A *mixed* splicing procedure for economic time series

Angel de la Fuente*

Instituto de Análisis Económico (CSIC)

December 2009

Abstract

This note develops a flexible methodology for splicing economic time series that avoids the extreme assumptions implicit in the procedures most commonly used in the literature. It allows the user to split the required correction to the older of the series being linked between its levels and growth rates on the basis what he knows or conjectures about the persistence of the factors that account for the discrepancy between the two series that emerges at their linking point. The time profile of the correction is derived from the assumption that the error in the older series reflects the inadequate coverage of emerging sectors or activities that grow faster than the aggregate.

Key words: linking, splicing, economic series
JEL Classification: C82, E01

* This paper is part of a research project cofinanced by the ERDF and Fundación Caixa Galicia. Additional financial support from the Spanish Ministry of Science and Innovation under project ECO2008-04837/ECON is also gratefully acknowledged. I would like to thank the participants in the Economic History Workshop at Universidad Carlos III for their useful comments.

1. Introduction

In order to construct long time series of economic aggregates, it is generally necessary to piece together several heterogeneous shorter series. Heterogeneity arises even in official national accounting data due to changes in benchmark years, which are often accompanied by methodological changes and by improvements in the quality of primary data sources and in estimation methods. Things are generally worse when we face the task of linking unofficial series constructed by historians and other researchers using incomplete data and different methodologies.

The problem has no easy solution. National statistical institutes can (and sometimes do) help mitigate it by recalculating back series of key aggregates using current methods and criteria in conjunction with detailed source data for earlier periods, but even in this case there is no sure way to know how earlier estimates would have changed if, for instance, new or improved data sources had been available earlier on. From the perspective of independent researchers, such detailed reconstructions are generally out of the question and the only feasible strategy involves the use of simple splicing or linking techniques for pasting together a set of series on a given variable.

The linking procedures commonly used in the literature generally involve the backward extrapolation of the most recent available series using the growth rates of older series (this is what I will call pure *retropolation* for short)[1] or interpolation between the benchmark years of successive series. The basic idea is similar in both cases: we correct the older series so that it matches the newer one at its starting point while retaining some of its features. The nature of the correction is, however, very different in each case. Retropolation preserves the period-by-period growth rates of the older series and places the entire burden of the correction on its levels, while interpolation preserves the starting (or benchmark) level of the older series and adjusts its growth rates as needed. Both procedures rewrite history but they do so in very different ways. As Prados (2006) warns, which method is chosen can make a very big difference, especially when we are dealing with long periods of time.[2]

The linking of economic time series is therefore a delicate exercise that should probably be handled with a bit more care than it often has been exercised in the literature. This note develops a new splicing procedure that may be a useful tool in this regard. The proposed method provides an intermediate option between the two standard methods sketched above that avoids their rather extreme implicit assumptions and allows the researcher to distribute the

[1] As far as I can tell, there is no settled standard terminology in this area. Different expressions (including *backcasting*, *backward projection*, *retropolation* and *back calculation*) are used to refer generically to the backward extrapolation of time series and to specific ways to go about it.

[2] This author analyzes the implications of applying different splicing procedures to Spanish GDP data covering the period 1954-2000. According to his calculations, pure retropolation of the most recent series leads to an upward revision of original GDP estimates for 1954 that exceeds 30%. He observes that such a large correction would significantly alter current views about Spain's relative income level in the mid 20th century in a direction that does not seem entirely plausible.

required correction between the levels and the growth rates of the older series on the basis of what he knows or conjectures about the persistence of the factors that account for the discrepancy between the old and new series that emerges at their linking point. It also derives the time path of the required correction from the assumption that such discrepancies arise from improvements in the coverage of emerging sectors or activities in the basic data underlying the new series. This may not always be the case, but it is certainly one of the usual suspects.

The rest of the note is organized as follows. Section 2 briefly describes the simplest versions of the two standard linking procedures used in the literature and highlights the assumptions implicit in each of them regarding the time profile of the "error" contained in the older series. Section 3 argues that these assumptions are too extreme to be plausible in many cases and introduces a new procedure that can accommodate intermediate situations. Section 4 concludes with a brief illustration of how the proposed procedure has been used in the construction of long employment series for Spain.

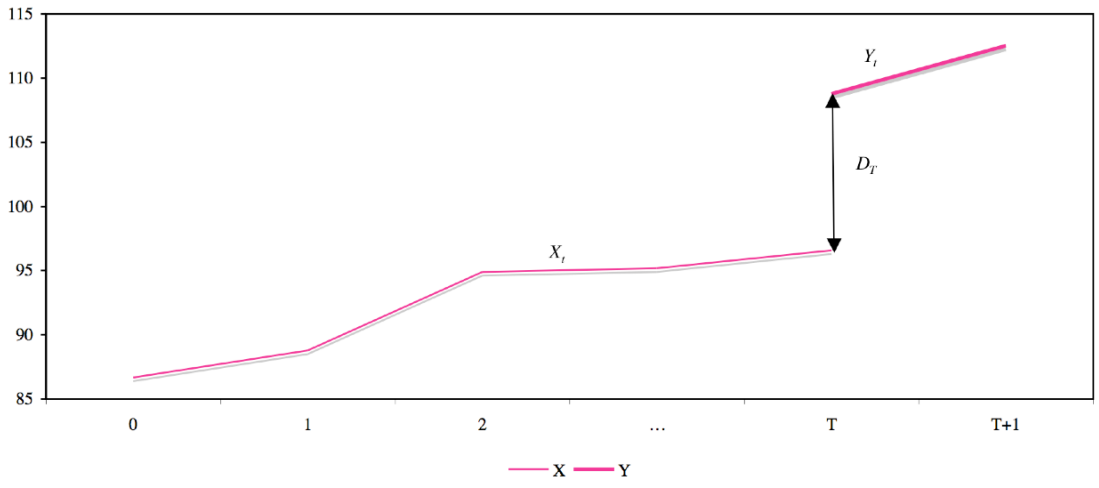
2. Simple splicing procedures: a quick review

Figure 1 illustrates a typical splicing problem. We have two series, X_t and Y_t, referring to the same economic aggregate. The older series, X_t, starts at 0 and extends until T. At this time a new and in principle better series, Y_t, is introduced. As is generally the case, the two series do not agree at their linking point, T. I will denote by D_T the discrepancy between the new and the old series at the linking point and by

$$(1)\ d_T = \ln Y_T - \ln X_T = y_T - x_T$$

the proportional or logarithmic difference between them. It is convenient (although not entirely accurate) to think of d_T as the "measurement error" contained in the older series at the linking point. In principle, this error may affect all terms of the older series but it can only be observed at T.

Figure 1: Two series to be spliced



The problem we face is that of extending the more recent series back to time *0* taking as a reference the older one. As noted above, there are two simple standard solutions that embody alternative hypotheses regarding the time profile of the "measurement error" contained in the older series: retropolation and interpolation.

Retropolation works by extending the new series backward from time T using the growth rates of the old series. As illustrated in Figure 2, the idea is to "raise" the older series by a constant proportion, respecting its time profile, until it matches the new series at the linking point. Using lower case letters to indicate that we are working with logarithms, the retropolation of Y_t taking X_t as a reference will be given by

$$(2)\ \hat{y}_t^r = x_t + (y_T - x_T) \equiv x_t + d_T \quad \text{for } 0 \le t \le T$$

Notice that the spliced series coincides with Y_t at time T and preserves the growth rates of X_t for the period before the linking point, that is

$$(3)\ \Delta\hat{y}_t^r = \Delta x_t \ \text{ for } \ 0 \le t \le T \quad \text{and} \quad \hat{y}_T^r = y_T$$

The implicit assumption is that the "error" contained in the older series

$$(4)\ d_t = y_t - x_t$$

remains constant over time -- that is, that it already existed at time *0* and that its magnitude, measured in proportional terms, has not changed between *0* and T. Hence, in order to recover the "correct" value of the magnitude of interest, all we have to do is add to the older series (measured in logs) the proportional discrepancy between the two series we observe at the linking point, d_T.

In the interpolation method, the series are linked by forcing the backward extension of the new series to go through a given point in the old one, say x_o, which will generally correspond to its base or benchmark year. (See Figure 2). The procedure assumes that the error in the older series has been generated entirely between *0* and T. In its simplest form, it also assumes that the proportional error increases linearly with time. Under these assumptions, the correct value of the variable prior to the linking point can be recovered by adding to the older series (in logs) a linear function of time:

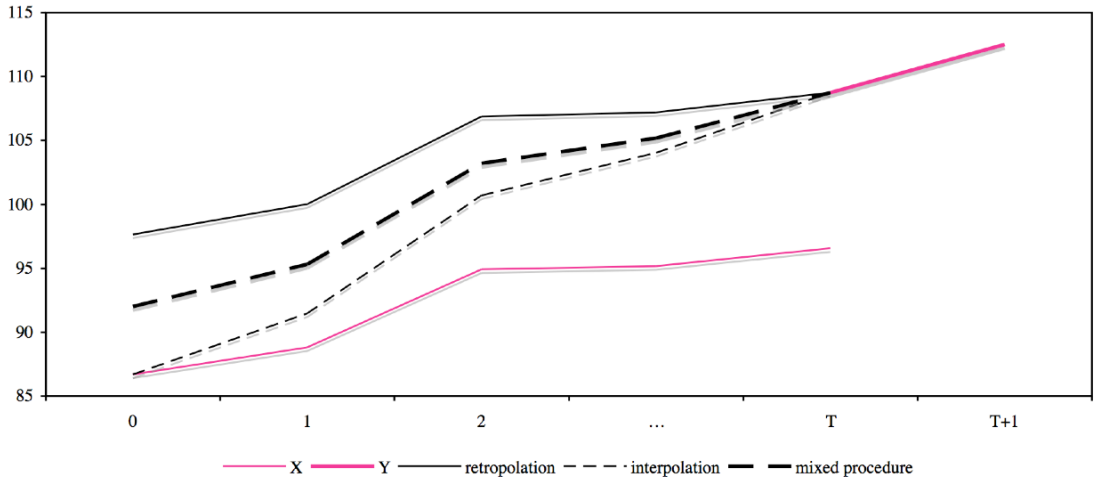
$$(5)\ \hat{y}_t^i = x_t + \frac{t}{T}(y_T - x_T) = x_t + \frac{t}{T} d_T$$

Proceeding in this manner, we preserve the original value of the older series in its base year, but not its growth rates, which are raised by a constant fraction of d_T that depends on the length of the period between the base year of the older series and the linking point,

$$(6)\ \Delta\hat{y}_t^i = \Delta x_t + \frac{1}{T} d_T \ \text{ for } 0 \le t \le T \ \text{ and } \ \hat{y}_o^i = x_o$$

As a result, the time profile of the spliced series can be quite different from that of the older series (see Figure 2).

Figure 2: Alternative splicing procedures



It is useful to note that both procedures are variations on a common theme. In both cases, the linked series is constructed by adding to the older series an estimate of the "error" contained in it that is based on the only direct observation of this magnitude available to the analyst: the one corresponding to the linking point. Hence, the spliced series obtained with procedure *j* will be given by

$$(7)\ \hat{y}_t^j = x_t + \hat{d}_t^j \quad \text{for } 0 \le t \le T \text{ with } \hat{d}_t^j = \begin{cases} \hat{d}_t^r = d_T \\ \hat{d}_t^i = \dfrac{t}{T} d_T \end{cases}$$

3. A *mixed* splicing procedure

When should each of the splicing procedures described in the previous section be used? In many countries, base-year estimates of GDP and other aggregates are built on a substantially more thorough analysis than estimates for non-benchmark years. Other things equal, this would be an important argument in favor of the interpolation procedure, which preserves base-year estimates. On the other hand, things are seldom equal for new base years are often accompanied by improvements in the primary data and in the estimation methods that are used to construct the national accounts.[3] As a result of such improvements, the estimated volume of activity is generally revised upward, presumably because better data and estimation methods allow national accountants to measure more accurately emerging activities or sectors that were not adequately covered in the older series. When this is the case, it seems plausible to conjecture that i) the error that emerges in the new base year already existed to some extent in previous years and will therefore affect the older series in its entirety and ii) that the size of such error has been growing over time because coverage problems tend to be especially severe in emerging sectors that have a growing weight in the aggregate.

[3] An additional source of discrepancies between the two series is the introduction of methodological changes. The most reasonable way to eliminate this type of discontinuity would be to reconstruct the older series using the new methodology prior to splicing it with the new series. I am assuming this has already been done to the extent that it is possible, so that remaining discrepancies between the series are due only to improvements in primary data and in estimation methods.

Hence, what we may expect to be a typical situation following the introduction of a new benchmark will not fit the assumptions that are implicit in standard linking procedures. This suggests that it may be a good idea to develop an alternative splicing method that can accommodate such situations. I will refer to the proposed procedure as the *mixed* splicing method because it will occupy an intermediate position between the two standard methods in the sense that the required correction to the older series will be distributed between its initial level and its growth rates.

One way to describe the difference between the two splicing procedures reviewed in the previous section is in terms of their assumptions concerning the size of the error in the older series at time *0*. Interpolation assumes that this error is zero ($d_o = 0$), while retropolation assumes that it is equal to the error observed at the linking point ($d_o = d_T$). A simple natural way to proceed when neither of these extreme assumptions seems plausible is to parameterize the initial error in a way that can accommodate any intermediate situation. I will assume, in particular, that

$$(8)\ \hat{d}_o^m = \rho d_T$$

where $\rho \in (0,1)$ is a free parameter that measures the magnitude of the initial error in the older series.

The second change I will introduce in the splicing procedure has to do with the time profile of the estimate of the error contained in the older series. In its simplest form, interpolation assumes that this error increases linearly with time. However, this is not the most plausible assumption if, as it often seems likely, the source of the error is the deficient coverage of certain activities whose weight in the aggregate increases over time. In this case, the time parth of the error will depend on the rate of growth of such activities relative to the rest of the economy. If we assume that the ratio between the relevant growth rates is approximately constant, we can model the evolution of the error in a simple way.

I will assume that the error contained in the older series, D_t, is a constant fraction θ of the volume of a set of activities, Z_t, that are deficiently measured. The "real value" of the series of interest will then be given by

$$(9)\ Y_t = X_t + D_t = X_t + \theta Z_t$$

I will also assume that the growth factor of Z_t is a constant multiple of the growth factor of X_t. That is, denoting by G_t the growth factor of X_t,

$$(10)\ \frac{X_{t+1}}{X_t} = G_t$$

I will assume that

$$(11)\ \frac{Z_{t+1}}{Z_t} = \mu G_t$$

where μ is a constant whose value will be determined later on. Letting $z = Z/X$, we obtain a simple difference equation

$$(12)\quad z_{t+1} = \frac{Z_{t+1}}{X_{t+1}} = \frac{\mu G_t Z_t}{G_t X_t} = \mu \frac{Z_t}{X_t} = \mu z_t$$

whose solution is of the form

$$(13)\quad z_t = z_o \mu^t$$

Dividing both sides of (9) by X_t, we have

$$(14)\quad \frac{Y_t}{X_t} = \frac{X_t + \theta Z_t}{X_t} = 1 + \theta z_t$$

Taking logs of this expression, we obtain a convenient approximation for d_t:

$$(15)\quad d_t = \ln \frac{Y_r}{X_t} = \ln\left(1 + \theta z_t\right) \cong \theta z_t$$

Using (13), this expression implies that

$$(16)\quad d_t \cong \theta z_t = \theta z_o \mu^t = d_o \mu^t$$

Next, we can recover the value of μ that is implicit in ρ. Evaluating (16) at time T and recalling that, by assumption, $d_o = \rho d_T$ we have

$$(17)\quad d_T \cong d_o \mu^T = \rho d_T \mu^T$$

where we can solve for μ,

$$(18)\quad 1 \cong \rho \mu^T \;\Rightarrow\; \mu = \left(\frac{1}{\rho}\right)^{\frac{1}{T}}$$

Finally, we substitute (18) into (16) to obtain the following expression for the time path of the proportional error in the older series:

$$(19)\quad d_t \cong d_o \left(\frac{1}{\rho}\right)^{\frac{t}{T}} = \rho d_T \left(\frac{1}{\rho}\right)^{\frac{t}{T}} = d_T \rho^{\frac{T-t}{T}}$$

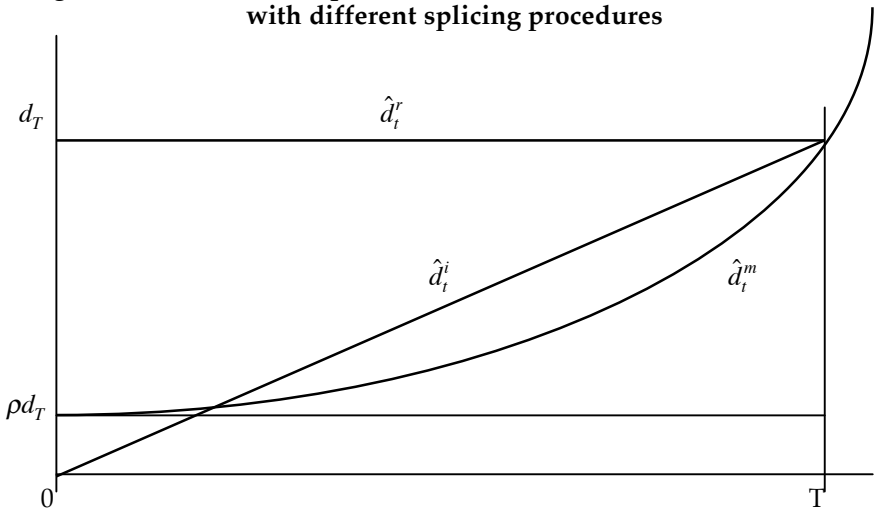
Hence, the spliced series will be given by

$$(20)\quad \hat{y}_t^m = x_t + \hat{d}_t^m \quad \text{for } 0 \le t \le T \text{ with } \hat{d}_t^m = d_T \rho^{\frac{T-t}{T}}$$

As illustrated in Figure 3, equation (19) describes a time path for the estimated error of the older series that is quite different from those derived from the analogous equations for the splicing procedures described above, given in equation (7). Under the hypotheses of this section, the error (d_t) will not be constant or a linear function of time but will grow at an increasing rate because the weight of the emerging activities that presumably are not well covered in the old

series will increase as we approach the linking year. This implies that the appropriate correction to the growth rate of the original series will not be constant but increasing in time and will be an increasing function of the parameter ρ that captures our hypothesis regarding the severity of the initial error.

Figure 3: Estimated time path of the error contained in the older series with different splicing procedures



The one decision the analyst has to make when using the mixed splicing procedure is that of assigning a value to the parameter ρ that measures the severity of the error in the older series at time *0*. While this is not generally an observable magnitude, the analyst may have access to external information that can be used to approximate its value. In this connection, it may be useful to note that the mixed procedure implicitly fixes the initial value of the spliced series, $\hat{y}_o^m$, as a weighted average of the values at time *0* of the older series and the series obtained by pure retropolation,

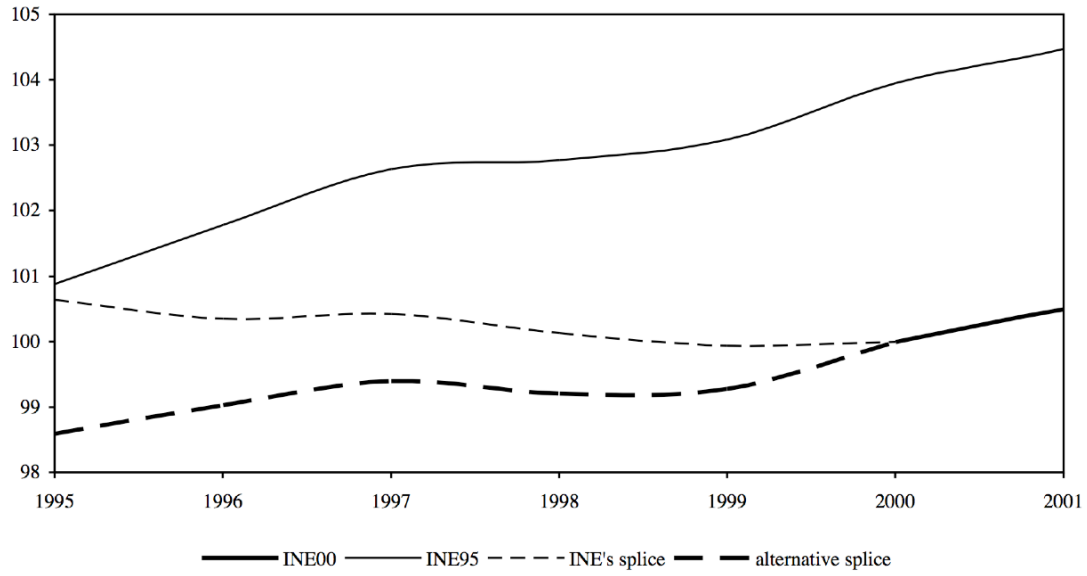
$$(21)\quad \hat{y}_o^m = x_o + \hat{d}_o^m = x_o + \rho d_T = (1+\rho-\rho)x_o + \rho d_T = (1-\rho)x_o + \rho(x_o + d_T) = (1-\rho)x_o + \rho\hat{y}_o^r$$

Hence, a comparison of the initial values of these two series with some outside estimate of the relevant magnitude at 0 or some nearby year may allow us to assess the relative plausibility of these two estimates and help us set the value of ρ. In the absence of outside estimates for time 0, it may still be possible to formulate a plausible conjecture about the value of ρ on the basis of whatever is known about the factors that account for the discrepancy between the two series at the linking point. While this is not an ideal situation, it seems preferable to having to opt between the two extreme assumptions implicit in the procedures discussed above. At the very least, the proposed procedure brings out in the open a problem that both of the standard methods hide – that we don't know the time profile of the error in the older series—and allows us to try to make a reasonable guess on the basis of whatever information may be available.

4. Some illustrations

A few comments on a specific application may perhaps be useful as an illustration of how the proposed procedure may be applied and of its potential advantages. The Spanish Statistical Institute (INE) has recently published new series of national and regional accounts that take as a benchmark the year 2000 (INE00) and a spliced series that extends this series back to 1995, which was the base year for the previous official series.[4] The spliced series has been constructed (essentially) by interpolation in order to respect the original estimate for 1995, which the INE considers quite reliable due to its "structural character" (INE, 2007). However, comparison of the two series in their common year of 2000 reveals a sizable upward revision of employment, with an increase of 7.55% in the estimated number of jobs. Since this discrepancy is distributed by the INE over only 5 years, the time profile of its spliced employment series is very different from that of the older one and not entirely plausible. As shown in Figure 5, for instance, INE's spliced series implies a reduction in average labor productivity between 1995 and 2000, while the previous official series (INE95) showed a positive, although modest, increase in this variable during the same period.

Figure 4: Average labor productivity in Spain
constant prices of 2000 (100 = 2000 in INE00)



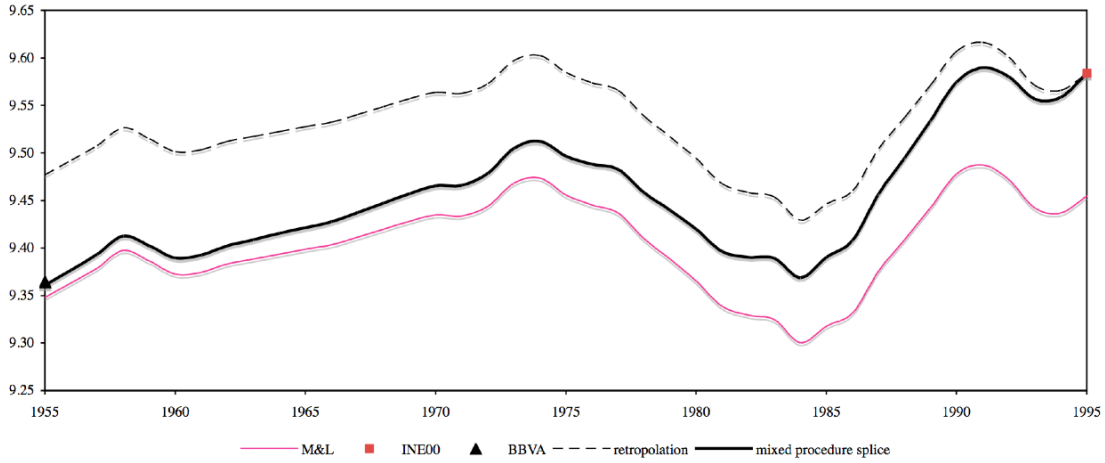
The upward revision of employment seems to be due mostly to improvements in the design of the Spanish Labor Force Survey (EPA) that have allowed a more accurate measurement of the number of part-time workers.[5] It seems hard to argue that problems with the measurement of part-time jobs did not arise until 1995, but it is probably true that the severity of the problem has increased over time as part-time work has become more frequent in Spanish society. Hence, neither of the extreme assumptions made by the standard linking procedures discussed in section 2 seems to be appropriate in this case. As far as I know, there are no hard data available

[4] Both series are available at INE (2009).
[5] See for instance Albacete and Laborda (2005).

that can be used to set the value of ρ. It seems likely, however, that the error due to this problem was only marginally smaller in 1995 than in 2000, which points to a value of ρ only slightly below 1. Figure 4 compares INE's spliced productivity series (INE's splice) with an alternative one that has been constructed by the mixed procedure with a ρ of 0.9 (see de la Fuente, 2009a). While there is absolutely no guarantee that this is the correct value of the parameter, the correction does at least yield a spliced series with a rather more plausible time profile.

A second illustration involves the projection of the Spanish employment series back to the mid 1950s. Pure retropolation of my spliced 1995-2000 series, using as a reference the historical series constructed by Maluquer and Llonch (M&L, 2005) working with national accounts data with benchmarks in 1986 and earlier years, would lead to a 13% upward revision of these authors' original estimate of Spanish employment in 1955, as illustrated in Figure 5.

Figure 5: Employment series for Spain
logarithmic scale



In order to determine whether such a large revision may be plausible, we need to analyze the sources of the discrepancy that existed between the two series in 1995 (which is carried back unchanged to 1955 by the retropolation procedure). It turns out that the break in the series in 1995 is due to two factors of approximately equal weight. One is the already mentioned change in the methodology of the Labor Force Survey that has led to the "emergence" in 2000 of a large number of part-time jobs held mostly by women-- 90% of which have been carried back to 1995 in my spliced series. The other is a change in the concept of employment used in the Spanish National Accounts, which supplied data on the number of employed workers until the 1995 base was introduced and switched at this point to the number of jobs.

To set the value of ρ we need to try to establish how relevant these two factors (part time employment and workers holding more than one job) were in the mid 50s. Fortunately, there are some outside sources that can be used to shed some light on this question. One is the 1950 Census, which tells us that only 0.55% of active workers declared to have a "secondary occupation" on top of their primary job (INE, 1950, volume II, Table V). The second is an independent estimate of the number of jobs in the year of interest constructed by the research department of a large Spanish bank (FBBV, 1999), which is considerably closer to M&L's

original estimate than to the retropolated series. Hence, the available information suggests that most of the "error" in the older series was not there in 1955 and points therefore to a low value of ρ. More informal evidence also points in the same direction, as we know that female labor force participation was much lower a few decades back and that the use of part-time contracts was much less common than it is now. On the basis of these considerations, I have chosen a value of 0.10 for ρ, obtaining the spliced series that is shown in Figure 5.[6]

References

Albacete, R. and A. Laborda (2005). "Cambios en la Encuesta de Población Activa y en la Contabilidad Nacional." *Cuadernos de Información Económica* 186, mayo-junio, pp. 44-55.

de la Fuente, A. (2009a). "Un enlace alternativo de los agregados de VAB y empleo de la CRE95 y la CRE00." Mimeo, Instituto de Análisis Económico, CSIC, Barcelona. http://www.fedea.es/030_Publicaciones_Principal.asp?z=3

de la Fuente, A. (2009b). "Series enlazadas de algunos agregados económicos nacionales y regionales, 1955-2007. Versión 2.1." Mimeo, Instituto de Análisis Económico, CSIC, Barcelona. http://www.fedea.es/030_Publicaciones_Principal.asp?z=3

Fundación BBV (FBBV, 1999). *Renta nacional de España y su distribución provincial. Serie homogénea. Años 1955 a 1993 y avances 1994 a 1997.* Bilbao.

Instituto Nacional de Estadística (INE, 1950). *Censo de la Población de España y territorios de su soberanía y protectorado según el empadronamiento realizado el 31 de diciembre de 1950.* In INEbase, electronic database: Demografía y población. Cifras de población y censos demográficos. Censos de población desde 1900. http://www.ine.es/inebaseweb/71807.do?language=0

Instituto Nacional de Estadística (INE, 2007). "Contabilidad regional de España, base 2000. Serie homogénea 1995-2006. Nota metodológica." In INEbase electronic database: Economía: Cuentas Económicas: Contabilidad Regional de España. Madrid. http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft35%2Fp010&file=inebase&L=0

Instituto Nacional de Estadística (INE, 2009). Contabilidad Regional de España. In INEbase electronic database: Economía: Cuentas Económicas: Contabilidad Regional de España. Madrid. http://www.ine.es/jaxi/menu.do?type=pcaxis&path=%2Ft35%2Fp010&file=inebase&L=0

Maluquer, J. and M. Llonch (2005). "Trabajo y relaciones laborales." En A. Carreras y X. Tafunell, coordinadores. *Estadísticas históricas de España, siglos XIX-XX*, second edition. Fundación BBVA, Bilbao, pp. 1155-1245.

Prados de la Escosura, L. (2006). "Assessing the bias in spliced GDP series: evidence from Spain's National Accounts, 1954-2000." Mimeo, Universidad Carlos III de Madrid.

[6] See de la Fuente (2009b) for additional details.