value of the technical efficiency (TE) measure across institutes. It appears that a large proportion of INSERM institutes are located around the mean (considering as input both T_RES and INSERM_RES), with a small tail of higher efficiency values. By contrast, at CNR most institutes are close to maximum efficiency values. So, apart from considerations of the *absolute* level of efficiency reached by the two systems, it appears that CNR is better able to obtain maximum relative effort (i.e. effort relative to the best institute) from its affiliates.

These conclusions are reinforced by looking at different combinations of inputs. In Figure 2 a DEA exercise is carried out using two types of personnel as inputs: total number of researchers (T_RES) and technical/administrative personnel (ITA). While INSERM roughly reproduces the familiar bell-shaped distribution of technical efficiency with a small tail (again this result is confirmed using as input INSERM_RES instead of T_RES), CNR exhibits a bimodal distribution with a significant portion of institutes located close to maximum efficiency.[23]



INSERM A

INSERM B

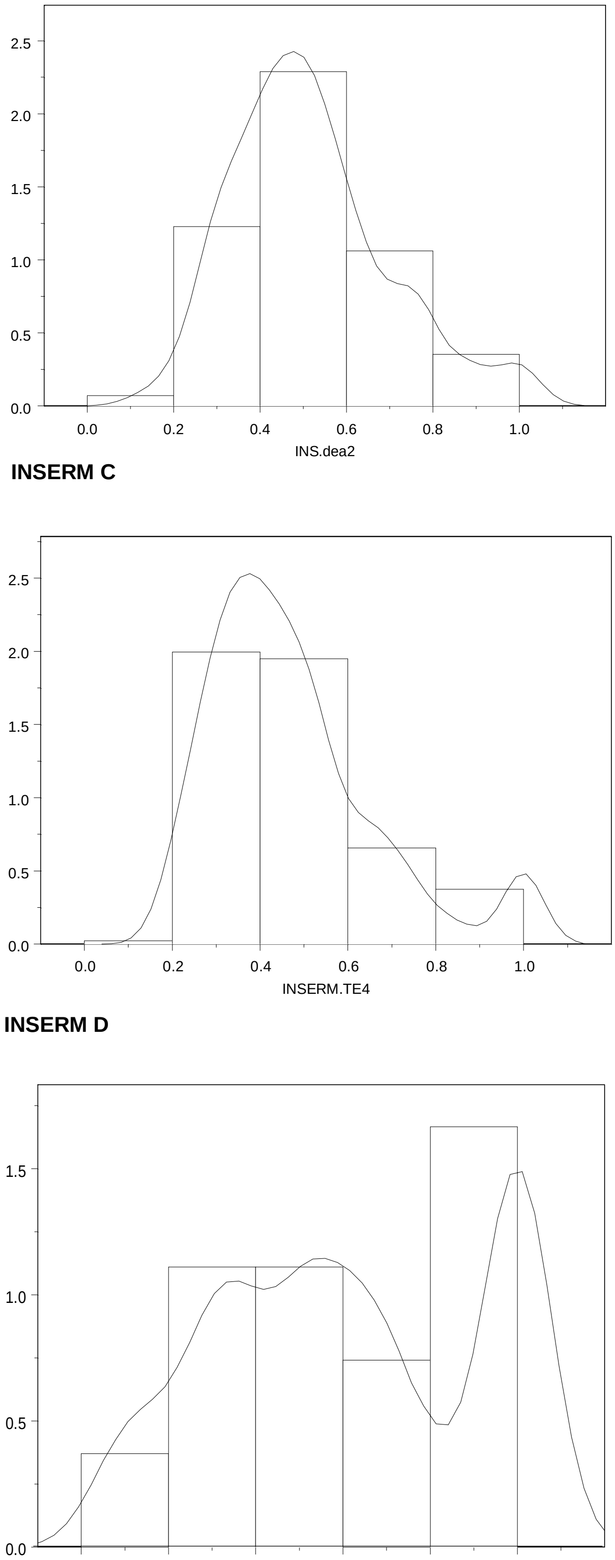
CNR

Figure 1. Relative frequency distribution of technical efficiency measure across institutes

Notes: Inputs: T_RES (or INSERM_RES for the INSERM B case), GAI

Output: INTPUB

The wavy lines represent the nonparametric density estimation obtained using a Gaussian kernel and a bandwidth determined applying the rule of thumb by Silverman (1986).



INSERM C

INSERM D

CNR

Figure 2. Relative frequency distribution of Technical Efficiency measure across institutes, with different inputs

Notes: Inputs: T_RES (or INSERM_RES for the INSERM D case), ITA

Output: INTPUB

The wavy lines represent the nonparametric density estimation obtained using a Gaussian kernel and a bandwidth determined applying the rule of thumb by Silverman (1986).

Robust nonparametric analysis

Table 7 shows the value of SPI(m) indicators for the three cases under analysis.[24] Since with robust indicators we can select the desired level of *robustness* (i.e. the percentage of extremely efficient institutes we want to exclude from the *relative* comparison) by fixing the percentage of observations that lie outside the order-*m* frontier, we fix this value at 5%, obtain the resulting value of *m* and then determine the value of indicators.

We computed the difference in efficiency between the unconditional — SPI(m) — and the conditional measure — SPI(m,g), and the proportion of institutes for which this difference is positive (i.e. adding GAI improves efficiency), negative or zero. For a large majority of INSERM institutes the geographical agglomeration is a significant factor in productivity. More than 88% of institutes at INSERM are sensitive to agglomeration factors, against a percentage of less than 45% for the CNR. Appendix A shows the plots of SPI(m) and SPI(m,g) against indicators of institute size.

Two results are striking:

1. Using a size *m* of the sample that ensures robustness, CNR exhibits higher internal efficiency than INSERM.
2. The difference between the unconditional efficiency SPI(m) and the conditional efficiency with geographical agglomeration is much larger in the case of INSERM.

Point 1 above confirms, with the maximum possible precision, that differences in efficiency between the two systems do not depend on size of the organisation. The result stated in point 2 is extremely interesting, since it gives an insight into possible explanations of the difference.

The fact that robust nonparametric techniques confirm the evidence from DEA means that the results do not depend on the effect of extreme values or outliers in the distribution of values of the organisations. Results from robust nonparametric techniques are the ultimate empirical support for efficiency analysis.

Table 7. Robust nonparametric indicators

		CNR	INSERM A	INSERM B
SPI(m)	Mean	0.913	0.592	0.665
	Std dev.	0.173	0.263	0.251
	Max	1.033	1.568	1.225
	Min	0.386	0.158	0.148

Notes: Indicators robust at 5%
CNR m =100, INSERM A m = 250, INSERM B m=150
CNR inputs: T_RES, GAI; output: INTPUB
INSERM A inputs: T_RES, GAI; output: INTPUB
INSERM B inputs: INSERM_RES, GAI; output: INTPUB

In the cases under analysis, because of the limitations of data, we do not draw any substantive implication regarding productivity, nor do we make any claim about national systems of public research

The use of DEA and its recently developed improvements makes it possible to go beyond crude evidence available through ratio analysis, without being subject to the severe methodological problems of multiple regression analysis based on the notion of production function. From a methodological point of view this is a very important achievement.

Of course, the substantive interpretation of the evidence relies entirely upon the accepted definition of inputs and outputs. In the cases under analysis, because of the limitations of data, we do not draw any substantive implication regarding productivity, nor do we make any claim about national systems of public research. The adoption of robust nonparametric techniques by the community of scientometrics scholars might lead to a stream of rigorous empirical evidence in the near future.

For the purposes of exploration of possible factors underlying differences in efficiency, in the following sections we discuss measurement errors and several possible explanations. Let us stress that also these explanations are contingent on the specific definition of inputs and outputs.

A comparison with measures of productivity

As we have seen in previous sections, several measures of productivity can be used in a comprehensive way in order to make a rigorous comparative productivity analysis. It is important to emphasize that each scientific productivity indicator (i.e. simple ratio, DEA index and SPI(m)) must be correctly interpreted, taking into account their economic meaning.

To highlight the relation existing between scientific productivity indicators, we compare the indicators applied in the illustrative example, using the following measures of correlation between the values at the level of research institutes:

- *Pearson correlation*, a measure of linear association;
- *Spearman correlation*, a commonly used nonparametric measure of correlation between two ordinal variables. For all of the cases, the values of each of the variables are ranked from smallest to largest, and the Spearman correlation coefficient is computed on the ranks;

- *Kendall's tau-b*, a nonparametric measure of association for ordinal or ranked variables that take ties into account. The sign of the coefficient indicates the direction of the relationship, and its absolute value indicates the strength, with larger absolute values indicating stronger relationships.

The results are shown in Tables 8–13.

Correlation analysis shows two general findings:

- All scientific productivity indicators are positively correlated to each other.
- While indicators based on nonparametric approaches (DEA, FDH, SPI) are highly positively correlated, they are weakly correlated with the crude ratio indicator (IPURES).

These findings confirm the importance of using several indicators and the superior performance of nonparametric techniques.

Table 8. Pearson correlations among productivity indicators – CNR results

	SPI	FDH	DEA	IPURES
SPI	1.000	.999**	.774**	.321
FDH		1.000	.765**	.317
DEA			1.000	.446*
IPURES				1.000

Notes: ** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).
We used the same inputs/outputs as the illustrative exercise.
SPI = scientific productivity indicator of order-*m*, input-oriented
FDH = free disposal hull, input-oriented
DEA = data envelopment analysis, input-oriented, VRS, BCC model
IPURES = INTPUB/T_RES

Table 9. Nonparametric correlations among productivity indicators – CNR results

		SPI	FDH	DEA	IPURES
Kendall's tau-b	SPI	1.000	.787**	.649**	.306*
	FDH		1.000	.613**	.279
	DEA			1.000	.381**
	IPURES				1.000
Spearman's rho	SPI	1.000	.831**	.752**	.364
	FDH		1.000	.706**	.330
	DEA			1.000	.456*
	IPURES				1.000

Notes: ** Correlation is significant at the .01 level (2-tailed).
* Correlation is significant at the .05 level (2-tailed).
We used the same inputs/outputs as the illustrative exercise.
SPI = scientific productivity indicator of order-m, input-oriented, robust at 5%
FDH = free disposal hull, input-oriented
DEA = data envelopment analysis, input-oriented, VRS, BCC model
IPURES = INTPUB/T_RES

Table 10. Pearson correlations among productivity indicators – INSERM (1) results

	SPI	FDH	DEA	IPURES (1)
SPI	1.000	.985**	.810**	.398**
FDH		1.000	.802**	.432**
DEA			1.000	.467**
IPURES (1)				1.000

Notes: ** Correlation is significant at the 0.01 level (2-tailed).
We used the same inputs/outputs as the illustrative exercise.
SPI = scientific productivity indicator of order-m, input-oriented, robust at 5%
FDH = free disposal hull, input-oriented;
DEA = data envelopment analysis, input-oriented, VRS, BCC model
IPURES (1) = INTPUB/T_RES

Table 11. Nonparametric correlations among productivity indicators – INSERM (1) results

		SPI	FDH	DEA	IPURES (1)
Kendall's tau-b	SPI	1.000	.952**	.680**	.237**
	FDH		1.000	.688**	.279**
	DEA			1.000	.306**
	IPURES (1)				1.000
Spearman's rho	SPI	1.000	.992**	.824**	.355**
	FDH		1.000	.818**	.406**
	DEA			1.000	.421**
	IPURES (1)				1.000

Notes: ** Correlation is significant at the .01 level (2-tailed).
We used the same inputs/outputs as the illustrative exercise.
SPI = scientific productivity indicator of order-m, input-oriented, robust at 5%
FDH = free disposal hull, input-oriented
DEA = data envelopment analysis, input-oriented, VRS, BCC model
IPURES (1)= INTPUB/T_RES

We also empirically confirm a recent result (see Chen and Iqbal Ali, 2002) according to which the top-ranked performance unit according to ratio analysis is a DEA frontier point (see Figures 9 and 10 in Appendix B). In fact, DEA subsumes the premise of ratio analyses, namely that a DMU that is most highly ranked with respect to the ratio of a single output to a single input dominates other DMUs. Such DMUs are easily identified as comprising a subset of frontier units in DEA. As noted in Chen and Iqbal Ali (2002), the ratio analysis fails to identify all types of dominating units as DEA does. A performance measure based on the ratio of a single output to a single input fails to capture the entirety of performance with respect to a set of outputs and inputs.

Table 12. Pearson correlations among productivity indicators – INSERM (2) results

	SPI	FDH	IPURES (2)	DEA
SPI	1.000	.986**	.580**	.904**
FDH		1.000	.624**	.876**
IPURES (2)			1.000	.558**
DEA				1.000

Notes: ** Correlation is significant at the 0.01 level (2-tailed).
We used the same inputs/outputs as the illustrative exercise.
SPI = scientific productivity indicator of order-m, input-oriented, robust at 5%
FDH = free disposal hull, input-oriented
IPURES (2)= INTPUB/INSERM_RES
DEA = data envelopment analysis, input-oriented, VRS, BCC model

Table 13. Nonparametric correlations among productivity indicators – INSERM (2) results

		SPI	FDH	IPURES	DEA
Kendall's tau-b	SPI	1.000	.918**	.418**	.852**
	FDH		1.000	.504**	.811**
	IPURES			1.000	.389**
	DEA				1.000
Spearman's rho	SPI	1.000	.981**	.571**	.964**
	FDH		1.000	.655**	.935**
	IPURES			1.000	.533**
	DEA				1.000

Notes: ** Correlation is significant at the .01 level (2-tailed).
We used the same inputs/outputs as the illustrative exercise.
SPI = scientific productivity indicator of order-m, input-oriented, robust at 5%
FDH = free disposal hull, input-oriented
IPURES (2) = INTPUB/INSERM_RES
DEA = data envelopment analysis, input-oriented, VRS, BCC model

Measurement errors

Before building a reasonable explanation of observed differences we must first take into account several sources of errors.[25]

Measurement errors on inputs

It is possible that the explanation of the observed difference is very simple: many people declared as members of INSERM institutes actually do *not* carry out research activity or, conversely, many people are involved in CNR research *without* being registered as members of institutes.[26] As a matter of fact, the gap between CNR and INSERM disappears if we assume that only INSERM researchers are actual inputs to the research process. The average researcher productivity at CNR (4.61 international papers per capita per year) lies in the range between INSERM productivity measured with all researchers (1.36 papers) and measured with internal researchers only (5.06).

It is very difficult to eliminate this type of measurement error. Qualitative observation on some INSERM institutes tells us that it is possible that some researchers from university and hospitals are declared as actively participating while their contribution is, in fact, minimal. A plausible reason for institute directors to overstate the number of researchers is to demonstrate a large volume of activity. But it is difficult to accept that this situation applies *uniformly* across all the INSERM system, so that in general the productivity of non-INSERM researchers is zero.

Another possibility is that productivity at CNR is enhanced by the contribution of university researchers that, on the contrary, are not declared as members. This is particularly true for some medical research centres in which close collaboration with the university is the norm.

Furthermore, while a *measurement error* is clearly possible from both sides, the magnitude of the difference in productivity is still very large.

If the main explanation of the difference is a measurement error on the input side, then there is room for an accurate rethinking of official statistics at both institutional and government level. A uniform international definition of research input should be adopted by all scientific institutions in all their official documentation. The *burden of proof* should not be placed on science policy scholars using available statistics, but on official sources.

Measurement errors on outputs

Another intriguing possibility is that there is a large measurement error in the output, leading to overestimating the production of CNR. This may take several forms:

- differences in the criteria of definition of publications in official sources;
- differences in patterns of co-authorship;
- strong heterogeneity in the nature and/or quality of publications.

The first possibility is difficult to evaluate. We restricted the examination to international publications, assuming that this definition does not create strong disparities. We consulted CNR reports in various years and tend to believe that, while the definition of total publications may be subject to overestimation (e.g. technical reports considered to be publications), the measurement of international publications should be reliable.

The second possibility is more serious. It is possible that papers at CNR have more co-authors, so that a researcher declares several papers as co-author while in fact the bulk of research has been done by others (perhaps at the university). The larger the number of co-authors the larger the possibility to inflate the contribution of CNR institutes. As a matter of fact, in the field of brain disorder, Lewison et al (2002) found that papers from Italian institutions have the largest number of co-authors in Europe. In order to check for this effect, the average number of co-authors of papers declared by CNR and INSERM should be computed. We carried out this task on the complete list of papers published by the two institutions in 1997, as available in the PubMed databank. The average number of co-authors is 6.01 for INSERM (total number of papers = 4,669) and 6.76 for CNR (total number of papers = 1,157). Although a difference is evident, it is hard to believe that it accounts for a large part of observed difference in productivity.

The third possibility is also complex. The argument goes as follows: CNR looks more productive, but in fact it produces papers with a strong clinical orientation, as opposed to basic research. Clinical papers are produced in greater quantities but are published in journals with a lower impact factor and/or overall quality. As a result there should not be a concern in the less productive institution, since they produce fewer papers but of higher quality.

We tried to check this hypothesis. First of all we interviewed a small number of scientists and historians of medicine and addressed this problem in order to receive qualitative feedback. The general opinion is that scientific production at CNR is, on the contrary, strongly oriented towards biochemistry, genetics and molecular biology, all disciplines in the basic research field. In order to test this effect more rigorously, we examined the complete list of journals in which papers were published in 1997 from the two institutions.

On the list of journals we tested:

- a measure of overlapping between the most important journals in the top 30;
- an inspection of the measure of clinical vs. basic research orientation, following the scale developed in Narin *et al* (1976).[27]

The outcomes of this analysis are as follows:

- Of the first 30 journals in the two institutions, seven are in common (*Journal of Biological Chemistry*, *Blood*, *British Journal of Haematology*, *Journal of Clinical Endocrinology and Metabolism*, *Biochemical Journal*, *Genomics*, and *International Journal of Cancer*).
- In the list of the top 30 INSERM journals, a good majority can be classified as clinical journals (around 57%).

The list of first 30 journals for the two institutions can be found in Appendix C. Although the analysis is far from being conclusive, the overall implication is that it is difficult to accept an explanation of observed differences based on an assumed orientation of INSERM output towards the basic research end of the spectrum and of CNR towards the clinical end.

Sample size bias

Another argument might be that in the comparison between a large institution such as INSERM and a small number of CNR institutes there is an inherent tendency towards penalising the large one. It may be that in a large number of institutes there are almost certainly many small or inefficient institutes, created in order to follow many emerging areas over time. It would be inevitably more difficult to preserve quality in large institutions.

Although the statistical merits of this argument are not clear, the criticism applies only to the direct comparison of average productivity measures. However, our robust nonparametric approach is not sensitive to the size of samples. As a matter of fact, one of the attractive properties of these indicators is exactly that they allow a productivity comparison between institutions or countries of different size.

Looking for candidates' explanatory factors

The complete elimination of all the aforementioned sources of error requires a detailed *ad hoc* study. In a further study we plan to compute *individual* productivity measures of scientists at both institutions, by downloading publications and accounting for patterns of co-authorship. Differences in individual productivity will then be linked to differences in productivity at institute level. So far, we are led to a situation where it is not possible to exclude that part of the observed difference in productivity is due to measurement errors, particularly in inputs and in the pattern of co-authorship.

However, since to eliminate the observed difference all measurement errors must have the *same* sign at the same time (i.e. reducing productivity at CNR or increasing productivity at INSERM), it is still legitimate to assume that part of the difference must have some substantial, as opposed to measurement-related, explanation. Furthermore, if we believe that measurement errors lie mainly in the input side, we should recall that robust nonparametric indicators show that INSERM is less efficient even when the adopted indicator includes *only* INSERM researchers, eliminating the source of error. In other words, with robust nonparametric techniques we are not comparing *absolute efficiency* of the two institutions, but *relative efficiency*, or the way in which institutes are distributed with respect to their own most efficient peers.

Size and agglomeration effects

We explored the possibility that INSERM institutes are more dependent on scale effects; that is, average productivity is decreased by a larger proportion of institutes being sub-optimal in size. In order to detect if size effects are in place, we applied a locally weighted least-squares (Loess) technique (see

Cleveland, 1993, 1994). It is a local regression technique that is a generalization of running means, which gets a predicted value at each point by fitting a weighted linear regression, where the weights decrease with distance from the point of interest.

Connecting these predicted values produces a smooth curve. The primary parameter affecting the smoothness of the fit is the span, which controls the speed with which the influence of points decreases with distance from the point of interest. Locally weighted least-squares is used for nonparametric curve fitting. This is essentially a noise-reduction smoothing algorithm. A 'locally weighted' linear regression is used to obtain smoothed values on a scatter plot of the associated points of value of y, given the values for x.

Scatter graphs compare INSERM and CNR with reference to how various indicators used in this study vary as a function of institute size, as measured by the total number of researchers (T_RES):

- robust nonparametric indicator of order-*m* (input: T_RES; output: INTPUB), Figures 3 and 4 in Appendix A;
- robust nonparametric indicator of order-*m* conditional to geographic agglomeration (GAI) (input: T_RES; output: INTPUB; environmental variable: GAI), Figures 5 and 6 in Appendix A.

First of all, careful inspection of Figures 3 to 6 in Appendix A shows an interesting pattern. The efficiency of INSERM institutes clearly decreases with size as far as we consider technical efficiency (TE) or a robust indicator of order-*m* that includes only researchers as inputs. Quite to the contrary, CNR institutes exhibit a U-shaped pattern, meaning that efficiency decreases with size but after a threshold it increases again.

However, a U-shaped pattern is apparent for INSERM if we include in the robust indicator an environmental variable, namely *geographical agglomeration*. A reasonable interpretation is as follows: in the French system agglomeration has a strong importance for scientific productivity. Isolated institutes are at a disadvantage, probably because they have less access to specialised large-scale equipment and high-quality students or young researchers. Given the lack of systematic interaction with universities, isolation means deprivation. This must be particularly true for large institutes, which utilise their resources in a strongly inefficient way.

The same effect does not apply to the Italian case. At CNR extremely productive institutes can be found at two extremes of the size range: in small and dynamic institutes and in large, well-organised institutes. Location and agglomeration effects do not play a great role.

Workforce composition effects

Another line of explanation refers to the structure of personnel. INSERM institutes have a significantly lower share of *technical personnel*. This means that researchers must perform technical tasks themselves, or employ students, or reduce the use of specialised equipment. In general, this decreases scientific productivity. Particularly in the new research regime after molecular biology, this limitation may be severe.

We are not in a position to compare the two institutions on the marginal impact of technical personnel, but we suggest the effect may be highly significant. In fact, if administrative personnel is a rather fixed proportion of the total, then almost all difference in the composition can be attributed to technicians. Going back to Figure 2, it is clear that technical efficiency of the use of both researchers and technical/administrative staff at INSERM is *relatively* inferior.

From an economic point of view the role of technicians may be substantial in determining increasing returns. If technicians are strictly complementary to researchers, when the size of institutes grows their number should grow approximately in the same proportion as that of researchers. If this is not the case, technicians act as a fixed factor and institutes enter into a regime of decreasing returns to scale. This may explain the finding on diseconomies of scale in the French system (see Figure 7 in Appendix A).

Conclusions and future research

The possibility that our analysis is vitiated by large *measurement errors* cannot be excluded. If this is the case, policymakers should correct them rapidly, since almost any decision would be based on wrong information. If, on the contrary, our data are fundamentally correct, we can ask the question whether observed differences in the relations between inputs and outputs available point to a more general difference in productivity.

This paper is mainly a methodological exercise, requiring further refining and extension. A future line of inquiry will be to build up indicators of individual productivity of scientists and explore the relationship between individual and organisational productivity.

Comparative analysis with advanced efficiency measurement techniques is useful to obtain rigorous evidence. Interpretation needs further accumulation of evidence. With a richer set of data, advanced nonparametric techniques offer a powerful tool to draw substantive conclusions on scientific productivity.

Appendix A. Loess plot of scientific indicators against size

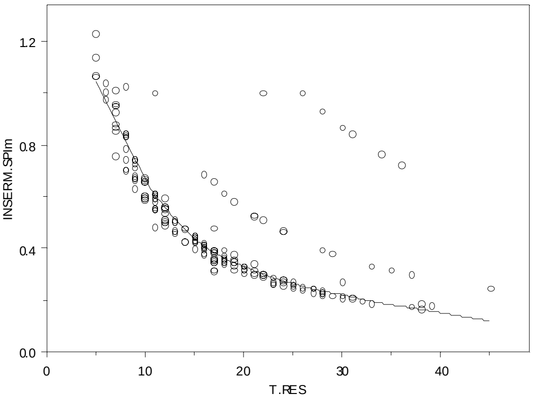


Figure 3. INSERM Loess plot of SPI(m) against T_RES (total researchers)

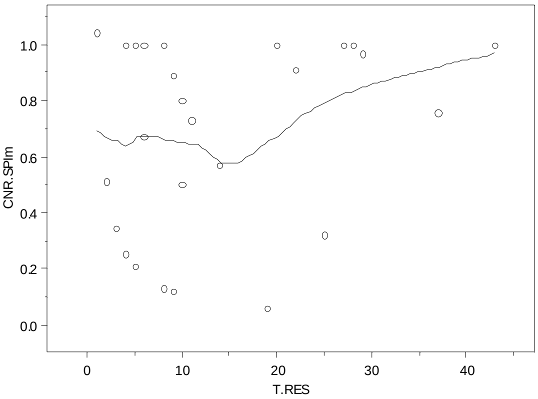


Figure 4. CNR Loess plot of SPI(m) against T_RES (total researchers)

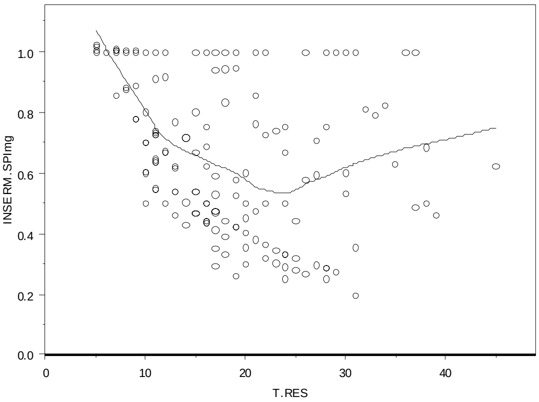


Figure 5. INSERM Loess plot of SPI(m,g) against T_RES (total researchers)

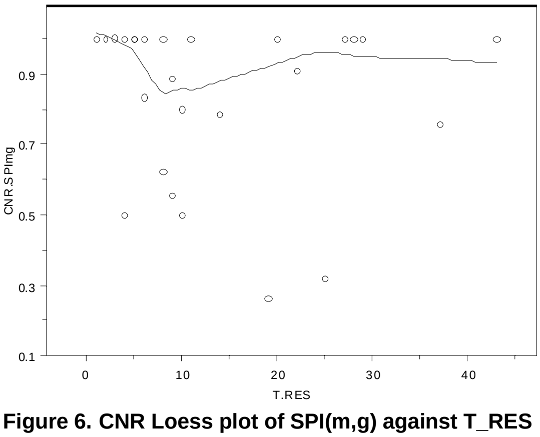


Figure 6. CNR Loess plot of SPI(m,g) against T_RES (total researchers)

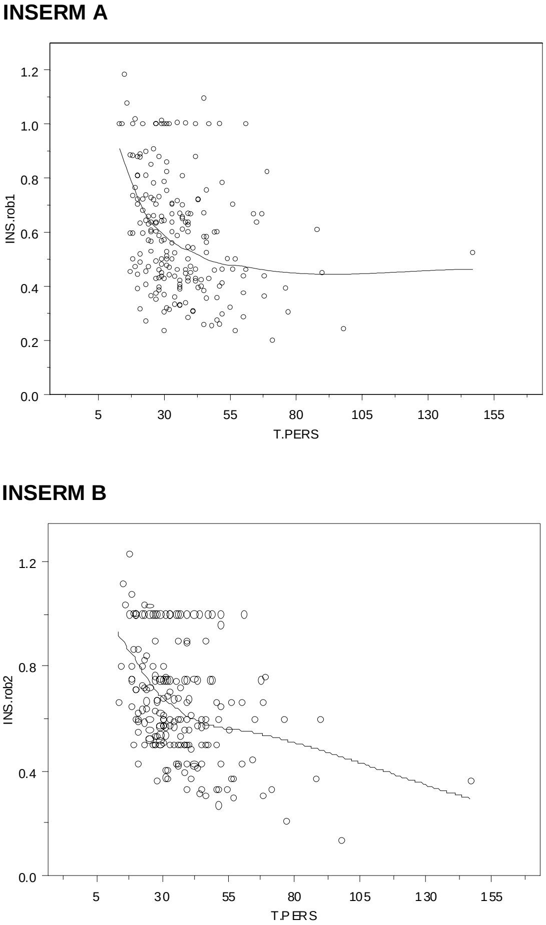


Figure 7. INSERM Loess plot of SPI(m) against T_PERS (total personnel)

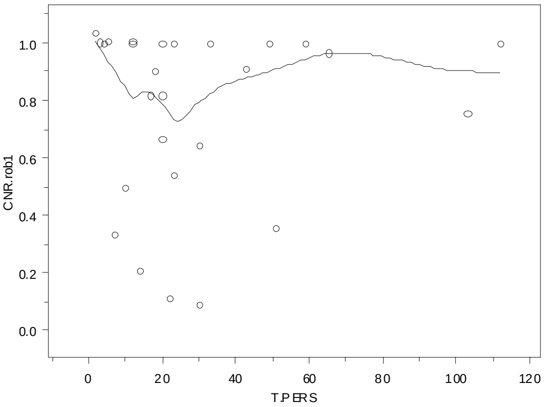


Figure 8. CNR Loess plot of SPI(m) against T_PERS (total personnel)

Appendix B. Line plots of DEA indices

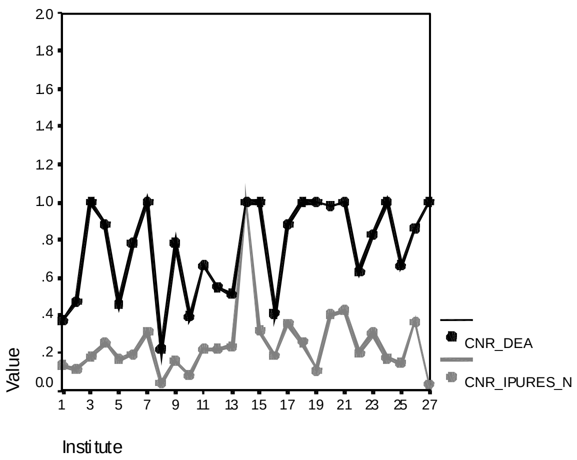
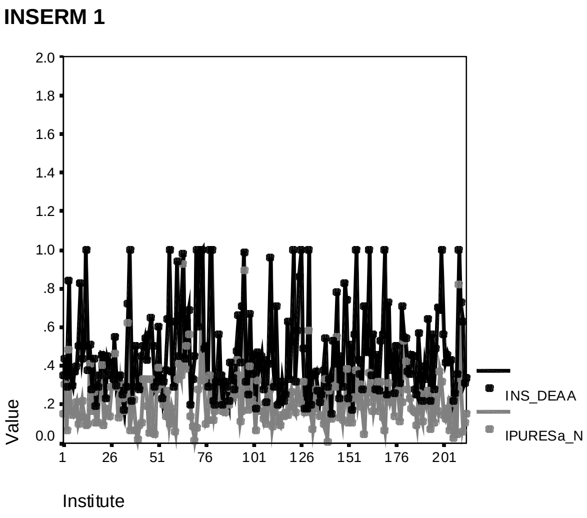


Figure 9. CNR Line Plot of DEA indices and ratio measures of productivity (IPURES_N)

Note: We have normalised the measure of IPURES in order to have a value between 0 and 1 and make the comparison with the DEA scores.



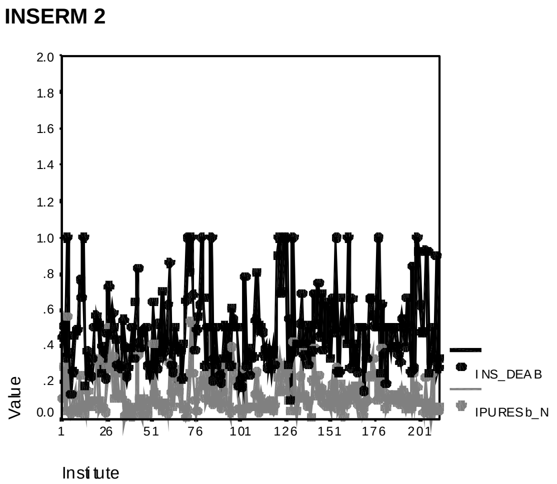


Figure 10. INSERM line plot of DEA indices and ratio measures of productivity (IPURES_N)

Notes: INSERM 1 input: T_RES
INSERM 2: input: INSERM_RES

Appendix C. Lists of journals, index of basic (B) vs clinical research (C), mean number of authors

Table 14. INSERM list of journals/publications, 1997

Rank	Journal	Index	Count	% of total	Mean number of authors
1	*Journal of Biological Chemistry*	B (4)	103	2.21	6.25
2	*Thrombosis and Haemostasis*	C	98	2.10	6.53
3	*Blood*	C	80	1.71	7.44
4	*Journal of Immunology*	C (3)	67	1.43	7.34
5	*American Journal of Human Genetics*	B	56	1.20	8.7
6	*Hepatology*	C	56	1.20	7.43
7	*Journal of the American Society of Nephrology*	C	45	0.96	5.62
8	*Journal of Hepatology*	C	38	0.81	7.26
9	*European Journal of Immunology*	C	37	0.79	6.38
10	*Gastroenterology*	C	35	0.75	6.97
11	*Medicine Sciences*	B	35	0.75	3.74
12	*American Journal of Physiology*	B (4)	33	0.71	5.36
13	*British Journal of Haematology*	C	32	0.69	7.94
14	*Journal of Investigative Dermatology*	C	32	0.69	6.06
15	*Oncogene*	C	32	0.69	6.88
16	*Brain Research*	B	31	0.66	5.1
17	*Febs Letters*	B (4)	31	0.66	6.42
18	*Journal of Clinical Endocrinology and Metabolism*	C	31	0.66	7.13
19	*Journal of Clinical Investigation*	C (3)	31	0.66	8
20	*Biochemical Journal*	B (4)	30	0.64	5.6
21	*Cytogenetics and Cell Genetics*	B	30	0.64	6.83
22	*American Journal of Respiratory and Critical Care Medicine*	C	29	0.62	6.72
23	*European Journal of Cancer*	C	28	0.60	7.54
24	*Gastroenterologie Clinique et Biologique*	C	28	0.60	5.75
25	*Neuroscience Letters*	B	28	0.60	4.71
26	*Genomics*	B	26	0.56	8.08
27	*European Journal of Biochemistry*	B	25	0.54	6.32
28	*International Journal of Cancer*	C	25	0.54	7.24
29	*Journal of Neuroscience*	B	25	0.54	5.2
30	*Neuroscience*	B	25	0.54	5
Total			4,669	1	6.01

Note: In this, and in Table 15 (3) and (4) are the levels of basic vs. clinical research given by the Narin *et al* (1976) classification of biomedical journals. Because of the early date of that paper, not all the biomedical journals were found in that paper; hence we report in brackets their levels for only some of the journals.

(*continued*)

Appendix C. (*continued*)

Table 15. CNR list of journals/publications, 1997

Rank	Journal	Index	Count	% of total	Mean number of authors
1	*Transplantation Proceedings*	C (3)	19	1.64	9.63
2	*Genomics*	B	19	1.64	7.21
3	*Gene*	B	16	1.38	8.06
4	*Biochemical Journal*	B (4)	14	1.21	8.36
5	*Radiol. Med. (Torino)*	C	14	1.21	7.21
6	*Journal of Clinical Endocrinology and Metabolism*	C	11	0.95	8.64
7	*Haematologica*	C	11	0.95	7.64
8	*International Journal of Cancer*	C (3)	10	0.86	7.6
9	*Circulation*	C	10	0.86	7.5
10	*Neuroreport*	B	10	0.86	5.3
11	*Human Molecular Genetics*	B	9	0.78	13.44
12	*Proceedings of the National Academy of Sciences of USA*	B	9	0.78	7.22
13	*European Heart Journal*	C	9	0.78	6.67
14	*Biochemical and Biophysical Research Communications*	B	9	0.78	6.44
15	*Recenti Prog. Med.*	C	9	0.78	1
16	*Blood*	C	8	0.69	10.38
17	*Journal of Biological Chemistry*	B (4)	8	0.69	7
18	*Human Immunology*	C	8	0.69	5.75
19	*American Journal of Medical Genetics*	B	8	0.69	5.13
20	*Clin. Ter.*	C	8	0.69	3.88
21	*European Respiratory Journal*	C	7	0.61	8
22	*Journal Med. Eng. Technol.*	B	7	0.61	6.43
23	*Journal of Molecular Biology*	B (4)	7	0.61	5.86
24	*Clinical Chemistry*	C (3)	7	0.61	5
25	*Giornale Italiano di Cardiologia*	C	6	0.52	11.83
26	*Leukemia and Lymphoma*	C	6	0.52	9.67
27	*British Journal of Haematology*	C	6	0.52	9.33
28	*Tumori*	C (2)	6	0.52	8.67
29	*Monaldi Archives for Chest Disease*	C	6	0.52	8.5
30	*Research in Virology*	B	6	0.52	6
Total			1,157	1	6.76

Acknowledgements

The article is a revised edition of a paper presented to the Seventh International Science and Technology Indicators Conference, Karlsruhe, 25–28 September 2002. Participants in the conference provided detailed comments and criticisms. Two anonymous referees offered suggestions that greatly improved the paper. Part of the evidence discussed in this paper had been presented at conferences and seminars at SPRU (NPR Net Conference), CNR (ISPRI, Rome) and INSERM (Institute de Cancerologie, Marseille). We are indebted to Daniele Archibugi, Bianca Maria Potì and Giorgio Sirilli (ISPRI-CNR), Luigi Donato (IFC-CNR), Claude Mawas (INSERM), Gilberto Corbellini (University of Rome La Sapienza) for helpful discussion, although only we are responsible for the content of this version. We thank Maria Adelaide Caligo (University of Pisa) and Roberto Girardello (Pharmacia) for help in classification of institutes and Marco Brancher and Francesca Pierotti (Sant'Anna School of Advanced Studies) for assistance in data analysis.

Notes

1. For a discussion on significant limitations of the production function approach in the evaluation of government-sponsored research projects see Link (1996).
2. See for example Adams and Griliches (2000).
3. For more details see Banker, Charnes and Cooper (1984).
4. Charnes, Cooper and Rhodes (1978) proposed a model that had an input orientation and assumed constant returns to scale (CRS). In their original study, they described DEA as a

> mathematical programming model applied to observational data that provides a new way of obtaining empirical estimates of extremal relations – such as the production functions and/or efficient production possibility surfaces that are a cornerstone of modern economics.

For a more in depth mathematical explanation see Seiford and Thrall (1990).

5. Banker, Charnes and Cooper (1984) proposed an extension of the CRS DEA model to account for variable returns to scale (VRS) situations. The Banker, Charnes and Cooper (1984) model (BCC hereafter) distinguishes between technical and scale inefficiencies by estimating pure technical efficiency at the given scale of operation. A *classical* reference on DEA is Fare, Grosskopf and Lovell (1994). For an updated review of the DEA models available in literature, see Cooper, Seiford and Tone (1999).
6. For a survey of recent contributions in stochastic frontier analysis see Kumbhakar and Lovell (2000).
7. For a comparison of DEA with least squares econometric production models, total factor productivity indices and stochastic frontiers, see Coelli, Rao and Battese (1998).
8. The specifications of more general distributional forms, such as the truncated-normal (Stevenson 1980) and the two-parameter gamma (Greene 1990), have partially alleviated this problem, but the resulting efficiency measures may still be sensitive to distributional assumptions.
9. We refer to the standard models, widely applied in literature, introduced by Charnes, Cooper and Rhodes (1978) and by BCC (1984).
10. For an overview on statistical inference in nonparametric frontier estimation see Simar and Wilson (2000a).
11. For a description of bootstrap applications in nonparametric efficiency models see Simar and Wilson (1998, 2000b).
12. As pointed out in Schubert and Braun (1996), comparative assessment of scientometric indicators has to be based on *normalized* scientometric indicators that first gauge them against a properly chosen reference standard, then compare their relative standing. There are basically two types of approaches in setting reference standards for cross-field normalization of scientometric indicators. The first type is based on a *prior* classification of units into science field categories of the required depth. In the second type for each unit to be assessed, a specific standard is set on the basis of automatic algorithms or human expertise. In this paper we decided to use the first approach that is easier to comprehend and accept, even if we lose in flexibility. For more details on *normalization* methods see Schubert and Braun (1996).
13. For a description of the test procedures see Simar and Wilson, 2001.
14. The FDH estimator has been proposed by Deprins, Simar and Tulkens (1984). For a survey on FDH applications see Van den Eeckaut (1997).
15. That is, it is allowed to destroy goods without costs. For a more formal definition see Deprins, Simar and Tulkens (1984).
16. The number of NN has been set to 10 in order to estimate the distribution/survivor function of DMUs. This choice reflects the trade-off between precision in the evaluation of the effects of environmental variables and a sufficient number of observations in order to carry out the estimation. In order to evaluate the influence of the choice of the number of NN we have done several estimations using smaller and higher values, controlling so that the obtained results are not affected by these choices.
17. A methodology to evaluate the effects of environmental variables based on sensitivity analysis (i.e. using several values of *m*) and using the difference of conditional and unconditional robust efficiency score is described in Daraio (2002) where an application to portfolio analysis is provided.
18. The complete database includes a number of other variables — such as research funds, personnel cost or various categories of CNR researchers — that do not have any correspondence in the INSERM case and are not therefore reported here. See Bonaccorsi and Daraio (2002) for discussion.
19. It is significant at less than 10% for T_RES and INTPUB, while is not significant for T_PERS, INTPUB/T_PERS and INTPUB/T_RES.
20. In Bonaccorsi and Daraio (2002) we performed a DEA exercise on the Italian CNR, including also research funds as input and the funds collected from market contracts as output. Here we have not included these variables as we do not have data for the INSERM case.
21. Since we have separate data for the two categories for CNR, but not for INSERM, we collapsed the two into a single unit. Results of the exercise are not reported at length in the paper; they are available from the authors on request.
22. For an analysis on knowledge spillovers in biotechnology, based on the ability of scientists to appropriate the value of knowledge embedded in their human capital along with the incentive structures, see Audretsch and Stephan (1999). Bonaccorsi (2002) describes the matching properties between research regimes and public research institutions, particularly illuminating for biomedical institutions.
23. We notice that the average value of TE for INSERM is 0.473 (Inputs: T_RES, ITA); CNR TE is 0.617; INSERM SE is 0.644, CNR SE is 0.519.
24. All the detailed computations are available from the authors on request.
25. On the role of ambiguity in measurement and analytical methods for exploring its impact see Bookstein and Wright (1997).
26. The latter possibility has been suggested by participants to the seminar of the first author at ISPRI-CNR in Rome (June 2002), pointing to the contribution of university professors to papers published jointly with CNR researchers.
27. In biomedical research the use of impact factor is subject to even more limitations than in other fields (see Schwartz and Hellin, 1996). Journal impact factors as a measure of quality have many general limitations (see Moed and Van Leeuwen, 1996; Seglen, 1997).

References

J D Adams and Z Griliches (2000), ‘Research productivity in a system of universities’, in D Encaoua *et al* (editors), *The Economics and Econometrics of Innovation* (Kluwer Academic Publishers, Netherlands), pages 105–140.

D J Aigner and S F Chu (1968), ‘On estimating the industry production function’, *American Economic Review*, 58, pages 826–839.

D J Aigner, C A K Lovell and P Schmidt (1977), ‘Formulation and estimation of stochastic frontier production function models’, *Journal of Econometrics*, 6, pages 21–37.

D B Audretsch and M P Feldman (1996), ‘R&D spillovers and the geography of innovation and production’, *American Economic Review*, 86, pages 630–640.

D B Audretsch and E S Stephan (1999), ‘Knowledge spillovers in biotechnology: sources and incentives’, *Journal of Evolutionary Economics*, 9, pages 97–107.

R D Banker, A Charnes and W W Cooper (1984), ‘Some models for estimating technical and scale inefficiencies in DEA’, *Management Science*, 32, pages 1613–1627.

A Bessent and W E Bessent (1980), ‘Determining the comparative efficiency of schools through DEA’, *Educational Administration Quarterly,* 16, pages 57–75.

A Bessent, W Bessent, J Kennington and B Reagan (1982), ‘An application of mathematical programming to assess productivity in the Houston independent school district’, *Management Science,* 28(12), pages 1355–1367.

A Bonaccorsi (2002), ‘Matching properties: research regimes and public research institutions’, paper presented to the workshop *Science as an Institution: The Institutions of Science*, Siena, Certosa di Pontignano, January 25–26.

A Bonaccorsi and C Daraio (2002) ‘The organization of science: size, agglomeration and age effects in scientific productivity’, paper presented to the conference *Rethinking Science Policy*, Brighton, SPRU.

A Bookstein and B Wright (1997), ‘Ambiguity in measurement’, *Scientometrics*, 40, pages 423–436.

M Bordons and M A Zulueta (1997), ‘Comparison of research team activity in two biomedical fields’, *Scientometrics*, 40, pages 423–436.

C Cazals, J-P Florens and L Simar (2002), ‘Nonparametric frontier estimation: a robust approach’, *Journal of Econometrics*, 106, pages 1–25.

A Charnes, W W Cooper and E Rhodes (1978), ‘Measuring the efficiency of decision making units’, *European Journal of Operational Research*, 2, pages 429–444.

Y Chen and A Iqbal Ali (2002), ‘Output-input ratio analysis and DEA frontier’, *European Journal of Operational Research,* 142, pages 476–479.

W S Cleveland (1993), *Visualizing Data* (Hobart Press).

W S Cleveland (1994), *The Elements of Graphing Data* (Hobart Press).

T Coelli (1996), *Assessing the Performance of Australian Universities using Data Envelopment Analysis*, mimeo (Centre for Efficiency and Productivity Analysis, University of New England, USA).

T Coelli, D S P Rao and G E Battese (1998), *An Introduction to Efficiency and Productivity Analysis* (Kluwer Academic Publishers).

P M D Collins (1991), *Quantitative Assessment of Departmental Research*, SEPSU Policy study no. 5 (The Royal Society, London).

W W Cooper, L M Seiford and K Tone (1999), *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software* (Kluwer Academic Publishers, Boston).

H D Daniel and R Fisch (1990), ‘Research performance evaluation in the German university sector’, *Scientometrics*, 19, page 349.

C Daraio (2002), ‘Portfolio analysis: estimation of efficient frontier using nonparametric robust methods’, *Memoire Des Stat* (Université Catholique de Louvain, Louvain-la-Neuve, Belgium).

G Debreu (1951), ‘The coefficient of resource utilisation’, *Econometrica,* 19, pages 273–292.

D Deprins, L Simar and H Tulkens (1984), ‘Measuring labor-efficiency in post offices’, in M Marchand, P Pestieau and H Tulkens (editors),*The Performance of Public Enterprises: Concepts and Measurement* (North-Holland, Amsterdam), pages 243–267.

R Fare, S Grosskopf and C A K Lovell (1994), *Production Frontiers* (Cambridge University Press, Cambridge).

R Fare, S Grosskopf and W Weber (1989), ‘Measuring school district performance’, *Public Finance Quarterly*, 17, pages 409–428.

M J Farrell (1957), ‘The measurement of the productive efficiency’, *Journal of the Royal Statistical Society, Series A*, CXX(3), pages 253–290.

W H Greene (1990), ‘A gamma-distributed stochastic frontier model’, *Journal of Econometrics,* 46, pages 141–164.

S Grosskopf, K Hayes, *et al* (1999), ‘Anticipating the consequences of school reform: a new use of DEA’, *Management Science*, 45, pages 608–620.

S Grosskopf and C Moutray (2001), ‘Evaluating performance in Chicago public high schools in the wake of decentralization’, *Economics of Education Review,* 20, pages 1–14.

S Katz (1994), ‘Geographical proximity and scientific collaboration’, *Scientometrics*, 31, pages 31–43.

T C Koopmans (1951), ‘An analysis of production as an efficient combination of activities’ in T C Koopmans (editor), *Activity Analysis of Production and Allocation,* Cowles Commission for Research in Economics, Monograph No. 13 (Wiley, New York).

P Korhonen, R Tainio and J Wallenius (2001), ‘Value efficiency analysis of academic research’, *European Journal of Operational Research,* 130, pages 121–132.

S C Kumbhakar and C A K Lovell (2000), *Stochastic Frontier Analysis* (Cambridge University Press, Cambridge).

G Lewison (1998), ‘New bibliometric techniques for the evaluation of medical schools’, *Scientometrics*, 41, pages 5–16.

G Lewison and G Dawson (1998), ‘The effect of funding on the outputs of biomedical research’, *Scientometrics*, 41, pages 17–27.

G Lewison, C Henderson and K Willcox-Jay (2002), ‘The anatomy of mental health research in 15 OECD countries’, paper presented at the Seventh International Science and Technology Indicators Conference, Karlsruhe, September 25–28.

A N Link (1996), ‘Economic performance measures for evaluating government sponsored research’, *Scientometrics*, 36, pages 325–342.

B R Martin (1996), ‘The use of multiple indicators in the assessment of basic research’, *Scientometrics*, 36, page 343.

W Meeusen and J van den Broeck (1977), ‘Efficiency estimation from Cobb-Douglas production functions with composed error’, *International Economic Review*, 18, pages 435–444.

H F Moed and T N van Leeuwen (1996), ‘Impact factors can mislead’, *Nature,* 381, pages 186.

F Narin and K S Hamilton (1996), ‘Bibliometric performance measures’, *Scientometrics*, 36, pages 293–310.

F Narin, G Pinski and H H Gee (1976), ‘Structure of the biomedical literature’, *Journal of the American Society for Information Science*, 27(1), pages 25–45.

P Ramsden (1994), ‘Describing and explaining research productivity’, *Higher Education*, 28.

S Rousseau and R Rousseau (1997), ‘Data envelopment analysis as a tool for constructing scientometric indicators’, *Scientometrics*, 40, pages 45–56.

S Rousseau and R Rousseau (1998), ‘The scientific wealth of European nations: taking effectiveness into account’, *Scientometrics*, 42, pages 75–87.

A Schubert and T Braun (1996), ‘Cross-field normalization of scientometric indicators’, *Scientometrics*, 36, pages 311–324.

S Schwartz and J L Hellin (1996), ‘Measuring the impact of scientific publications: the case of the biomedical sciences’, *Scientometrics*, 35, pages 119–132.

P O Seglen (1997), ‘Why the impact factor of journals should not be used for evaluating research’, *BMJ*, 314, pages 498–502.

L M Seiford and R M Thrall (1990), ‘Recent developments in DEA: the mathematical approach to frontier analysis’, *Journal of Econometrics*, 46, pages 7–38.

B W Silverman (1986), *Density Estimation for Statistics and Data Analysis* (Chapman & Hall, London).

L Simar and P Wilson (1998), ‘Sensitivity of efficiency scores: how to bootstrap in non-parametric frontier models’, *Management Sciences,* 44(1), pages 49–61.

L Simar and P W Wilson (2001), ‘Testing restrictions in nonparametric efficiency models’, *Communications in Statistics*, 30(1), pages 159–184.

L Simar and P Wilson (2000a), ‘Statistical inference in nonparametric frontier models: the state of the art’, *Journal of Productivity Analysis,* 13, pages 49–78.

L Simar and P Wilson (2000b), ‘A general methodology for bootstrapping in nonparametric frontier models’, *Journal of Applied Statistics*, 27(6), pages 779–802.

R E Stevenson (1980), ‘Likelihood functions for generalised stochastic frontier estimation’, *Journal of Econometrics*, 13, pages 57–66.

J G Thursby and S Kemp (2002), ‘Growth and productive efficiency of university intellectual property licensing’, *Research Policy*, 31(1), pages 109–124.

P van den Eeckaut (1997), ‘Free disposal hull and measurement of efficiency: theory, application and software’, PhD thesis, Faculté de Sciences Economiques, Sociales et Politiques, Nouvelle Série (229) (Université Catholique de Louvain, Louvain-la-Neuve, Belgium).

A F J van Raan (1993), ‘Advanced bibliometric methods to assess research performance and scientific development: basic principles and recent practical applications’, *Research Evaluation*, 3, page 151.

A F J van Raan (1997), ‘Scientometrics: state of the art’, *Scientometrics*, 38, pages 205–218.

L G Zucker, M R Darby and J Armstrong (1998), ‘Geographically localized knowledge: spillovers or markets?’ *Economic Inquiry*, XXXVI, pages 65–86.