Private returns to R&D in the presence of spillovers, revisited – Online appendix

Giovanni Millo

Abstract

This is the Online Appendix to Millo (2018), containing the full results, tables and comments. The accompanying software and data, together with other Open Source software from the R Project, allows for the full replication of this document.

Keywords: panel time series, R.

This document is the Online Appendix to Millo (2018), in turn a replication and critical extension of Eberhardt, Helmers, and Strauss (2013) (henceforth EHS). As such, it presents both a full replication of the original paper EHS and a number of alternative or supplementary results, most of which were not included in Millo (2018) because of space constraints.

In the following sections, we first summarize the analysis in EHS, to set the stage for the replication and the extension of their work. Next, we highlight the main findings of the study. The subsequent section details the full replication, listing all the results' tables. Computational details and Conclusions follow.

This document, together with the original EHS data (a copy of which is included, with the kind consent of the authors, in the R package **pder**), with the accompanying software and Open Source software from the R Project (R Core Team 2014), is self-contained and entirely replicable, as detailed in Section 4.

1. Summary

The empirical analysis in EHS regards the estimation of static and dynamic versions of the following production function. The basic specification as in EHS, Formula (4) is defined as

$$y_{it} = \alpha l_{it} + \beta k_{it} + \gamma r_{it} + \lambda_t + \psi_i + \varepsilon_{it}$$

where y, l, k and r are, respectively, the logs of production and of the three inputs: labour, capital and research and development.

The estimators employed can be categorized in two ways: according to the hypotheses on the technology parameters β (homogeneous or heterogeneous); and, analogously, according to those on the impact of unobservables, as summarized in EHS, Table 4.

The main empirics of the paper are contained in four tables (originally Tables 5-8 in EHS):

Table 5 "Pooled Production Functions (Static)" compares the following static homogeneous specifications: pooled OLS with time fixed effects (POLS), two-ways fixed effects (2FE), first differences with time effects (FD), and common correlated effects pooled *a la Pesaran*

(2006) (CCEP) with or without the explicit inclusion of time fixed effects. Coefficients and standard errors are reported. The constant returns to scale hypothesis (CRS), i.e. $H_0 : \alpha + \beta + \gamma = 0$ is tested based on the above. Diagnostics are included for overall fit (root mean square error), serial correlation (m_p *test* Arellano and Bond 1991), cross-sectional correlation (CD test Pesaran 2004) and unit roots int he residuals ($CIPS$ test Pesaran 2007).

Table 6 "Pooled Production Functions (Dynamic)" compares autoregressive distributed lags (ARDL) dynamic versions of the five above specifications, substituting FD with the generalized method of moments (GMM) dynamic panel estimator of Arellano and Bond (1991) in the "system" version of Blundell and Bond (1998) (BB). The full dynamic models are not reported; in their place estimates of the long-run coefficients are reported in two possible versions according to the results of a common factor (COMFAC) test: either *unrestricted* long-run coefficients, with standard errors calculated by the Delta method, or *restricted* versions where COMFAC restrictions are applied *ex post* through a minimum distance procedure. Diagnostics as in Table 5 are also reported, regarding the unrestricted ARDL specifications.

Table 7 "Heterogeneous Production Functions (Static)" presents four heterogeneous parameter specifications: mean groups *a la* Pesaran and Smith (1995) (MG), the same on cross-sectionally demeaned data (CDMG), and two versions of Pesaran (2006)'s common correlated effects mean groups (CMG) with or without the inclusion of individual country trends. Next to the CRS Wald statistic, diagnostics as above are reported, substituting the serial correlation $m_p - test$ with a Fisher-type combination of individual Ljung-Box tests.

Table 8 "Heterogeneous Production Functions (Dynamic)" compares four dynamic specifications as in the above Table 6, this time reporting only restricted versions of the four models, which all pass the COMFAC test. The usual diagnostics, again regarding the unrestricted ARDL versions, are reported, with the exception of a serial correlation test.

We omit the (unproblematic) descriptive statistics tables (EHS Table 3) and concentrate on the estimation results and tests reported in Tables 5, 6, 7 and 8 in EHS, Section VI.

This replication exercise consists in:

- reproducing Tables 5-8 as faithfully as possible using the original dataset but different software
- suggesting alternatives to estimators and tests where we see it appropriate and providing our modified versions of said tables, and commenting on the consequences on the economic results of the paper
- as the conclusions of EHS suggest, providing directions for a first assessment of the direction of spillovers based on a more sophisticated analysis of cross-sectional and spatial dependence in the original models.

The following Sections, one dedicated to the main findings and the next to the full replication, are therefore both organized into subsections according to the structure of EHS, i.e. they contain one subsection for each table 5 to 8 in EHS. The Higlights section begins with a general *caveat* on the issues arising when estimating dynamic specifications with CCE.

2. Highlights

Before going through the full replication, we provide a section summarizing the main findings of the replication.

Estimating dynamic models by CCE There are issues with dynamic CCE models which were not known at the time EHS was published. Recent research showed that neither the CCEMG (Chudik and Pesaran 2015) nor the CCEP estimator (Everaert and De Groote 2016) are consistent for fixed T in the context of a dynamic model with persistent common factors.

A partial solution has been proposed by Chudik and Pesaran (2015) for the dynamic CCEMG, consisting in augmenting the model with a number of lags of the cross-sectional averages equal at least to the (unknown) number of common factors, with a practical rule of $T^{\frac{1}{3}}$. Nevertheless, while the augmented CCEMG estimates of the mean slopes of the regressors are quite robust even to small T and N, the mean coefficient of the lagged dependent variable still suffers from severe bias unless T is sufficiently large with respect to N. According to the simulations by the authors, small sample biases are substantial even for T as large as 50. Moreover, the EHS dataset is highly unbalanced, making the calculation of many lags of the cross-sectional averages problematic. As for the CCEP estimator, Everaert and De Groote (2016) show that the dynamic CCEP estimator is inconsistent for fixed T in a persistent common factors scenario, the bias being again substantial up to $T = 50$.

Given that consistently estimating the coefficient on the dynamic term is essential to the calculation of the long-run solutions presented in EHS, we conclude that the EHS dataset is simply much too "short" for employing dynamic CCE estimators.[1]

Pooled production functions, static (EHS Table 5) In the original work, the coefficient on R&D is of comparable magnitude (between 0.5 and 1) across all specifications and highly significant with the only exception of the FD model. Residuals are nonstationary for POLS and 2FE, while – surprisingly – cross-sectional correlation is only detected in the CCEP and, marginally, in the CCEPt models. Constant returns to scale are rejected in the POLS and CCEPt specifications. EHS conclude that POLS and 2FE are seriously misspecified, CCEP and CCEPt fail to address cross-sectional dependence and FD is the preferred model on grounds of its favourable diagnostics.

For the SEs of the **POLS** model, despite the evidence of serial correlation, EHS use the heteroskedasticity correction of White (1980). A clustered covariance *a la* Arellano (1987) would be in order.[2] Clustering by individual yields dramatically higher SEs, although all coefficients are still significant. The same applies to the **2FE** model, where employing clustered SEs R&D is not significant any more.

In the **FD** specification, again, SEs should, in our view, be clustered by group. Nevertheless, despite this omission the SEs in EHS are of comparable magnitude: differencing seems to have

[1] We thank an anonymous referee for pointing out these recent results.

[2] By contrast, the CRS test reported in EHS is ostensibly computed using the clustering-robust covariance.

effectively eliminated the persistence of errors within each individual cluster. The replication is otherwise successful but for some small differences (a missing dummy).

The **CCEP** and **CCEPt** estimators are obtained in EHS by interacting the cross-sectional average of the regressand and of each explanatory variable with every individual dummy, for a total of at least $K + N + (K + 1) \times N = 598$ regressors. This procedure allows to perform CCEP estimation with a standard regression, but causes numerical instability ad unpredictable behaviour because of the large number of dummies and interaction terms. While this method still comes very close to the parameter estimates one obtains from coding the estimator in matrix form, it does not provide the correct SEs. Employing the formula from the original CCE paper (Pesaran 2006, Th.3), the latter are much wider than those reported by EHS. A thorough selection of alternative estimates is reported (later in the paper) in Table 3 for the sake of comparison.

Controlling for clustering, both homogeneous models accounting for individual heterogeneity (FE2, FD) now again have SEs of magnitude comparable to that of the CCEP models, provided the latter are also computed according to Pesaran (2006). The main economic result is that, unlike the original findings of EHS, own R&D expenditure is consistently insignificant as soon as one controls for any kind of individual heterogeneity. This reconciles the findings of EHS, Table 5 with those of the dynamic and/or heterogeneous models in EHS, Tables 6-8. See further below, as Table 2, our preferred version of EHS, Table 5.

Summing up, Table 5 in EHS can be entirely replicated. Yet there are arguably better ways to estimate the standard errors, leading to wider confidence intervals. The consequences on the economic interpretation are that, when accounting at least for individual effects – i.e. in any model bar POLS – R&D is never significant any more and constant returns to scale (CRS) are never rejected any more. All in all, the results from the different models are much more easily reconciled than in the original contribution. Substantial residual cross-sectional dependence supports the use of CCEP over FD, but the results are qualitatively very similar. Local dependence diagnostics suggest geographical (within-country) proximity as the main direction for the spillovers.

Diagnostics EHS employ m_p tests for residual serial correlation (Arellano and Bond 1991), which assume cross-sectional independence. We substitute them with the procedure in Wooldridge (2010a), running a (panel) autoregression on the model residuals and evaluating the significance of the AR coefficients; which test, importantly, can also be robustified as appropriate, e.g. against cross-sectional correlation (Wooldridge 2010a, 10.5.4). Qualitative conclusions do not change. Unit root tests are in turn all successfully replicated and the conclusions unchanged.

As regards cross-sectional dependence diagnostics, the CD test is well known to lose power if the data are cross-sectionally demeaned, which centers the average of correlation coefficients $\bar{\rho}$ on zero (see Sarafidis, Yamagata, and Robertson 2009, 2.2). In this case, from test results one cannot tell whether pairwise correlation coefficients are generally low in absolute value or rather if positive and negative coefficients are compensating. As a first indication, we compare the average of correlation coefficients with that of their absolute values. Then we perform the rank-based sibling of the CD test, the Frees test (Frees 1995), which does not share this weakness.

The $\bar{\rho}$ is near zero for all models, which as expected affects the power of the CD test. By contrast, the average absolute correlation $|\bar{\rho}|$, although bigger for POLS and 2FE, is of sizeable magnitude for all models, hinting at substantial cross-sectional dependence which is actually detected by the Frees rank-based test. Unlike EHS, we conclude that cross-sectional dependence is present in all models (see Table 5).

The latter can, *a priori*, be of either *global* (pervasive, factor-type) or *local* (decaying, spatial-type) nature. Notice that the CCEP estimator is a) designed to control for common factor dependence, and b) consistent in the presence of spatial dependence. It is therefore no wonder that cross-sectional dependence be left in the CCEP residuals, as long as it is due to local correlation and not to a pervasive factor structure. [3] In order to check for this, we apply the exponent of cross-sectional dependence defined in Bailey, Kapetanios, and Pesaran (2016) to the CCEP residuals. The results (reported later in the paper in Table 6) are consistent with the hypothesis that no pervasive factor structure remains in the residuals. The residual spatial dependence is instead discussed in the following. [4]

In the light of the conclusions of EHS on the importance of identifying the direction of spillovers from R&D, we assess the degree of *local* cross-sectional dependence between "neighbouring" observations with respect to two naturally defined "proximity" dimensions in the EHS data: country and industry. As a first step, we apply the local, or spatial, CDp test (Pesaran 2004, Section 7); the null hypothesis of independence is consistently rejected for either definition of space, and across all specifications. It must nevertheless be observed that global dependence related to unobserved common factors can cause the CDp test to reject as well (Holly, Pesaran, and Yamagata 2010; Moscone and Tosetti 2010; Millo 2017); therefore the latter can be safely used only on the residuals from estimators which do control for common factors. Moreover, neither the CD nor the CDp test tolerate serial correlation, which seems indeed present in all the static models of EHS.

Factor-robust testing procedures are compared and discussed at length in Millo (2017), where a new test (RW) based on randomizing the neighbourhood relationships in the proximity matrix is proposed. Differently from the CDp, the RW test can be safely applied to any of the specifications considered by EHS; and, interestingly in view of the above concerns, the RW test is also insensitive to serial correlation. Its pseudo-p values (symmetric version, 1000 random draws; hence 0.002 is the minimum) are reported in the last row of the table. From this evidence we conclude that all models have *locally* correlated residuals within single countries; and, but the evidence is much weaker, probably also within industrial sectors.

Pooled production functions, dynamic (EHS Table 6) EHS estimate the unrestricted dynamic ARDL models and then, based on diagnostics from the latter, resent either the unrestricted long-run estimates (Table 6A in EHS) or, for those models that pass the COMFAC test, the restricted ones (Table 6B in EHS). Only the 2FE passes the restriction. POLS and BB raise suspicions of nonstationarity; moreover, BB gives unrealistic estimates and shows signs of cross-sectional dependence, which would invalidate it. The remaining models, (restricted) 2FE and (unrestricted) CCEP, all estimate the effect of R&D next to zero. CRS is never rejected. As above, we argue that SEs for the **POLS** and **2FE** models should at a

[3] We thank an anonymous reviewer for raising this point.

[4] The bias-corrected version of $\mathring{\alpha}$ (Bailey *et al.* 2016, Eq. 13), allowing for persistent factors but no spatial correlation$\mathring{\alpha}$ is estimated at 0.616, its 95% confidence bands at $(0.574, 0.658)$. All other versions are computed with largely similar results, reported in Table 6. The exponent is not computable for the residuals of models containing time fixed effects, therefore it is here applied only to the CCEP.

minimum be corrected for clustering by `id`. Doing this, SEs increase substantially.

The system-GMM specification of Blundell and Bond (1998) (**BB**)is estimated by EHS according to the two-step method, using the popular correction by Windmeijer; all regressors are used as GMM-type instruments. As EHS observe, the **BB** estimator is generally inappropriate for the problem at hand. Despite limiting the number of lags to avoid instrument proliferation, numerical problems emerge: Stata drops some instruments because of collinearity and, most importantly, resorts to a generalized inverse. Our R replication is close but not exact [5].

The dynamic **CCEP** estimator, again, is obtained by augmenting the specification with interactions of individual dummies and cross-sectional averages of regressand and regressors, yielding $7 + 119 + 8 \times 119 = 1078$ variables which become 1103 if adding $T - 1 = 25$ time effects. All considerations from the static case apply. The results can be replicated in R with reasonable accuracy, given how ill-conditioned the problem is. Again, though, the SEs reported are incorrect, the ones from the original CCE formula (Pesaran 2006) being much wider. This has the important consequence that the COMFAC test, which rejected based on the exceedingly narrow SEs in EHS, does now not reject.

Table 6 in EHS can be entirely replicated; but the considerations on Table 5 apply here as well, in particular regarding estimation of standard errors. Revised accordingly, the CCEP model passes the COMFAC test and therefore the restricted estimates of the long-run parameters are to be reported. The CCEP model thus migrates from Panel A of Table 6 in EHS to Panel B together with its special case, 2FE. Although important differences remain (above all, in the estimated productivity of capital), and taking heed of the general issues with dynamic CCEPs mentioned above, the main conclusions to be drawn from the two preferred models are now similar.

Heterogeneous production functions, static and dynamic (EHS Tables 7-8) Mean groups estimators are less problematic that the pooled ones in the Tables 5-6 of EHS, also thanks to the availability in Stata of a user-contributed function, `xtmg`, written by one of the authors (Eberhardt 2012). All results in Table 7, EHS are replicated exactly, with one small exception[6].

The same goes for estimates and diagnostics from the dynamic model (EHS, Table 8) [7].

3. Replication

In this Section we first attempt at replicating EHS as faithfully as possible, then we comment on the appropriate points and provide alternatives, again following the main structure of EHS with one subsection dedicated to each Table from 5 to 8.

[5] The R function `pgmm` in turn resorts to a generalized inverse in the second step. Given the numerical difficulties, we did not expect to replicate results exactly; yet the two implementations are not far from each other either in terms of coefficients or of standard errors.

[6] In the Ljung-Box-Fisher serial correlation tests, we find serial correlation in all models, where for the CMG and CMGt EHS report a strange p-value of 1.00 which is likely to be a typo.

[7] Here, the only discrepancy is due to a typo: the results of the CRS test reported in EHS are actually those computed on the *un*restricted models. We replicated them exactly as such; the results are not reported here but the code can be found in the replication script as chunk *t8.crstests.u*. Computing the CRS tests on the restricted models, constant returns to scale are never rejected.

3.1. Pooled production functions, static (EHS Table 5)

In this subsection we reproduce the results from static production functions with homogeneous coefficients, as in EHS, Table 5.

	POLS	2FE	FD	CCEP	CCEPt
ln Lit	0.464	0.608	0.635	0.562	0.582
	40.946	18.944	18.085	20.714	21.002
ln Kit	0.465	0.487	0.279	0.289	0.203
	37.802	10.908	3.431	7.946	4.972
ln Rit	0.096	0.063	0.045	0.084	0.064
	22.923	4.544	1.698	4.925	3.662
CRS	0.000	0.001	0.614	0.077	0.000
AB-AR(1)	0.000	0.000	0.005	0.001	
AB-AR(2)	0.000	0.000	0.188	0.682	
CD	0.116	0.143	0.110	0.010	0.055
CIPS	0.100	0.100	0.010	0.010	0.010
RMSE	0.277	0.161	0.061	0.052	0.052

Table 1: Static homogeneous models (Table 5 in EHS). POLS: pooled OLS with time FEs; 2FE: two-way fixed effects; FD: first differences with time FEs; CCEP: pooled common correlated effects. CRS: Wald test for H_0 of constant returns to scale; AB-AR(1) and (2): Arellano-Bond (1991) m_1 and m_2 tests for residual serial correlation; CIPS: Pesaran (2007) unit root test: H_A is stationarity, upper limits are reported for the p-value. RMSE: root mean square error.

Observations on EHS, Table 5

Pooled model with time effects (POLS) In the POLS model, which is actually estimated with the addition of time fixed effects, EHS use Stata with the `robust` option, which corresponds to the heteroskedasticity correction of White (1980). Given the panel nature of the dataset, and for homogeneity with the other specifications, a clustered covariance *a la* Arellano (1987) would be in order. Clustering by individual yields dramatically higher standard errors, although all coefficients are still significant.

Two-ways fixed effects (2FE) As above (POLS), standard errors are computed according to the White method without clustering, hence are not robust to intra-group serial correlation (which is detected by the AB test). Again, we argue that they should at a minimum be corrected for clustering by `id`. Doing this, standard errors increase fourfold as for POLS. R&D is not significant any more.

Interestingly, the reported p-value of the CRS test is obtained using the clustering-robust covariance (and, as such, is replicated in the modified table instead of the original one). This behaviour is probably due to having tested the CRS restriction after (re)estimating the 2FE model as individual fixed effects plus time dummies, in which case the 'robust' option produces clustering-robust SEs a la Arellano (1987); while the original model was estimated as pooled

	POLS	2FE	FD	CCEP
ln Lit	0.464	0.608	0.646	0.562
	11.839	5.567	16.543	6.379
ln Kit	0.465	0.487	0.262	0.289
	11.294	3.057	2.771	1.802
ln Rit	0.096	0.063	0.045	0.084
	6.704	1.351	1.447	1.239
CRS	0.068	0.338	0.674	0.675
W-AR(1)	0.000	0.000	0.000	0.000
W-AR(2)	0.000	0.000	0.082	0.189
Frees	0.000	0.000	0.000	0.000
CIPS	0.100	0.100	0.010	0.010
RMSE	0.277	0.161	0.061	0.052

Table 2: Static homogeneous models, modified table. t-statistics in parentheses, constructed from clustered (POLS, 2FE, FD) viz. nonparametric (CCEP) standard errors. W-AR(p) test: Wooldridge (2010) test for H_0 of no residual serial correlation at the p-th order, robustified; Frees (1995) test for H_0 of cross-sectionally independent residuals; CRS, CIPS, RMSE: see preceding Table

OLS with both individual and time dummies: in this case producing the heteroskedasticity-consistent estimator in the sense of White (1980) (see the replication code by the authors). This issue testifies on the need to carefully check the consistency of the default behaviour of statistical software.[8]

First difference model (FD) The FD specification in EHS includes time dummies.

A small typo in the Stata code for the FD model originally made results hard to reproduce. Given the initial observation for each individual which is lost in differencing, EHS omitted *two* time dummies instead of one; but then Stata ostensibly also dropped one itself. Hence the exact results of EHS can be replicated in R by explicitly adding time effects for periods 3 to 26. The effect on estimation is minimal.

Again, standard errors should, in our view, be clustered by group. Interestingly, doing so does not inflate the standard errors: the t-ratios are of comparable magnitude with those in EHS, Table 5. Differencing seems to have effectively eliminated the persistence of errors within each individual cluster.

Common correlated effects pooled (CCEP, CCEPt) In EHS, the CCEP estimator, which allows for heterogeneous effects of common factors, is obtained interacting the cross-sectional average of the regressand and of each explanatory variable with every individual dummy, for a total of $K + N + (K + 1) \times N = 3 + 119 + 4 \times 119 = 598$ regressors; and estimating this specification by OLS.

This procedure allows to perform CCEP estimation with standard regression software; but, because of the large number of dummies and interaction terms, it is heavy on the machine

[8] In R the same would have happened if estimating one model with `lm` and the other with `plm`; while using `plm` for both, with the different options 'pooled' and 'within', would have provided consistent results.

and numerically unstable: in fact, some dummies are dropped by Stata because of extreme collinearity. Even worse in the case where 25 time dummies are added too (CCEPt). In fact, collinearity between the time dummies and the cross-sectional averages is to be expected. Regression software will generally manage the ensuing collinearity in soem automatic way, either by dropping variables or by employing a generalized inverse, which can entail a loss of control by the researcher over what is actually done. For this reason we have dropped the CCEPt from our revised estimates altogether.

Actually, this method comes very close to the parameter estimates one obtains from coding the estimator in matrix form, at least unless time dummies are also added. We replicated it in R and report the results. Yet, as it turns out, it does not directly provide the correct standard errors, because not all of the added terms are orthogonal to the explanatory variables lnL, lnK, lnR: in fact, their cross-sectional averages are not. Employing the formula from the original CCE paper (Pesaran 2006), the standard errors are much wider than those reported by EHS.

Controlling for clustering, both homogeneous models accounting for individual heterogeneity (FE2, FD) now again have standard errors of magnitude comparable to that of the CCEP models, provided the latter are also computed according to Pesaran (2006).

The main economic result is that, unlike the original findings of EHS, own R&D expenditure is consistently insignificant as soon as one controls for any kind of individual heterogeneity. This reconciles the findings of Table 5 with those of the dynamic and/or heterogeneous models in Tables 6-8.

Computing the standard errors of CCEP estimators We compare the standard errors and resulting t-statistics for significance from the nonparametric estimator proposed in Pesaran (2006, 5.2, Th. 3) (henceforth *nonparametric SEs*) with two sets of alternatives:

- *parametric*: “sandwich” estimates in the spirit of White (1980) and Arellano (1987) based on different types of “meat”
- *bootstrapped*: estimates based on various types of clustered resampling procedures

It should be borne in mind that nonparametric SEs are based on looser assumptions than parametric ones. The former are justified by Pesaran (2006, Th. 3) as requiring $(N, T) \rightarrow \inf$ jointly, in no particular order; the latter - at least in the Newey-West form, by Pesaran (Th. 4 in 2006) which assumes $T/N \rightarrow \inf$ as $(N, T) \rightarrow \inf$ jointly, so that the sandwich estimator is appropriate only for relatively “short” panels. By contrast, the theoretical justification of the nonparametric estimator from Pesaran (Th. 3 in 2006) becomes problematic under the homogeneity hypothesis $\beta_i = \beta \forall i$. If the homogeneity hypothesis $\beta_i = \beta \forall i$ holds (Pesaran 2006, Th. 4), and if $N >> T$, the covariance matrix of $\hat{\beta}_{CCEP}$ can be consistently estimated by the sandwich-type estimator in Pesaran (2006, Formula 74), where the “meat” of the sandwich is calculated according to a panel version of the Newey and West estimator (Pesaran 2006, Formulae 51-52).

According to Pesaran (2006) and simulation evidence therein, nevertheless, the nonparametric estimator performs well in general, homogeneity scenario included, and therefore it is the suggested alternative. We compare the nonparametric SEs with three parametric alternatives of the sandwich type, differing by the calculation method for the “meat”: White-Arellano, Newey-West and its generalization, robust to cross-sectional dependence, the Driscoll and Kraay "SCC" estimator.

Bootstrap methods are useful when parametric standard errors are difficult to derive; or they can improve on the finite-sample properties of known parametric estimators. Lastly, they can be useful when the parametric SEs are based on some regularity assumptions (homoskedasticity, incorrelation) which are not met. In the general case of clustered sandwich estimators, the latter are well known and easily computable; moreover, they do already accommodate heteroskedasticity and (groupwise) correlation of unknown form; so the need for bootstrapping reduces to the case in which the small sample properties of the parametric alternative are problematic, i.e., when the number of clusters is too small (Cameron, Gelbach, and Miller 2008). In the case of EHS, judging by the simulation evidence in Cameron *et al.* (2008), the number of clusters is large enough to make the asymptotic refinement of bootstrap-t methods redundant. By contrast, cluster bootstrapping, while accommodating factor-related dependence, still relies on an assumption of spatial independence, which is mimicked in the independent resampling of individuals (Kapetanios 2008, p. 11 WP). In the case of the CCEP estimator, available cluster-bootstrap methods are no panacea because they rely on similar assumptions on independence, sample size and parameter homogeneity as their parametric counterparts. In the specific case at hand, they are substantially useless because the number of clusters is big enough for the clustered sandwich estimator of the standard errors to be reliable; hence they do simply reproduce its results in a computationally more intensive way.

	ln Lit			ln Kit			ln Rit		
EHS	0.027	0	***	0.036	0	***	0.017	0	***
Nonparametric	0.088	0	***	0.161	0.07148	.	0.068	0.21526	
HC (cluster by firm)	0.045	0	***	0.077	0.00017	***	0.033	0.01023	*
Newey-West	0.031	0	***	0.045	0	***	0.02	3e-05	***
Driscoll-Kraay	0.042	0	***	0.076	0.00015	***	0.019	1e-05	***
Wild bootstrap	0.045	0	***	0.077	0.00018	***	0.034	0.01285	*
Pairs bootstrap	0.047	0	***	0.078	0.00021	***	0.035	0.01473	*
Wild t-bootstrap (HC)		0	***		0.002	**		0.02	*
Pairs t-bootstrap (HC)		0	***		0	***		0.016	*

Table 3: Original (EHS) vs. alternative standard errors for the CCEP model from Table 5 in EHS, with corresponding significance diagnostics. Nonparametric: as in Pesaran 2006, Th.3; HC, NW and DK: 'sandwich' estimates as in Pesaran 2006, Th.4, with the 'meat' calculated by the respective methods; bootstrap ($M = 999$): 'pairs' or 'wild' bootstrap estimates, clustered by firm, as defined in CGM; t-bootstrap: 'pairs' or 'wild' resampling of cluster-robust t-statistics. For each variable, first column: standard error; second column: p-value for the t-test.

General remarks on diagnostics The Arellano and Bond (1991) m_p tests for residual serial correlation are only appropriate if residuals are independent across individuals. In the light of the above discussion, this hypothesis is unlikely to hold here. Hence we propose to substitute the m_1 and m_2 tests presented in EHS, Table 5 with the procedure in Wooldridge (2010b), running a (panel) autoregression on the model residuals and evaluating the significance of the AR(p) coefficients for $p = 1, 2$. Such procedure allows both testing each order of serial correlation individually and also employing a robust covariance matrix, e.g. to account for cross-sectional correlation.

As Wooldridge (2010b) notes, the procedure is to be adjusted for the induced correlation from transforming the data, which under the null hypothesis of no serial correlation in the original errors is, respectively, -0.5 at the first order and 0 at the second for the FD model (Wooldridge 2010b, 10.6.3) (see also Drukker *et al.* 2003); and $-\frac{1}{1-T}$ at any order for the FE one (Wooldridge 2010b, Eq. 10.52). The serial incorrelation test for the original errors thus can be performed as a t-test for this serial correlation coefficient (Wooldridge 2010b, 10.5.4).

This induced correlation is another reason to be suspicious of the results of the m_{1-2} tests: as EHS (p. 444) observe, AR(1) is to be expected in the FD errors; unsurprisingly, we add, the errors of the 2FE model turn out correlated at both orders because of the equicorrelation induced by the FE transformation. Nevertheless, substituting the Wooldridge-type procedure for the original m tests does not qualitatively change the conclusions regarding residual serial correlation in any of the models. The results are reported in Table 4.

	POLS	2FE	FD	CCEP
W AR(1), HC	0.000	0.000	0.000	0.000
W AR(2), HC	0.000	0.000	0.082	0.189
W AR(1), SCC	0.000	0.000	0.000	0.000
W AR(2), SCC	0.000	0.000	0.028	0.397

Table 4: Wooldridge's informal test for serial correlation, with either HC (Arellano) or SCC (Driscoll-Kraay) covariance; H0: no serial correlation at given order in untransformed errors.

Unit root tests are in turn all successfully replicated and the conclusions unchanged.

As regards cross-sectional dependence diagnostics, the CD test is well known to lose power if the data are cross-sectionally demeaned, which centers the average of correlation coefficients $\bar{\rho}$ on zero (see Sarafidis *et al.* 2009, 2.2).[9] In this case, from test results one cannot tell whether pairwise correlation coefficients are generally low in absolute value or rather if positive and negative coefficients are compensating. As a first indication, we compare the average of signed correlation coefficients with that of their absolute values. Then, again for comparison purposes, we perform the rank-based sibling of the CD test, the Frees test (Frees 1995), which does not share this weakness.

As discussed in the Highlights section, we assess the degree of *local* cross-sectional dependence between "neighbouring" observations with respect to country and industry, applying the local, or spatial, CDp test (Pesaran 2004, Section 7) and the RW test of Millo (2017). Results are reported in a dedicated Table 5.

We also apply the exponent of cross-sectional dependence defined in Bailey *et al.* (2016) to the CCEP residuals. Different flavours of the bias-corrected version of $\mathring{\alpha}$ (Bailey *et al.* 2016, Eq. 13), possibly allowing for either persistent factors or/and spatial correlation, are computed and presented below in Table 6. For computing the spatial version, missing data are interpolated through the Beckers and Rixen (2003) procedure through the `dineof` function from the **sinkr** package (Taylor 2017) (see also Taylor, Losch, Wenzel, and Schröter 2013).

[9] This explains the somewhat strange finding that models accounting for cross-sectional dependence in a simplistic way through time dummies (POLS, 2FE, FD) turn out less affected than more sophisticated CCEP ones.

	POLS	2FE	FD	CCEP
Global CD	-1.570	-1.464	-1.598	2.588
	0.116	0.143	0.110	0.010
avg. rho	-0.004	-0.005	-0.005	0.005
avg. |rho|	0.503	0.502	0.218	0.263
Frees rank test	21.959	22.575	-8.308	-2.013
	0.000	0.000	0.000	0.000
CD(1), sector	8.631	8.677	2.910	3.937
	0.000	0.000	0.004	0.000
CD(1), country	24.584	22.579	16.409	20.092
	0.000	0.000	0.000	0.000
RW, sector	0.002	0.002	0.008	0.026
RW, country	0.002	0.002	0.002	0.002

Table 5: Comparison of cross section dependence diagnostics for the models in Table 5. Global CD is Pesaran's (2015) test; rho is the mean of pairwise correlation coefficients; $|rho|$ the mean of their absolute values; Frees (1995) is the rank test for cross-sectional dependence; CD(1) sector and, respectively, country are the local CD tests (Pesaran, 2004) when defining neighbourhood as sharing either the same industrial sector or the same country. RW is Millo's (2017) randomization test for spatial correlation robust to common factors (pseudo-p values from 999 random draws are reported).

Although the estimate of $\mathring{\alpha}$ is slightly higher than 0.5, we still take the results as suggesting that no pervasive factor structure remains in the residuals (a short discussion in Millo 2018).

	a.o.lwr95	a.o	a.o.upr95	N.alfa	N
None	0.578	0.616	0.654	18	119
Persistent factors, NW	0.574	0.616	0.658	18	119
Persistent factors, AR	0.565	0.616	0.667	18	119
Spatial correlation (reduce)	0.624	0.667	0.710	24	119
Spatial correlation (interpolate)	0.636	0.662	0.725	26	119
Both (NW + reduce)	0.619	0.667	0.714	24	119

Table 6: Exponent of cross-sectional dependence estimated on the CCEP residuals, allowing for (in row order): no persistence or spatial residual correlation; persistence, Newey-West style variance estimator; persistence, variance estimated through AR(4); spatial correlation in remainder errors, data balanced by reduction; spatial correlation, balanced by interpolation; both persistent factors (NW) and spatial correlation (by reduction). In column order: lower 95 percent bound, central estimate and upper bound for the bias-corrected version of the exponent; estimated number of significant factor loadings; total number of cross-sectional units (memorandum item).

3.2. Pooled production functions, dynamic (EHS Table 6)

EHS proceed to estimate the unrestricted dynamic ARDL models and then, based on diagnostics from the latter, present either the unrestricted long-run estimates (Table 6A in EHS) or, for those models that pass the COMFAC test, the restricted ones (Table 6B in EHS). Only the 2FE passes the restriction. POLS and BB raise suspicions of nonstationarity; moreover, BB gives unrealistic estimates and shows signs of cross-sectional dependence, which would invalidate it. The remaining models, (restricted) 2FE and (unrestricted) CCEP, all estimate the effect of R&D next to zero. CRS is never rejected.

As done above with the static models, we now proceed to assessing local dependence in the dimensions of countries and sectors.

Results suggest that local cross-sectional dependence is substantial, and remains in the residuals of CCEP estimators as well (although these are consistent in its presence). This is true for both within sector and within country dependence. As for the intensity (measured by the absolute value of the CD statistics), it is worth noticing that: within-sector is weaker, and is rather successfully accounted for by time fixed effects, suggesting that the factor loadings of different firms to sector-wide shocks are reasonably homogeneous; within-country dependence is generally stronger, and by contrast to the former it is better accounted for by CCEP models, suggesting a greater deal of heterogeneity in factor loadings.

As EHS proceed to estimate the unrestricted dynamic ARDL models and then, based on diagnostics from the latter, present either the unrestricetd long-run estimates (Table 6A in EHS) or, for those models that pass the COMFAC test, the restricted ones (Table 6B in EHS), not being bound by compactness requirements, we here present first the dynamic ARDLs together with the relevant diagnostics, in order to assess their statistical properties as done in EHS; then, in the following table(s), either the unrestricted or the restricted estimates with the relevant CRS test.

Observations on EHS, Table 6

Pooled model with time effects (POLS) Standard errors for the POLS model are computed according to the White method without clustering, hence are not robust to intra-group serial correlation (which is detected by the AB test). We argue that they should at a minimum be corrected for clustering by `id`. Doing this, standard errors increase fourfold.

Two-ways fixed effects (2FE) Standard errors for 2FE are computed as for POLS. Again, we argue that they should at a minimum be corrected for clustering by `id`. Doing this, standard errors increase fourfold as for POLS.

System GMM model (BB) The System-GMM specification of Blundell and Bond (1998) is estimated by EHS according to the two-step method, using the popular correction by Windmeijer; all regressors are used as GMM-type instruments. Proliferation of instruments is clearly an issue here, given the sizeable time dimension (REF.: Baltagi?); to control it, EHS limit the number of lags used for instrumentation to 3 and collapse the instrument matrix as suggested in Roodman (see refs. therein). Numerical problems nevertheless emerge, Stata dropping some instruments because of collinearity and, most importantly, resorting to

	POLS	2FE	BB	CCEP	CCEPt
ln Yit-1	0.981	0.934	0.935	0.408	0.402
	220.715	126.800	24.293	14.560	14.527
ln Lit	0.660	0.660	0.564	0.670	0.663
	20.409	19.682	3.052	13.056	16.251
ln Lit-1	-0.654	-0.617	-0.616	-0.452	-0.445
	-20.388	-18.284	-3.147	-8.934	-25.345
ln Kit	0.208	0.062	0.175	0.503	0.503
	4.554	1.085	0.726	5.238	0.154
ln Kit-1	-0.205	-0.046	-0.082	-0.285	-0.331
	-4.525	-0.823	-0.347	-3.302	-0.102
ln Rit	0.078	0.018	-0.025	0.020	0.020
	3.879	0.748	-0.220	0.503	0.006
ln Rit-1	-0.069	-0.019	0.040	0.020	0.019
	-3.469	-0.796	0.354	0.504	0.006
COMFAC	0.000	0.688	0.030	0.000	0.000
CD	0.125	0.097	0.000	0.229	0.891
CIPS	0.010	0.010		0.010	0.010
RMSE	0.060	0.055	0.075	0.034	0.032

Table 7: Dynamic homogeneous models. Replication of unreported table, basis for Table 6 in EHS. Diagnostics denoted as in the preceding tables.

a generalized inverse. The R function `pgmm` in turn resorts to a generalized inverse in the second step. Given the numerical difficulties, we did not expect to replicate results exactly; yet the two implementations are not far from each other either in terms of coefficients or of standard errors.

Common correlated effects pooled (CCEP, CCEPt) *The considerations in the following only regard replication of the original dynamic CCEP models in EHS. They must be seen in the light of the "Estimating dynamic models by CCE" paragraph on page 3, which have a more general bearing on the statistical appropriateness of the whole estimation.*

The dynamic CCEP estimator, again, is obtained by augmenting the specification with interactions of individual dummies and cross-sectional averages of regressand and regressors. In the dynamic model, there are 7 regressors, so this yields an even larger design matrix, with $7 + 119 + 8 \times 119 = 1078$ variables, which become 1103 if adding $T - 1 = 25$ time effects.

All considerations from the static case apply. The results can be replicated in R with reasonable accuracy, given how ill-conditioned the problem is. Again, though, the standard errors reported are incorrect, the ones from the original CCE formula (Pesaran 2006) being much wider. This has the important consequence that the COMFAC test, which rejected based on the exceedingly narrow SEs in EHS, does now not reject.

Diagnostics As done above with the static models, we assessed local dependence in the dimensions of countries and sectors. Results (not shown, to be found in the Online Appendix) suggest that local cross-sectional dependence is substantial, and remains in the residuals of

	POLS	2FE	BB	CCEP
ln Yit-1	0.981	0.934	0.935	0.393
	96.077	39.039	24.293	2.942
ln Lit	0.660	0.660	0.564	0.687
	18.045	16.061	3.052	5.639
ln Lit-1	-0.654	-0.617	-0.616	-0.459
	-17.707	-15.698	-3.147	-3.137
ln Kit	0.208	0.062	0.175	0.493
	2.687	0.739	0.726	1.870
ln Kit-1	-0.205	-0.046	-0.082	-0.311
	-2.617	-0.557	-0.347	-1.191
ln Rit	0.078	0.018	-0.025	-0.022
	2.471	0.528	-0.220	-0.192
ln Rit-1	-0.069	-0.019	0.040	0.055
	-2.246	-0.565	0.354	0.423
COMFAC	0.000	0.750	0.030	0.427
Frees	0.000	0.000	0.000	0.000
CIPS	0.010	0.010		0.010
RMSE	0.060	0.055	0.075	0.034

Table 8: Dynamic homogeneous models, modified table. Diagnostics denoted as in the preceding tables.

CCEP estimators as well (although these are consistent in its presence). This is true for both within sector and within country dependence. As for the intensity (measured by the absolute value of the CD statistics), it is worth noticing that: within-sector is weaker, and is rather successfully accounted for by time fixed effects, suggesting that the factor loadings of different firms to sector-wide shocks are reasonably homogeneous; within-country dependence is generally stronger, and by contrast to the former it is better accounted for by CCEP models, suggesting a greater deal of heterogeneity in factor loadings.

Again, Table 6 in EHS can be entirely replicated; but considerations on Table 5 apply here as well, in particular regarding estimation of standard errors. As a consequence, the revised CCEP model passes the COMFAC test and therefore the restricted estimates of the long-run parameters are to be reported. The CCEP model thus migrates from Panel A of Table 6 in EHS to Panel B together with its special case, 2FE. Although important differences remain (above all, the estimate of the productivity of capital), the main conclusions to be drawn from the two preferred models are now similar.

In the following we report both our replication of the original, and our preferred modified versions, of the two panels A and B of Table 6 in EHS.

	POLS	2FE	BB	CCEP	CCEPt
ln Lit	0.338		-0.792	0.369	0.364
	2.495		-0.927	5.256	4.996
ln Kit	0.173		1.409	0.367	0.287
	0.869		2.041	3.746	2.644
ln Rit	0.462		0.222	0.066	0.064
	2.788		1.594	1.668	1.600
CRS	0.593		0.868	0.033	0.008

Table 9: Dynamic homogeneous models, unrestricted long-run solutions (Replication of Table 6, Panel A in EHS).

	POLS	2FE	BB	CCEP	CCEPt
ln Lit		0.654			
		19.761			
ln Kit		0.078			
		1.447			
ln Rit		0.019			
		0.815			
CRS		0.000			

Table 10: Dynamic homogeneous models, long run solutions imposing COMFAC restriction (Replication of Table 6, Panel B in EHS).

	POLS	2FE	BB	CCEP
ln Lit	0.338		-0.792	
	2.495		-0.927	
ln Kit	0.173		1.409	
	0.869		2.041	
ln Rit	0.462		0.222	
	2.788		1.594	
CRS	0.593		0.868	

Table 11: Dynamic homogeneous models, unrestricted long-run solutions, modified version of Panel A.

	POLS	2FE	BB	CCEP
ln Lit		0.654		0.646
		19.761		5.425
ln Kit		0.078		0.355
		1.447		1.600
ln Rit		0.019		0.011
		0.815		0.139
CRS		0.020		0.962

Table 12: Dynamic homogeneous models, long run solutions imposing COMFAC restriction, modified version of Panel B.

3.3. Heterogeneous production functions, static (EHS Table 7)

Uniform demeaning techniques (CDMG) seem to account for within-sector dependence just as well as do models allowing for heterogeneous factor loadings. On the contrary, within-country dependence is still stronger in the CDMG residuals than in CMG ones. The former situation is consistent with an assumption of uniform effect of factors, which is actually more reasonable for technological progress within a particular industry than for the effect of idiosyncratic national factors.

	MG	CDMG	CMG	CMGt
ln Lit	0.568	0.557	0.599	0.698
	6.569	7.628	9.000	8.236
ln Kit	0.117	0.445	0.244	0.149
	0.955	5.008	1.702	1.004
ln Rit	-0.058	0.089	0.035	-0.050
	-0.728	2.123	0.445	-0.601
CRS	0.000	0.092	0.468	0.277
Ljung-box AR(1)	0.000	0.000	0.000	0.000
CD	0.000	0.046	0.505	0.347
CIPS	0.010	0.010	0.010	0.010
RMSE	0.051	0.068	0.037	0.035

Table 13: Static heterogeneous models (Table 7 in EHS). Diagnostics denoted as in the preceding tables, but for Ljung-Box AR(1) which is a Fisher-type combination of individual Ljung-Box tests for serial correlation.

	MG	CDMG	CMG	CMGt
Global CD	22.668	1.992	0.667	0.940
	0.000	0.046	0.505	0.347
avg. rho	0.059	0.005	0.002	0.003
avg. |rho|	0.226	0.240	0.223	0.228
Frees rank test	-5.527	-4.363	-5.715	-5.486
	0.000	0.000	0.000	0.000
CD(1), sector	9.569	4.129	5.035	4.051
	0.000	0.000	0.000	0.000
CD(1), country	20.660	16.568	12.830	12.507
	0.000	0.000	0.000	0.000
RW, sector	0.002	0.008	0.002	0.004
RW, country	0.002	0.002	0.002	0.002

Table 14: Comparison of cross section dependence diagnostics for the models in Table 7. Global CD test, rho, $|rho|$, Frees rank test, CD(1) sector and, respectively, country, and RW randomization test defined as in Table eftab:xsd.

Observations on EHS, Table 7

Mean groups estimators are much less problematic that the pooled ones in the Tables 5-6 of EHS, also thanks to the availability in Stata of a user-contributed function, `xtmg`, written by one of the authors (Eberhardt 2012).

All results in Table 7, EHS have been replicated exactly (at least up to the decimals reported), with the only exception of the Ljung-Box-Fisher serial correlation tests. The latter are combination-type tests whose null hypothesis is that none of the individual time series of residuals are serially correlated, hence they can be expected to be very conservative. In fact, in our replication, serial correlation is found in all four models; on the contrary, for the CMG and CMGt models EHS report a strange p-value of 1.00 which is likely to be a typo.

3.4. Heterogeneous production functions, dynamic (EHS Table 8)

	MG	CDMG	CMG	CMGt
ln Yit-1	0.153	0.470	-0.022	-0.046
	1.658	10.954	-0.519	-0.992
ln Lit	0.690	0.563	0.642	0.671
	5.723	9.519	9.073	8.843
ln Lit-1	-0.237	-0.350	-0.094	-0.037
	-2.730	-5.928	-1.307	-0.463
ln Kit	0.278	0.184	0.295	0.229
	1.616	1.950	1.749	1.326
ln Kit-1	-0.146	0.000	-0.261	-0.264
	-0.647	0.002	-1.648	-1.609
ln Rit	-0.096	0.139	-0.086	-0.094
	-0.701	3.961	-0.665	-0.726
ln Rit-1	-0.032	-0.082	0.046	0.010
	-0.168	-2.234	0.263	0.054
COMFAC	0.716	0.480	0.969	0.849
CD	0.000	0.065	0.078	0.061
CIPS	0.010	0.010	0.010	0.010
RMSE	0.035	0.038	0.022	0.020

Table 15: Dynamic heterogeneous models.

	MG	CDMG	CMG	CMGt
ln Lit	0.703	0.567	0.642	0.678
	6.152	10.011	9.386	9.432
ln Kit	0.277	0.245	0.276	0.172
	1.867	3.373	1.709	1.088
ln Rit	-0.107	0.139	-0.084	-0.088
	-0.953	3.947	-0.945	-0.964
CRS	0.469	0.540	0.274	0.142

Table 16: Dynamic heterogeneous models, long run solutions imposing COMFAC restrictions (Table 8 in EHS).

Observations on EHS, Table 8

The considerations in the following only regard replication of the original dynamic CCEMG models in EHS. They must be seen in the light of the "Estimating dynamic models by CCE" paragraph on page 3, which have a more general bearing on the statistical appropriateness of the whole estimation.

As observed, MG estimators in Stata match those in R very well. All diagnostics from the dynamic model are replicated exactly. All models pass the COMFAC test, thus only restricted versions of the long-run models are reported in Table 8, EHS. The latter is in turn exactly replicated in R, but for the last row. Here, the only discrepancy is due to a typo: the results of the CRS test reported in EHS are actually those computed on the *un*restricted

	MG	CDMG	CMG	CMGt
Global CD	12.336	1.843	-1.761	-1.875
	0.000	0.065	0.078	0.061
avg. rho	0.033	0.005	-0.006	-0.006
avg. |rho|	0.227	0.225	0.211	0.220
Frees rank test	-7.799	-7.826	1.557	2.019
	0.000	0.000	0.000	0.000
CD(1), sector	6.102	4.023	2.938	3.053
	0.000	0.000	0.003	0.002
CD(1), country	11.262	10.567	6.307	6.346
	0.000	0.000	0.000	0.000
RW, sector	0.016	0.002	0.006	0.010
RW, country	0.002	0.002	0.002	0.002

Table 17: Comparison of cross section dependence diagnostics for the models in Table 8. Diagnostics defined as in earlier Tables.

models[10]. Computing the CRS tests on the restricted models, constant returns to scale are never rejected.

4. Computational details

This appendix has been produced as a dynamic document with the Sweave utility (Leisch 2002) in order to be replicable in the sense of Peng (2011): as such, it is self-contained and fully self-reproducing without the need for any further resources other than the files provided with it and FOS software from the R project (R Core Team 2014).

All calculations were originally performed in R 3.0.2 (R Core Team 2014), in particular using packages **plm** 1.3-3 (Croissant and Millo 2008), **lmtest** (Zeileis and Hothorn 2002), **car** (Fox and Weisberg 2011), **msm** (Jackson 2011) and **sinkr** (Taylor 2017), on the following systems: Ubuntu Linux 16.04, Windows 7. Results were identical across platforms. Tables have been automatically typeset in LaTeX by the **xtable** package (Dahl 2014). Full replication scripts are available in the accompanying materials to this Appendix, together with extended versions of some published functions from packages **plm** and **splm**, as well as some entirely new procedures (most of them meant to be included in said packages in the future).

4.1. Data sources and materials

Data and procedures from the original paper have been retrieved from the personal page of Markus Eberhardt, `www.medevecon.com`. They are available from the data archive of the Review of Economics and Statistics as well. With the kind agreement of the authors, the dataset has also been included in the **pder** package for R (Croissant and Millo 2017) (a companion package to the forthcoming book by the same name), whence they can be retrieved with `data(RDSpillovers, package="pder")`.

[10]We replicated them exactly as such; the results are not reported here but the code can be found in the replication script as chunk *t8.crstests.u*.

4.2. Computational robustness checks

In the following we discuss the possibility of errors in the R code, and the replication of non-matching results with different procedures and/or different software as a robustness check. All relevant scripts are available upon request. Results are discussed in order of appearance.

While our R implementation of the minimum distance estimator for the COMFAC restricted models is a port of the Stata routines of Mans Soderbom (`www.soderbom.net/Resources.htm`), the Arellano-Bond m-tests have been rewritten from scratch starting from the original paper, instead of porting the `abar` Stata routine by Doorman. The former in fact replicates exactly the original results, while the latter does not. Our implementation of the m-tests, both m_1 and m_2, has been checked for consistency by means of a small Monte Carlo simulation, in which it turned out to produce test statistics that are correctly sized and have the expected power properties as the serial correlation coefficient at, respectively, the first and second order goes from 0 to near 1 (we tested the following values: 0, 0.1, 0.2, 0.3, 0.5 and 0.99, at a confidence level of 5%, obtaining an empirical size in the region of 0.04 and monotonically increasing power).

5. Conclusions

The paper by EHS is a rich piece of empirical work which exposes a number of issues whose interest goes beyond that of the application at hand to bear more generally on the practice of modelling panel time series with common factors.

We have successfully replicated all procedures in the paper in open source R with only minor discrepancies. Next to the replication of original results, we have provided alternatives for estimators and diagnostics for some problematic cases. [11] The most important aspect is perhaps our critique of the "augmentation" approach to computing CCEP estimators. While very convenient and applicable through standard regression, the technique yields grossly underestimated standard errors. Nevertheless, the economic conclusions of the paper are still upheld. Our corrected estimates actually reconcile some of the original findings which were at odds with each other. Constant returns to scale (CRS) restrictions are now not rejected throughout all preferred specifications of EHS, hereby confirming the consistency of their results beyond the original presentation, where heterogeneous common-factor augmented models (CMG, CMG with trends) rejected CRS while all other preferred models (homogeneous static: FD, homogeneous dynamic: CCEP, heterogeneous static: CMG, CMG with trends) did not.

As for the cross-sectional correlation which is such a substantial feature of the original study, we highlight the shortcomings of relying on CD tests in presence of time fixed effects and other cross-sectional augmentations, and suggest alternatives. We assess the effectiveness of defactoring in CCEP models computing the exponent of cross-sectional (residual) dependence. Lastly, we provide a first assessment of the direction of spillovers by means of local CDp and RW tests over the dimensions of countries and industries, finding that correlation is highest

[11] The availability of the original dataset *and* Stata code, published by the authors, made this replication and critique of the main results in the original paper possible; in particular, the availability of code clarified the procedures used far beyond what could be gauged from the paper itself, a testimony to the importance of complementing published results with *linked* data and *executable* code, in the spirit of the "gold standard of replication" (Peng 2011).

within each country.

References

Arellano M (1987). "Computing Robust Standard Errors for Within Estimators." *Oxford bulletin of Economics and Statistics*, **49**(4), 431–434.

Arellano M, Bond S (1991). "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations." *The Review of Economic Studies*, **58**(2), 277–297.

Bailey N, Kapetanios G, Pesaran MH (2016). "Exponent of Cross-Sectional Dependence: Estimation and Inference." *Journal of Applied Econometrics*, **31**(6), 929–960.

Beckers JM, Rixen M (2003). "EOF calculations and data filling from incomplete oceanographic datasets." *Journal of Atmospheric and oceanic technology*, **20**(12), 1839–1856.

Blundell R, Bond S (1998). "Initial conditions and moment restrictions in dynamic panel data models." *Journal of econometrics*, **87**(1), 115–143.

Cameron AC, Gelbach JB, Miller DL (2008). "Bootstrap-based improvements for inference with clustered errors." *The Review of Economics and Statistics*, **90**(3), 414–427.

Chudik A, Pesaran MH (2015). "Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors." *Journal of Econometrics*, **188**(2), 393–420.

Croissant Y, Millo G (2008). "Panel data econometrics in R: The plm package." *Journal of Statistical Software*, **27**(2), 1–43.

Croissant Y, Millo G (2017). *pder: Panel Data Econometrics with R*. R package version 1.0-0, URL `https://CRAN.R-project.org/package=pder`.

Dahl DB (2014). *xtable: Export tables to LaTeX or HTML*. R package version 1.7-4, URL `http://CRAN.R-project.org/package=xtable`.

Drukker DM, *et al.* (2003). "Testing for serial correlation in linear panel-data models." *Stata Journal*, **3**(2), 168–177.

Eberhardt M (2012). "Estimating panel time-series models with heterogeneous slopes." *Stata Journal*, **12**(1), 61–71.

Eberhardt M, Helmers C, Strauss H (2013). "Do spillovers matter when estimating private returns to R&D?" *Review of Economics and Statistics*, **95**(2), 436–448.

Everaert G, De Groote T (2016). "Common correlated effects estimation of dynamic panels with cross-sectional dependence." *Econometric Reviews*, **35**(3), 428–463.

Fox J, Weisberg S (2011). *An R Companion to Applied Regression*. Second edition. Sage, Thousand Oaks CA. URL `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`.

Frees EW (1995). "Assessing cross-sectional correlation in panel data." *Journal of econometrics*, **69**(2), 393–414.

Holly S, Pesaran MH, Yamagata T (2010). "A spatio-temporal model of house prices in the USA." *Journal of Econometrics*, **158**(1), 160–173.

Jackson CH (2011). "Multi-State Models for Panel Data: The msm Package for R." *Journal of Statistical Software*, **38**(8), 1–29. URL `http://www.jstatsoft.org/v38/i08/`.

Kapetanios G (2008). "A bootstrap procedure for panel data sets with many cross-sectional units." *The Econometrics Journal*, **11**(2), 377–395.

Leisch F (2002). "Sweave: Dynamic generation of statistical reports using literate data analysis." In *Compstat*, pp. 575–580. Springer.

Millo G (2017). "A simple randomization test for spatial correlation in the presence of common factors and serial correlation." *Regional Science and Urban Economics*, **66**, 28–38.

Millo G (2018). "Private Returns to R&D in the Presence of Spillovers, Revisited." *Journal of Applied Econometrics*.

Moscone F, Tosetti E (2010). "Testing for error cross section independence with an application to US health expenditure." *Regional Science and Urban Economics*, **40**(5), 283–291.

Peng RD (2011). "Reproducible research in computational science." *Science (New York, Ny)*, **334**(6060), 1226–1227.

Pesaran MH (2004). "General diagnostic tests for cross section dependence in panels."

Pesaran MH (2006). "Estimation and inference in large heterogeneous panels with a multifactor error structure." *Econometrica*, **74**(4), 967–1012.

Pesaran MH (2007). "A simple panel unit root test in the presence of cross-section dependence." *Journal of Applied Econometrics*, **22**(2), 265–312.

Pesaran MH, Smith R (1995). "Estimating long-run relationships from dynamic heterogeneous panels." *Journal of econometrics*, **68**(1), 79–113.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Sarafidis V, Yamagata T, Robertson D (2009). "A test of cross section dependence for a linear dynamic panel model with regressors." *Journal of econometrics*, **148**(2), 149–161.

Taylor M (2017). *Collection of functions with emphasis in multivariate data analysis*. R package version 0.6, URL `https://github.com/marchtaylor/sinkr`.

Taylor MH, Losch M, Wenzel M, Schröter J (2013). "On the sensitivity of field reconstruction and prediction using empirical orthogonal functions derived from gappy data." *Journal of Climate*, **26**(22), 9194–9205.

White H (1980). "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica: Journal of the Econometric Society*, pp. 817–838.

Wooldridge JM (2010a). *Econometric analysis of cross section and panel data*. MIT press.

Wooldridge JM (2010b). *Econometric analysis of cross section and panel data*. MIT press.

Zeileis A, Hothorn T (2002). "Diagnostic Checking in Regression Relationships." *R News*, **2**(3), 7–10. URL http://CRAN.R-project.org/doc/Rnews/.

Affiliation:

Giovanni Millo
Group Insurance Research, Generali SpA
Via Machiavelli 3
34132 Trieste (Italy)
Telephone: +39/040/671184
Fax: +39/040/671160
E-mail: Giovanni.Millo@Generali.com