EXPORT DIVERSIFICATION: A NONPARAMETRIC EXAMINATION

CHRIS PAPAGEORGIOU AND CHRISTOPHER F. PARMETER

Abstract. Export diversification has recently been found as a key component of economic development. This has spawned an empirical literature which documents a quadratic relationship in the development-diversification nexus. Using cutting edge nonparametric panel data estimators and inferential tools, we examine the shape of this nexus using several different measures of diversification. We question the robustness of the quadratic shape of the development-diversification nexus that has dominated the applied trade diversification literature. Accounting for time specific heterogeneity reveals that this relationship is tenuous. Additional parametric robustness checks further lend support that taking the quadratic relationship as a stylized fact may not be advisable. Finally, we deploy both nonparametric panel estimators and consistent model specifications tests and present further evidence that the formulation of the development-diversification nexus is not apt to be characterized as quadratic.

1. Introduction

Economic theory suggests that export concentration is optimal, and yet we see a wide array of diversification behavior across countries and time. Consistent with the seminal work of Imbs & Wacziarg (2003), as countries migrate through different stages of development, trade diversification takes on differing levels of importance. Recently, Klinger & Lederman (2006), Parteka (2007) and Cadot, Carrère & Strauss-Kahn (2011) empirically investigate the presence, or lack thereof, of a quadratic, 'Kuznets' type relationship between export diversification (more appropriately concentration) and economic output (measured as GDP per capita). If the development story of Imbs & Wacziarg (2003) holds then one expects for low levels of income, countries are highly concentrated, then countries diversify as they grow, and once they reach a certain point of development they reconcentrate to exploit

International Monetary Fund and University of Miami
Date: March 14, 2018.
Chris Papageorgiou, International Monetary Fund; e-mail: cpapageorgiou@imf.org
Christopher F. Parmeter, Corresponding Author, Department of Economics, University of Miami; e-mail: cparmeter@bus.miami.edu.
* We benefited from discussions with Olivier Cadot and Nikola Spatafora. The views expressed in this study are the sole responsibility of the authors and should not be attributed to the International Monetary Fund, its Executive Board, or its management. This document was compiled in `knitr` using RStudio 0.98.977.

their natural comparative advantages. An important policy concern is where this level of development is.

Given the concentrate-diversify-reconcentrate argument, a quadratic-type relationship between income per capita and trade diversification should emerge. The preeminent study on this topic is Cadot et al. (2011) who use an unbalanced panel of over 140 countries, across four different measures of trade diversification and estimate a turning point in the development-diversification nexus around $25,000 which is remarkably consistent across the four measures and robust to alternative econometric approaches which they take to the data. This finding is qualitatively consistent across the literature, dating back to Imbs & Wacziarg (2003).

An important extension of the quadratic shape of the diversification-development nexus is that an additive decomposition is available that allows one to determine if diversification is occurring along the intensive (new products-new trade partners) or extensive (higher volumes) margin (this stems from the fundamental finding of Cadot et al. (2011) when the Theil index is used as the measure of diversification). This decomposition is remarkable in its simplicity and ability to shed deeper insight into key policy issues regarding trade diversification.

With the importance of trade in the global economy, the study of the trade diversification-income nexus is paramount. Here we reexamine the existence of the shape of this relationship. We do this across several dimensions. First, we replicate the initial estimates of Cadot et al. (2011). Next, we argue that the quadratic relationship that is taken as a benchmark in this literature is fragile and present evidence along many dimensions which reinforces this view.

We begin in this endeavor by considering several alternative modeling exercises of the preferred quadratic parametric specification of the field. This is then supplemented using a new dataset that exists at a higher level of granularity to assess if the manner in which diversification is measured impacts the key findings; we also try alternative time averaging schemes. Finally, we deploy a recently developed nonparametric panel data estimator to formally estimate and test this conjectured quadratic relationship.

Overall, we are able to successfully replicate the main findings of Cadot et al. (2011). In the course of further analysis we note that the inclusion, or lack thereof, of time specific heterogeneity in the panel model leads one to draw into question both the presence and location of the turning point in the diversification-development nexus. While one can argue that the presence of time effects is immaterial, we confirm the presence of time specific heterogeneity using both well accepted significance tests as well as model selection methods.

Further, several alternative estimation strategies do not help to fully rectify this uncertainty over the presence of a quadratic relationship.

Additionally, we also use a new dataset which offers a thicker granularity of export products. This allows us to extend our data in the time dimension (while giving up little in the country dimension). With this more aggregated data we still reach the same conclusion: there is considerable uncertainty about the existence and location of the turning point within the diversification-development nexus.

Finally, using recently developed nonparametric panel data estimators and inferential tools, we reject the quadratic specification and find that while our nonparametric estimator suggests the presence of a "hump" the location of the turning point is much lower than either parametric or cross-sectional nonparametric estimates propound. We find a turning point that occurs at a much lower level of development than has been previously found and a long, flat portion of the diversification-development nexus. This relatively flat portion of the relationship is an overlooked aspect and one that hopefully, will receive both theoretical and empirical attention moving forward.

The importance of our results lies in the fact that in the 15 years since the seminal work of Imbs & Wacziarg (2003) it appears that this perceived U-shape, while mesmerizing in its simplicity and intuitive appeal, is illusory and does not necessarily represent a real pattern of development across a wide array of countries. Our results, using recently developed model selection, consistent parametric specification inference and a cutting edge nonparametric panel estimator, suggests that the results change drastically from the standard benchmark results with the U-shape turning point shifting quite substantially and in fact the entire U-shape morphing into some other shape which is not appropriately characterized by a U. The main thrust of our findings is that if there was indeed such a prominent (and stable) relationship between diversification and development then this result should stand an attack of the form waged on it here. What is clear from these results is that more work is needed to further investigate this important issue.

The remainder of the paper is structured as follows. In Section 2 we present results which (nearly) replicate the baseline findings presented in Cadot et al. (2011). Section 3 uses several alternative time averaging approaches as well as a different dataset based on a rougher level of granularity of the diversification data to further examine the relationship. Section 4 presents results from consistent model specification tests and nonparametric estimates in a proper panel data framework to examine the diversification-development nexus. Section 5 offers concluding remarks and avenues for future research.

2. The Initial Focus

2.1. **The Baseline Empirical Specification.** In order to discern if a hump shape exists in the income-diversification nexus, Imbs & Wacziarg (2003) (and others) estimate a quadratic in income with trade diversification for country i at year t:

$$TD_{it} = \beta_0 + \beta_1 GDPpp_{it} + \beta_2 GDPpp_{it}^2 + \alpha_i + \lambda_t + \varepsilon_{it}. \tag{1}$$

Here, following Cadot et al. (2011), four different measures of TD_{it} are used: either a Theil (T), Gini, or Herfindahl (HHI) index, or the number of product lines (Nber) that country i exports at time t and $GDPpp$ is gross domestic product per capita in 2005 purchasing power parity (PPP) dollars (scaled by \$10,000). A U-shaped export diversification path would require that $\beta_1 < 0$ and $\beta_2 > 0$ for the first three indices while the opposite is expected to be true for the number of open product lines. Using UNCTAD's COMTRADE HS6 (the finest level of granularity) over 4,991 product lines for 141 countries over the period 1988-2006, a turning point estimate of approximately \$20,000-\$30,000 is found across all four measures of diversification for a variety of assumptions placed on the estimation model in (1).

2.2. **The Empirical Results.** First, we simply aim to replicate the baseline findings of Cadot et al. (2011) by estimating model (1), and using, as these authors did, an unbalanced panel of 141 countries over the period 1988-2006. The data are sourced from the World Bank for income per capita and the UNCTAD COMTRADE database for trade measures.

The main findings of Cadot et al. (2011) are replicated successfully in 1. As in Table 2 in Cadot et al. (2011) we estimate model (1) under three different scenarios for each of the four measures of trade diversification. First, we ignore the panel structure and estimate the model as a pooled panel with time dummies. Second, we exploit the panel structure and estimate the model under the fixed effects framework, but do not include time dummies to capture unobserved time heterogeneity. Third, we estimate a between effects version of model(1) by aggregating over time.[1]

2.3. **Including Time Fixed Effects.** To our surprise though the most appropriate panel specification that would include both country *and* time fixed effect has not been considered in the literature. Given the use of annual data (19 year horizon) and the level of granularity of the HS6 measurement, time effects would help smooth out year-to year fluctuations that might unduly influence the overall estimates. In fact correcting for both country and time

[1]To be exact, there is a small inconsistency in our results: our coefficient estimate for $GDPpcppp^2$ for the number of product lines (Nber) differs somewhat from that obtained by the authors, leading to a higher estimate of the turning point. This is inconsequential, however, as the main finding of Cadot et al. (2011) holds up, namely that there is a hump shape in the income-trade diversification nexus.

Table 1. Replication of Within Estimates from Table 2 in Cadot et al. (2011). HHI is the Herfindahl index for trade diversification.

	Theil	HHI	Gini	Nber
GDPpc	-0.7794	-0.0650	-0.0263	3898.7918
	(0.1587)	(0.0293)	(0.0028)	(162.6998)
GDPpc^2	0.1832	0.0138	0.0059	-568.1568
	(0.0297)	(0.0055)	(0.0005)	(30.4266)
Turning Point ($)	21,269	23,531	22,447	34,311

Note: Using HS6 aggregation data from Cadot et al. (2011).

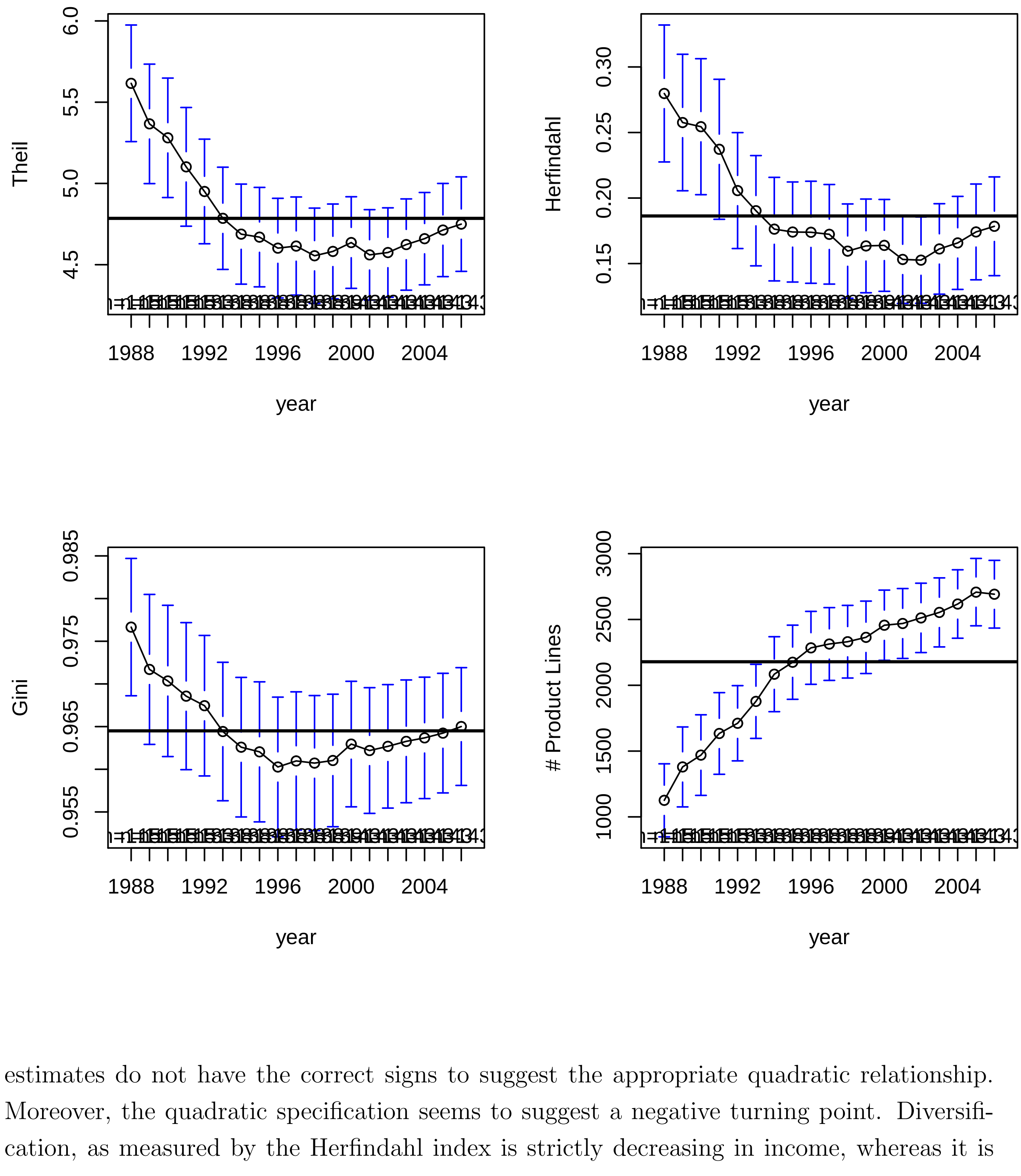
fixed effects has become the baseline specification in the vast majority of cross-country regressions — and we do not see any theoretical or empirical reason why this paper should be an exception.

Perhaps one might argue that the time effects are simply not necessary and their inclusion only leads to a model which obfuscates the diversification-development nexus. Looking at the average of each of the four measures of diversification across time reveals interesting patterns (Figure 1). We see that for our three indices they are all decreasing over time whereas the number of product lines is increasing over time. Moreover, there appears to be more variation in each of the four measures in the early years of the data. Another reason to account for time effects is that during the 1990s the technology boom was in full force and so new products, and means of constructing and delivering products was taking shape during this period. This is confirmed by looking at Figure 2, which plots the time period specific standard deviations of each of the four measures of trade diversification after standardization. Values greater than 1 indicate greater than normal time variation whereas values less than one indicate less than normal time variation, relative to the entire time frame.

Both the Theil and Herfindahl indices display greater than normal time variation at the start of the study, whereas the Gini index has greater than normal time variation in the middle of the time period. The number of product lines has the least amount of variation across the time frame. With this graphical information, we therefore re-estimate model (1) accounting for time variation through the fixed effects framework and present estimates in Table 2.

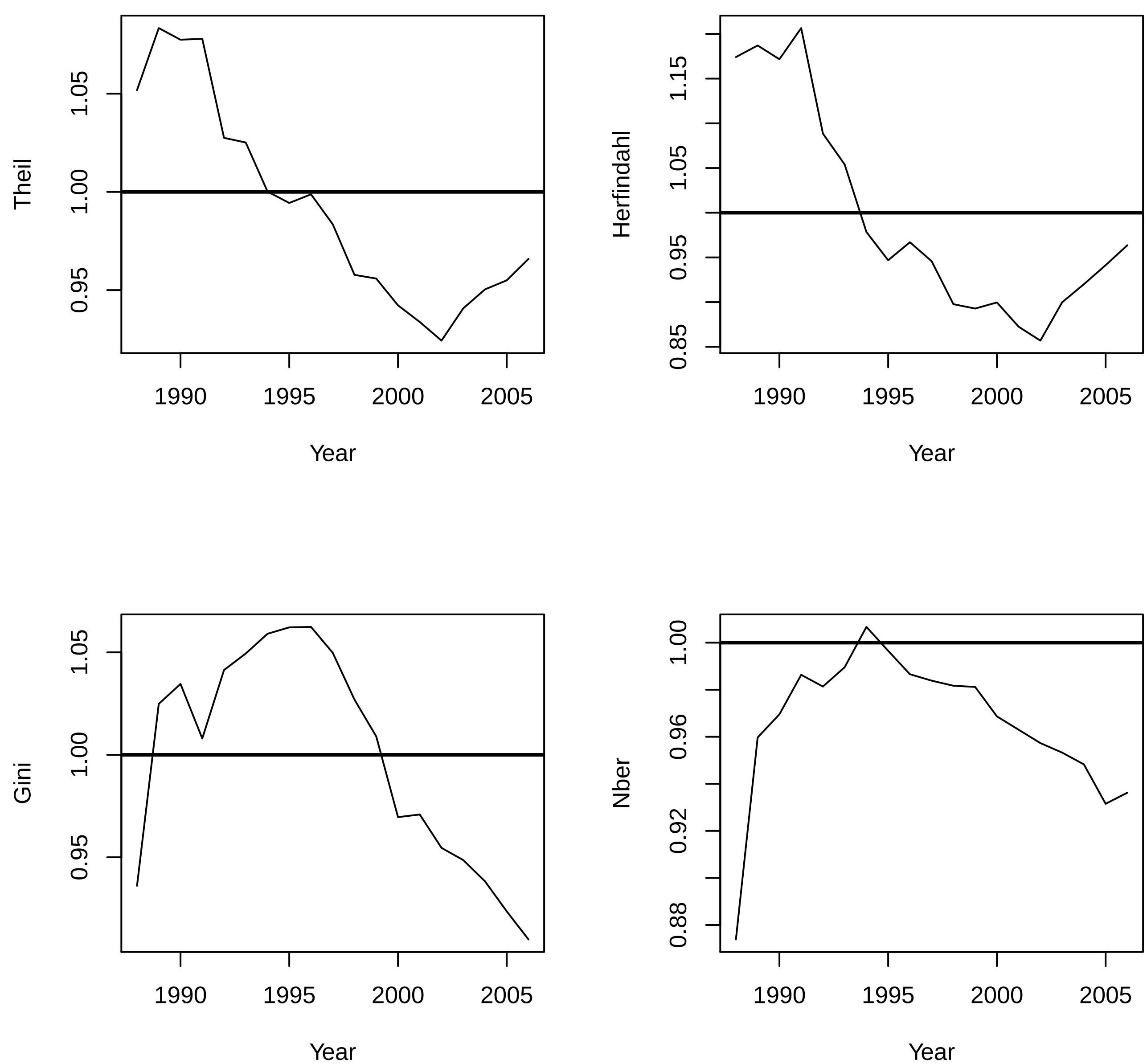
What is immediately noticeable is that coefficient estimates for the income turning point are strikingly different from those obtained in the baseline specification. For both the Theil and Herfindahl indices, there is no hump shaped relationship; in fact the relevant coefficient

Figure 1. Time Heterogeneity in Average of Trade Diversification Indices



estimates do not have the correct signs to suggest the appropriate quadratic relationship. Moreover, the quadratic specification seems to suggest a negative turning point. Diversification, as measured by the Herfindahl index is strictly decreasing in income, whereas it is

Figure 2. Time Heterogeneity in Variation in Trade Diversification Indices



strictly increasing for the Theil index over all positive income. Both the Gini index and the number of product lines confirm to initial findings of Cadot et al. (2011) though the estimated turning point is almost 50% smaller than what was originally estimated. In this

Table 2. Within Estimates including Time Effects. HHI is the Herfindahl index for trade diversification.

	Theil	HHI	Gini	Nber
GDPpc	0.7358	0.1313	-0.0107	1272.7926
	(0.1483)	(0.0312)	(0.0028)	(119.3932)
GDPpc^2	0.0349	-0.0057	0.0045	-299.5895
	(0.0255)	(0.0054)	(0.0005)	(20.5185)
Turning Point ($)	-105,360	115,763	12,006	21,242
Time Effects	0.000	0.000	0.000	0.000

Note: Using HS6 aggregation data from Cadot et al. (2011).

case, there are an additional 17 countries above the turning point: Bahrain, Cyprus, Czech Republic, Gabon, Hungary, Iceland, Kuwait, Luxembourg, New Zealand, Oman, Portugal, Qatar, Saudi Arabia, Slovak Republic, Slovenia, South Korea and United Arab Emirates.[2]

One potential reason that the time effects may wipe away the quadratic impact of GDP is simply that they do not belong in the model. We can easily test for this using a standard F-test for the joint significance of the time effects. For each of the four within models, we reject at all conventional levels the null hypothesis that the time effects are jointly statistically insignificant, confirming their presence in the model. Further, the presence of the time effects does not render the impact of GDP on trade diversification insignificant. While the hump may disappear, in none of the four models is GDP a statistically insignificant predictor of trade diversification. It may simply be that there is no empirical support for a hump in the data. We note that this result in noway invalidates the seminal contribution of Cadot et al. (2011) on the decomposition of the Theil index into the intensive and extensive margin. We simply mention that the inclusion of time specific heterogeneity makes it difficult to detect the classic diversification-development nexus.

Lastly, we also used the model selection approach of Lu & Su (2017) to determine if country-specific, time-specific or both effects should appear in the four models. This model selection device works by calculating a leave-one-observation-out cross-validation function for each of the models under consideration (pooled, only country-specific, one time-specific, and time and country-specific). The model which produces the lowest cross-validation score

[2]We note that a working paper version of Cadot et al. (2011) (Cadot, Carrère & Strauss-Kahn 2007) does include time fixed effects. However, almost no discussion of the results are provided, with the authors mentioning "Indeed, the second block of Table 5, which reports estimates with time and country fixed effects, shows no turning point at all." Further, it seems somewhat inconsistent to include time effects when estimating the pooling model and then drop them when gravitating towards the fixed effects framework.

TABLE 3. Model selection criterion across alternative measures of diversification.

Measure	Model	AIC	BIC	CV
	Pooled	-3.023	-3.016	0.049
HHI	Individual	-4.586	-4.252	0.010
	Time	-3.024	-2.975	0.049
	Two-way	-4.689	-4.314	0.009
	Preferred	Two-way	Two-way	Two-way
	Pooled	0.761	0.768	2.142
Theil	Individual	-1.167	-0.834	0.319
	Time	0.752	0.801	2.121
	Two-way	-1.426	-1.050	0.248
	Preferred	Two-way	Two-way	Two-way
	Pooled	-6.825	-6.818	0.001
Gini	Individual	-9.260	-8.927	0
	Time	-6.823	-6.774	0.001
	Two-way	-9.428	-9.053	0
	Preferred	Two-way	Two-way	Two-way
	Pooled	13.969	13.976	1165667.412
Nber	Individual	12.741	13.075	354682.436
	Time	13.875	13.924	1061907.725
	Two-way	12.232	12.607	213201.217
	Preferred	Two-way	Two-way	Two-way

is deemed the best model. Lu & Su (2017) show that this model selection device out performs more popular selection criteria like AIC and BIC, and also does not place restrictive conditions on the rates at which n and t pass to infinity. More specifically, Lu & Su (2017) need $T = O(\ln N)$ for their asymptotic results to hold.

Table 3 presents the preferred models over Lu & Su's (2017) criterion as well as the more commonly known AIC an BIC criterion. In all cases the model which includes both country and time specific effects is deemed optimal. Again, this result points to the fact that the empirical model which should be deployed is the two-way fixed effects setup which produced the estimates in Table 2.

3. Alternative Checks

3.1. **Alternative Specifications.** Perhaps it is unreasonable to think that trade diversification across countries varies much on an annual basis that inclusion of time specific heterogeneity assists in overfitting the model. However, it may be the case that changes occur over more lengthy periods, perhaps three to five years, where the HS classification has

Table 4. Within Estimates for (1) using 3 and 6 year averaging. HHI is the Herfindahl index for trade diversification.

	3-year Averaging				6-year Averaging			
	Theil	HHI	Gini	Nber	Theil	HHI	Gini	Nber
GDPpc	-0.5702	0.0348	-0.0207	2626.8371	-1.0936	-0.0201	-0.0298	3189.0510
	(0.2140)	(0.0376)	(0.0038)	(234.4974)	(0.3410)	(0.0583)	(0.0060)	(378.8831)
$GDPpc^2$	0.1248	-0.0017	0.0038	-310.6439	0.2091	0.0070	0.0048	-394.0497
	(0.0399)	(0.0070)	(0.0007)	(43.7313)	(0.0658)	(0.0113)	(0.0012)	(73.1577)
Turning Point ($)	22,848	99,907	27,489	42,281	26,145	14,320	30,847	40,465

Note: Using HS6 aggregation data from Cadot et al. (2011).

undergone changes in product lines which may affect overall accounting of diversification. A simple approach to this would be to use a time averaged version of the model in (1). Given that we have 19 years of data, we use three and six year averaging, dropping the 1988 observations from the data.[3]

Our estimates of the four within models under this scenario appear in Table 4. We see that in this case our estimates for both Theil and Gini indices are roughly consistent with Cadot et al.'s (2011) full panel results ignoring time effects. However, the estimated turning point for the number of product lines is considerably larger than their initial estimates. In fact, no countries would be above this turning point over the entire period. Lastly, we see that the HHI index of trade diversification again fails to display robustness, with the coefficient estimates having the wrong signs and the estimated turning point being almost four times as large as initially reported. Moreover, the estimated relationship is such that as countries grow beyond the turning point that diversification increases (as opposed to the general intuition that diversification decreases after the turning point).

3.2. **Using an Alternative Measure of Diversification.** Not obtaining the expected U-shaped relationship when using both country and time fixed effects for either the Theil or Herfindahl index (the most common indices in existing work) raises questions as to whether the relationship is indeed as robust as most of the literature believes. A consideration worth investigating further is the relatively short time dimension of the dataset used in Cadot et al. (2011). That is, perhaps the short period of study (1988-2006) could only partially capture the experience of different countries along different stages of their development process and therefore failing to adequately reflect the entire U-shape path of diversification.

[3]Our findings are qualitatively identical to dropping 2007, the last period as well.

A possible way to improve the time-dimension issue is by considering instead the HS4 level trade data. Moving to this lower level of granularity has the advantage of extending the time frame both backwards, to 1962, and forwards, to 2010, to determine if the hump appears over a longer time horizon. While there is loss of information at the cross-sectional dimension there is an significant increase in the time dimension (from 19 to 49 years). We use the data in Papageorgiou & Spatafora (2012) to reestimate model (1) focusing on the Theil index (that most commonly used in the literature).

Table 5. Estimates using HS4 level aggregation, Theil Index Only.

	Pooled		Within no time effects		Within time effects	
	1962-2010	1988-2007	1962-2010	1988-2007	1962-2010	1988-2007
GDPpc	-1.8857	-1.8956	-0.5282	0.1840	-0.0128	0.4357
	(0.0413)	(0.0646)	(0.0398)	(0.0812)	(0.0431)	(0.0892)
$GDPpc^2$	0.36917	0.37807	0.12283	0.00989	0.05947	-0.01744
	(0.01211)	(0.01919)	(0.00830)	(0.01482)	(0.00828)	(0.01524)
Turning Point ($)	25,539	25,070	21,502	**-92,967**	**1,078**	**124,934**
# Countries	131	134	131	134	131	134
# Observations	5,534	2,441	5,534	2,441	5,534	2,441

Note: Using HS4 aggregation data from Papageorgiou & Spatafora (2012).

Table 5 presents both pooled and within estimates using the HS4 trade data constructed Theil index. We present our estimates both for the full time frame available, 1962-2010, as well as the time frame specific to Cadot et al.'s (2011) analysis, 1988-2006. The within estimates are calculated both including and excluding time effects. The results are stark. First, using the 1988-2006 time frame (i.e. using HS4 data instead of HS6 data for the same time horizon as the one used by Cadot et al.'s (2011)) we show that the pooled data estimates are consistent with the expected hump shape and an estimated turning point of around $25,000, while when we allow for country and time fixed effects continue to show no evidence of the hump. Most importantly, when we use the HS4 level data of the entire time period of 1962-2010 including both time and country effects we once again, and against our prior, fail to uncover the hump shape.

4. A Nonparametric Examination

Given that there is some skepticism as to the existence of a quadratic shape in the development-diversification nexus, resorting to nonparametric methods may help to shed more light on this relationship. We note that Cadot et al. (2011) provided graphical evidence of a quadratic relationship using pooled nonparametric methods. Here, we elect to use recently developed nonparametric panel data methods to appropriately control for country-specific heterogeneity. While a range of methods have recently been proposed (a thorough review is provide in Parmeter & Racine 2018), we deploy the profile least-squares estimator of Gao & Li (2013), Li, Peng & Tong (2013), and Lin, Li & Sun (2014).[4] All three of these papers have developed the asymptotic theory of the kernel smoothed regression estimator of the model

$$y_{it} = m(\boldsymbol{x}_{it}) + \alpha_i + \varepsilon_{it}. \tag{2}$$

In our setup y_{it} would be one of the four measures of trade diversification and $\boldsymbol{x}_{it}$ would represent GDP per capita. For generality we will assume that $\boldsymbol{x}_{it}$ is of dimension q. We are also operating in the fixed effects setting, thus $E[\alpha_i|\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT}] \neq E[\alpha_i]$.

To detail how the estimator can be implemented we note that the profile least-squares estimator works in three steps. First, the α_i are assumed to be known; in this case they can be subtracted from y_{it} and we engage in standard, cross-sectional nonparametric kernel regression. For the local-linear setting, define $M(\boldsymbol{z}) = (m(\boldsymbol{z}), h \odot \dot{m}(\boldsymbol{z})')'$ (where $\odot$ represents Hadamard multiplication) which represents the $(q+1) \times 1$ vector composed of the function itself and the bandwidth scaled derivatives. An estimate of $M(\cdot)$ can be constructed from

$$M_\alpha(\boldsymbol{x}) = \arg \min_{M \in \mathbb{R}^{q+1}} \big(Y - D\alpha - D_{\boldsymbol{x}} M\big)' \mathcal{K}_{\boldsymbol{x}} \big(Y - D\alpha - D_{\boldsymbol{x}} M\big), \tag{3}$$

where Y is the vector of response variables, α is the vector of individual effects, $D_{\boldsymbol{x}}$ is the $n \times (q+1)$ matrix whose first column is all ones and the remaining q columns are composed of $X - \boldsymbol{x}$ where X is the matrix of observations of the data, $D = (I_n \otimes i_T) \cdot d_n$, $d_n = [-i_{n-1}, I_{n-1}]'$, and i_n is a $n \times 1$ vector of ones. The construction of D is to ensure that $N^{-1} \sum_{i=1}^{N} \alpha_i = 0$, which is a necessary identification condition for estimation of the unknown function.

Define the smoothing matrix

$$S(\boldsymbol{x}) = (D'_{\boldsymbol{x}} \mathcal{K}_{\boldsymbol{x}} D_{\boldsymbol{x}})^{-1} D'_{\boldsymbol{x}} \mathcal{K}_{\boldsymbol{x}},$$

where $\mathcal{K}_{\boldsymbol{x}}$ is a diagonal matrix which contains kernel weights along the diagonal. While there are many methods in which to calculate the kernel weights, we adopt the product kernel with individual bandwidths for each of the covariates (Li & Racine 2007).[5] The estimator which

[4]These papers are all based on Su & Ullah (2006) and Sun, Carroll & Li (2009) who originally studied estimators for a partly linear panel data model.

[5]This is trivial here as we only have a single covariate, GDP per capita, but in more advanced settings it is recommend to use a different smoothing parameter for each variable in one's model.

solves the minimization problem in Equation (3) is then

$$\widehat{M}_\alpha(\boldsymbol{x}) = S(\boldsymbol{x})\ddot{\varepsilon}$$

where $\ddot{\varepsilon} = Y - D\alpha$. $\widehat{M}_\alpha(\boldsymbol{x})$ contains the estimator of the conditional mean of inefficiency, $\widehat{m}_\alpha(\boldsymbol{x})$ as well as the $q \times 1$ vector of estimated first derivatives, scaled by the appropriate bandwidth, $h \odot \widehat{m}_\alpha(\boldsymbol{x})'$. Define $s(\boldsymbol{x})' = e'S(\boldsymbol{x})$ with $e = (1, 0, \ldots, 0)'$ the $(q+1) \times 1$ vector. Then $\widehat{m}_\alpha(\boldsymbol{x}) = s(\boldsymbol{x})'\ddot{\varepsilon}$.

Once the estimator of $m(\boldsymbol{x})$ is obtained, the second step focuses on the estimation of α, which is done through profile least squares estimation[6] of the first stage residuals on the matrix of dummy variables for the firms. More specifically, our estimator is found from

$$\widehat{\alpha} = \arg\min_{\alpha} \big(Y - \widehat{m}_\alpha(\boldsymbol{x}) - D\alpha\big)'\mathcal{K}_{\boldsymbol{x}}\big(Y - \widehat{m}_\alpha(\boldsymbol{x}) - D\alpha\big),$$

with $\widehat{m}_\alpha(\boldsymbol{x}) = (\widehat{m}_\alpha(\boldsymbol{x}_{11}), \widehat{m}_\alpha(\boldsymbol{x}_{12}), \ldots, \widehat{m}_\alpha(\boldsymbol{x}_{1T}), \widehat{m}_\alpha(\boldsymbol{x}_{21}), \ldots, \widehat{m}_\alpha(\boldsymbol{x}_{NT}))'$. Lin et al. (2014) show that the profile least squares estimator (the minimizer of the profile least squares problem) is

$$\widehat{\alpha} = \left(\widetilde{D}'\widetilde{D}\right)^{-1} \widetilde{D}'\widetilde{Y},$$

with $\widetilde{D} = (I_{NT} - S)D$ and $\widetilde{Y} = (I_{NT} - S)Y$. Here $S = (s(\boldsymbol{x}_{11}), s(\boldsymbol{x}_{12}), \ldots, s(\boldsymbol{x}_{1T}), s(\boldsymbol{x}_{21}), \ldots, s(\boldsymbol{x}_{NT}))'$ and $\widehat{\alpha}_1 = -\sum_{i=2}^{n} \widehat{\alpha}_i$.

The third and final step is to estimate $M(\boldsymbol{x})$ using the estimates for α. In this setting the profile least-squares estimator is

$$\widehat{M}(\boldsymbol{x}) = \widehat{M}_{\widehat{\alpha}}(\boldsymbol{x}) = S(\boldsymbol{x})\widehat{\ddot{\varepsilon}} \tag{4}$$

with $\widehat{\ddot{\varepsilon}} = Y - D\widehat{\alpha}$.

4.1. **Estimation with Unbalanced Panel Data.** In our setting, and as is common in many economic settings, unbalanced panel data presents itself. In this case the profile least-squares estimator just described only requires minor modifications in notation to be operational; specifically, the construction of the matrix D. In the balanced setting D is an $nT \times (n-1)$ matrix. With unbalanced panel data D is a $\check{T} \times (n-1)$ matrix where $\check{T} = \sum_{i=1}^{n} T_i$ is the total number of observations in the data set and T_i is the number of time periods that country i appears in the data.

To understand how exactly D changes in the presence of unbalanced panel data, define the $T \times (n-1)$ matrix Δ_1 consisting of all -1s and the $T \times (n-1)$ matrix$\Delta_\jmath$, with $\jmath \in (2, \ldots, n)$

[6]Here the profiling is over the $\boldsymbol{x}$s.

which has all entries 0 except for the $\jmath - 1$ column, which contains 1s. In the balanced case we can write D as

$$D_{bal} = \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_n \end{bmatrix}.$$

In the unbalanced setting let $e_\jmath$ be the vector of 1s and 0s representing which of the T time periods individual $\jmath$ appears in and define the $T_\jmath \times T$ matrix with 1s along the main diagonal and 0s everywhere else as $\Gamma_\jmath$. Γ_1 is the matrix composed of vertical concatenation of all the $e_\jmath$ vectors. Assuming that the 1st individual appears T times, then in the unbalanced case we have

$$D_{unbal} = \begin{bmatrix} \Gamma_1 \odot \Delta_1 \\ \Gamma_2 \Delta_2 \\ \vdots \\ \Gamma_n \Delta_n \end{bmatrix}.$$

Aside from the modification to the D matrix, there are no additional changes to the profile least-squares estimator.

4.2. **Additional Extensions.** The profile least-squares estimator just described is appealing, as noted by Parmeter & Racine (2018), it is easy to implement, does not require iterative methods (as many earlier estimators that have proposed do) and has a straightforward link with existing kernel methods. However, there are many possible areas where the estimator could be improved upon. First, we have detailed local-linear estimation. Extending the estimator to the local-polynomial setting is certainly worthwhile as it is well known that higher order polynomial fits serve to reduce bias in the estimator (Li & Racine 2007). Additionally, few of the existing approaches on kernel estimation in the fixed effects framework explicitly detail how best to select the vector of smoothing parameters, h. A popular applied tool is to use a rule-of-thumb bandwidth, such as $h_\jmath \sim 1.059 \cdot \widehat{\sigma}_{x_\jmath} \check{T}^{-1/(4+q)}$, where $\widehat{\sigma}_{x_\jmath}$ is the standard error of the jth covariate. Henderson & Parmeter (2015) mention the use of least-squares cross-validation (LSCV), however, in the case where an entire cross-section is left out, the prediction of $alpha_i$ remains an issue.

Lastly, an interesting, an still unexplored extension would be to consider the two-way specification of the model. That is,

$$y_{it} = m(\boldsymbol{x}_{it}) + \alpha_i + \lambda_t + \varepsilon_{it}. \quad (5)$$

A profile least-squares estimator could be developed for this setting following the lines detailed above. In this case, one would need to assume that both α and λ were known, which would necessitate including two D matrices, one for individual effects and another for time effects. This would be a useful extension for practitioners but is beyond the scope of the paper. We leave this for future research. We note however, that the one-way fixed effects framework detailed in (2) is consistent with the preferred approach of the main studies in this field.

4.3. **Specification Testing.** Beyond direct nonparametric estimation of the development-diversification nexus, we are also interested in formal specification testing of the quadratic relationship proposed by Cadot et al. (2011). A direct test of correct specification was formulated by Lin et al. (2014) through use of an integrated squared error statistic. Their test is based on local-constant estimation of the panel data model in (2) under the fixed effects framework. The null hypothesis of correct parametric specification is

$$H_0\colon Pr\left\{m(\boldsymbol{x}) = \boldsymbol{x}\beta_0\right\} = 1,$$

for some $\beta_0 \in \mathbb{R}^q$ against the alternative hypothesis

$$H_1\colon Pr\left\{m(\boldsymbol{x}) = \boldsymbol{x}\beta_0\right\} < 1,$$

for any $\beta_0 \in \mathbb{R}^q$. Denote the parametric estimator of β_0 as $\widehat{\beta}$, which could be the within or least-squares dummy variable (LSDV) estimator and $\widehat{m}(\boldsymbol{x})$ the profile least-squares estimator just described. A consistent test for H_0 can be constructed through the sample equivalent of squared error:

$$\int \left(\widehat{m}(\boldsymbol{x}) - \boldsymbol{x}\widehat{\beta}\right)^2 d\boldsymbol{x}.$$

This test statistic cannot be directly deployed as there are several technical issues surrounding it, namely the appearance of non-zero centering terms which produce an asymptotic bias if not properly attended to. Lin et al. (2014) use the approach of Härdle & Mammen (1993) to avoid these non-zero center terms (see also Henderson & Parmeter 2015) and smooth $\boldsymbol{x}_{it}\widehat{\beta}$ using kernel weighting. Specifically, estimate $m(\boldsymbol{x})$ using local-constant least-squares as

$$\widehat{m}(\boldsymbol{x}) = \left(\imath_{NT}' S(\boldsymbol{x})\imath_{NT}\right)^{-1} \imath_{NT}' S(\boldsymbol{x}) y \tag{6}$$

where $S(\boldsymbol{x}) = Q(\boldsymbol{x})'\mathcal{K}_{\boldsymbol{x}} Q(\boldsymbol{x})$ and $Q(\boldsymbol{x}) = I_{NT} - D\left(D'\mathcal{K}_{\boldsymbol{x}}D\right)^{-1} D'\mathcal{K}_{\boldsymbol{x}}$ with $\imath_{NT}$ an $NT \times 1$ vector of ones. Note that we smooth over y in Equation (6) as $Q(\boldsymbol{x})D = 0$ which will eliminate the presence of the fixed effects as in Equation (4). The same smoothing is applied to the parametric estimates to produce

$$\widehat{m}_{para}(\boldsymbol{x}) = \left(\imath_{NT}' S(\boldsymbol{x}) \imath_{NT}\right)^{-1} \imath_{NT}' S(\boldsymbol{x}) \left(\boldsymbol{x}_{it}\widehat{\beta}\right).$$

Let $\widehat{\epsilon}_{it} = y_{it} - \boldsymbol{x}_{it}\widehat{\beta}$ (the parametric residuals, free of the fixed effects). Then it holds that $\widehat{m}(\boldsymbol{x}) - \widehat{m}_{para}(\boldsymbol{x}) = \left(\imath_{NT}' S(\boldsymbol{x}) \imath_{NT}\right)^{-1} \imath_{NT}' S(\boldsymbol{x}) \widehat{\epsilon}$. Lin et al. (2014) propose a final test-statistic for H_0

$$\widehat{I}_{NT} = \frac{1}{N^2|h|} \sum_{i=1}^{N} \sum_{j\neq i}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} \widehat{\widetilde{\epsilon}}_{it} \widehat{\widetilde{\epsilon}}_{js} K_{itjsh}, \tag{7}$$

where $\widehat{\widetilde{\epsilon}}_{it} = \widehat{\epsilon}_{it} - \widehat{\epsilon}_{i\cdot}$ are the within transformed residuals and K_{itjsh} is the kernel function evaluate at the point $\frac{x_{it} - x_{js}}{h}$. Lin et al. (2014) demonstrate that an appropriately normalized version of $\widehat{I}_{NT}$ has an asymptotically normal distribution.[7]

While critical values can be calculated directly from a normal table once the test statistic in (7) has been calculated, it is well known that kernel-based tests possess poor finite sample size and power (Li & Racine 2007, Henderson & Parmeter 2015). The classic remedy to these poor asymptotic approximations is to rely on resampling plans, which offer asymptotic refinements. An easy to implement wild bootstrap resampling plan is given in Lin et al. (2014) to approximate the distribution of $\widehat{J}_{NT} = N\sqrt{|h|}\widehat{I}_{NT}/\sqrt{\widehat{\sigma}_0^2}$ with

$$\widehat{\sigma}_0^2 = \frac{2}{N^2|h|} \sum_{i=1}^{N} \sum_{j\neq i}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} \widehat{\widetilde{\epsilon}}_{it}^{2} \widehat{\widetilde{\epsilon}}_{js}^{2} K_{itjsh}^2. \tag{8}$$

The bootstrap procedure proposed by Lin et al. (2014) is (borrowing from Parmeter & Racine (2018)):

(1) Estimate the linear panel data model under the fixed effects framework using the within estimator and obtain the residuals $\widehat{\epsilon}_{it} = y_{it} - \boldsymbol{x}_{it}\widehat{\beta}$.
(2) For $i = 1, 2, \ldots N$ and $t = 1, 2, \ldots T$ generate the two-point wild bootstrap error $\epsilon_{it}^* = \frac{(1-\sqrt{5})}{2}\left(\widehat{\varepsilon}_{it} - \overline{\widehat{\varepsilon}}\right)$ with probability $p = \frac{(1+\sqrt{5})}{2\sqrt{5}}$ and $u_{it}^* = \frac{(1+\sqrt{5})}{2}\left(\widehat{\epsilon}_{it} - \overline{\widehat{\epsilon}}\right)$ with probability $1-p$. Then construct $y_{it}^* = \boldsymbol{x}_{it}\widehat{\beta} + \epsilon_{it}^*$. Call $(y_{it}^*, \boldsymbol{x}_{it})$ for $i = 1, 2, \ldots N$ and $t = 1, 2, \ldots T$ the bootstrap sample.
(3) Use the bootstrap sample to estimate β deploying the within estimator. Calculate the residuals $\widehat{\epsilon}_{it}^* = y_{it}^* - \boldsymbol{x}_{it}\widehat{\beta}^*$.
(4) Compute J_{NT}^*, where J_{NT}^* is obtained from J_{NT} using the residuals $\widehat{\widetilde{\epsilon}}_{it}^*$.

[7]They focus on the asymptotic behavior of $\widehat{I}_{NT}$ for $N \to \infty$ but the theory can be established if both N and T are increasing.

TABLE 6. Consistent Model Specification Test Results. Each test is conducted using $B = 1000$ Monte Carlo replications with bandwidths selected via least-squares cross-validation.

Measure	$\widehat{J_{NT}}$	Bootstrap p-value
Herfindahl	3.612	<0.01
Theil	15.406	<0.01
Gini	22.278	<0.01
Product Lines	30.253	<0.01

(5) Repeat steps (2)-(4) a large number (B) of times and reject H_0 if the estimated test statistic $\widehat{J}_{NT}$ is greater than the upper α-percentile of the bootstrapped test statistics.

In both theory and simulations, Lin et al. (2014) demonstrate that the proposed bootstrap approach dominates the asymptotic version of the test. More specifically, in their simulations, the bootstrap test has correct size in both univariate and bivariate settings, suggesting practical merit for applications. The bootstrap test also has a high degree of power.

This test can be readily implemented in existing statistical software. Lin et al.'s (2014) test is identical in construction to the consistent model specification test of Li & Wang (1998), albeit in the panel setting. However, if the model is estimated using the within transformation, then there is no difference. This means that one only needs to engage in kernel weighting of the within residuals from parametric estimation. More specifically, the `np` package (Hayfield & Racine 2008) in the `R` programming environment offers a test of consistent model specification in the cross-sectional setting, `npcmstest()`. The adept user could simply within transform their data prior to calling this command, making implementation straightforward; this is the approach we follow here.

4.4. **Our Findings.**

4.4.1. *Inference for Correct Specification.* Prior to nonparametric estimation of the diversification-development nexus, we implement the bootstrap test of Lin et al. (2014) just described. We test for correct quadratic specification, consistent with Cadot et al. (2011), for each of the four different measures of diversification (Theil, Herfindahl and Gini indices along with number of open product lines). We use a bandwidth selected through least-squares cross-validation and $B = 1000$ replications. All four of the tests soundly reject the null of correct quadratic specification. The values of the normalized test statistics and their corresponding bootstrap p-values appear in Table 6.

These results should not be taken to mean that a hump type relationship does not exist in the diversification-development nexus. Rather, this results suggest that the quadratic specification is not fully capturing the shape of the conditional mean. To obtain a clearer picture of what this relationship looks like, while accounting for country specific heterogeneity, we now turn to nonparametric estimation.

4.4.2. *Nonparametric Estimation.* We implement both pooled cross-section and pure panel data nonparametric estimators to graphically assess the diversification-development nexus. We do so for each of the four different measures of diversification. In each setting we use rule-of-thumb bandwidths. Both estimators (pooled cross-sections and panel) are implemented in the local-linear framework.

The Theil Index. Figure 3 presents the nonparametric estimates for both pooled cross-sections and the local-linear profile least squares estimator (in panel (a)) and the estimated gradients (in panel (b)). There are several notable features. First, both the pooled cross-sections and the profile least-squares estimators suggest a hump in the diversification-development nexus, but they differ dramatically on where this hump occurs.

The pooled cross-sections estimate of the hump is consistent with Cadot et al. (2011) occurring between $24,000 and $25,000, while the location of the hump for the proper nonparametric panel data estimator is much lower, around $7,000. Moreover, from the estimated gradients, it appears that the intensity of the diversification that occurs have the hump fades. This roughly flattening out of the curve could be the source of the rejection of the consistent model specification test.

The Herfindahl Index. Figure 4 presents the same set of nonparametric estimates for both pooled cross-sections and the local-linear profile least squares estimator (in panel (a)) and the estimated gradients (in panel (b)) for the Herfindahl index. The insights from the Herfindahl index are nearly identical to those from the Theil index. Both pooled and panel estimates suggest a hump type relationship, but the location of the turning point again differs dramatically, though the pooled cross-section actually displays two humps, one that is quite close to the panel estimates and one more inline with the parametric results discussed earlier.

Again, the gradient estimates suggest that for the proper panel specification that the relationship flattens out as an economy grows larger. This is most likely the underlying source of the rejection of correct specification for the Herfindahl index that appears in Table 6. From $0 to approximately $15,000, both sets of estimates are remarkably similar. It is the movement beyond this point where the difference between a cross-sectional approach and a panel approach begins to manifest.

One can argue that the quadratic relationship is simply not there, and has never been there. The flattening of the relationship later on in the development of any country could simply be that there is less noise in the data as with better quality data for advanced economies as opposed to rather noisy source level data in developing economies and particularly LICs. As we will see in the sequel, what may lie underneath the data pattern that we see is nothing but mere shadow cross-sectional relationship that carries no value when looked at from the perspective of cross-sectional variation. Put differently, looking at the time series dimension of this problem could in fact reveal that very few countries exist that go through the "U-shape" path.

The Gini Index. Figure 5 uses the Gini index to examine the development-diversification nexus, again comparing the estimates from both pooled cross-sections and local-linear profile least squares (in panel (a)) and the estimated gradients (in panel (b)). The estimates in this case follow roughly the same pattern as the two earlier indices. The panel model has a hump shaped relationship, but after $15,000 it flattens out and the difference in the location of the humps differs by almost $20,000 in GDP per capita between the pooled cross-sections estimates and the panel estimates.

Judging from the gradients in panel (b) both sets of estimates suggest a relatively flat relationship between diversification and development. Again, this flatness is most likely the driving force behind the rejection of the quadratic specification for the Gini index. Moreover, while a general quadratic shape appears in the pooled cross-sections estimates, it has a relatively low intensity. At least from the prospect of how much each product composes the entire diversification portfolio, it seems that this relationship is quite weak. The panel estimates suggest a more fluid and robust relationship up to roughly $15,000, as has been consistent with the other two indices.

Open Product Lines. Our last comparison uses the number of active or open product lines to study the development-diversification nexus. Figure 6 presents the estimates from both pooled cross-sections and local-linear profile least squares (in panel (a)) and the estimated gradients (in panel (b)). Here the estimates are quite similar between the two methods. While a hump does appear in the panel estimates, this is so close to zero and judging by the cluster of the data, it seems that this hump is spurious.

Both curves steadily rise and reach a peak right around $28,000. The gradient estimates also bare out this similarity. Here the development-diversification nexus appears to be robust and consistent with earlier findings in the literature. Even though the panel estimates continue upward, closer inspection reveals that this is due to several observations with exceedingly large GDP per capita and so little can be taken from so few points to base the relationship off of.

5. Conclusion

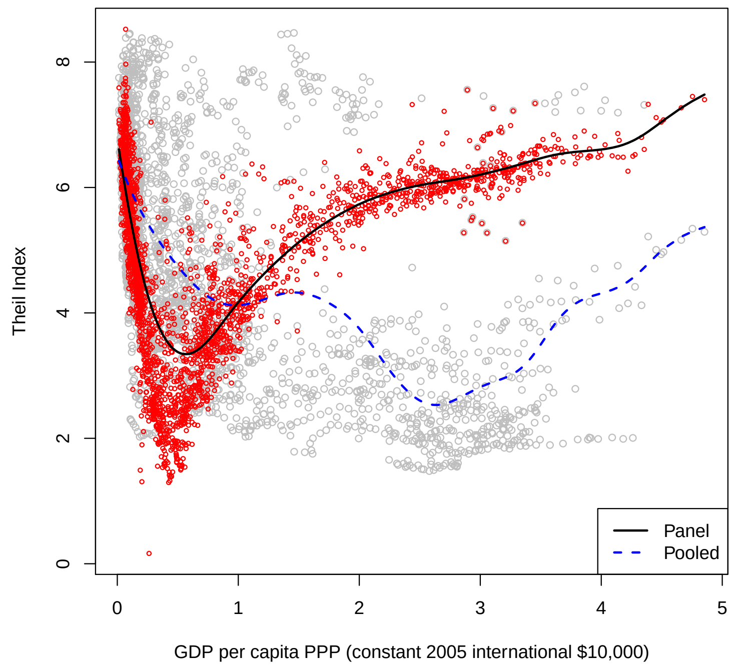
This work has investigated the shape of the trade diversification-development nexus. Following the seminal contribution of Imbs & Wacziarg (2003), it appears that when subjected to a variety of econometric "stress" tests, the perceived, and widely held view of a U-shape disappears. This occurs across many dimensions. For instance, the inclusion of time effects within the panel structure draws into question the seeming robustness of the trade diversification hump. What do the congeries of estimates and inferences mean?

First, the lack of empirical robustness of various measures of trade diversification suggests that the stylized fact of a hump shape in the trade diversification-development nexus requires more care and thought. Second, rather than the use of annual measures, more aggregated measures could be deployed to smooth out product lines or channels of trade that are short lived to aptly capture the long term effects of trade. Indeed, we see that, using the Theil index, the prominent hump shape is robust to both three and six year averaging even though several other measures are less resistant in this setting. Making strong arguments about reconcentration as a path that most countries will follow requires further investigation - and perhaps be more closely linked to the emerging literature of quality upgrading (indices using trade data on prices rather than volumes).

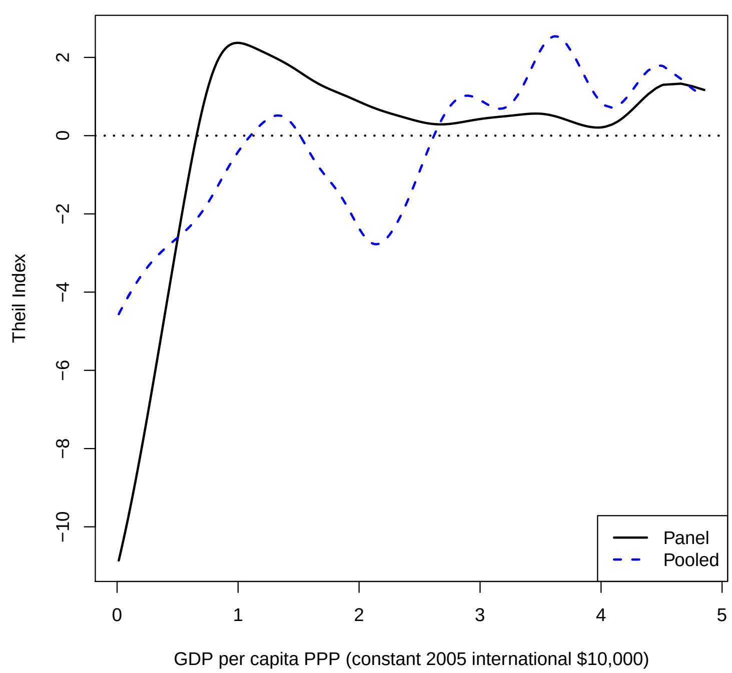
Lastly, while our model specification tests reject the pure "U-shape" relationship that is consistent with a quadratic form, our nonparametric estimates do reveal a hump shape, though one where the hump both occurs at a much lower level of development and one that considerably flattens out as a country becomes more developed. This last result is important as it is suggestive that once countries reach a certain threshold of development that product line thinning (or concentration) is less of an issue that at lower levels of development.

All told, the results here suggest that the "U-shape" is not robust to empirical scrutiny and more work needs to be undertaken at the theoretical level to provide insights into the form of this relationship. As newer panel data estimators and inferential procedures are put forth, they can, and should, be deployed to help further provide evidence in this exciting arena.

Figure 3. Cross-section and panel estimates of the diversification-development nexus using the Theil index.

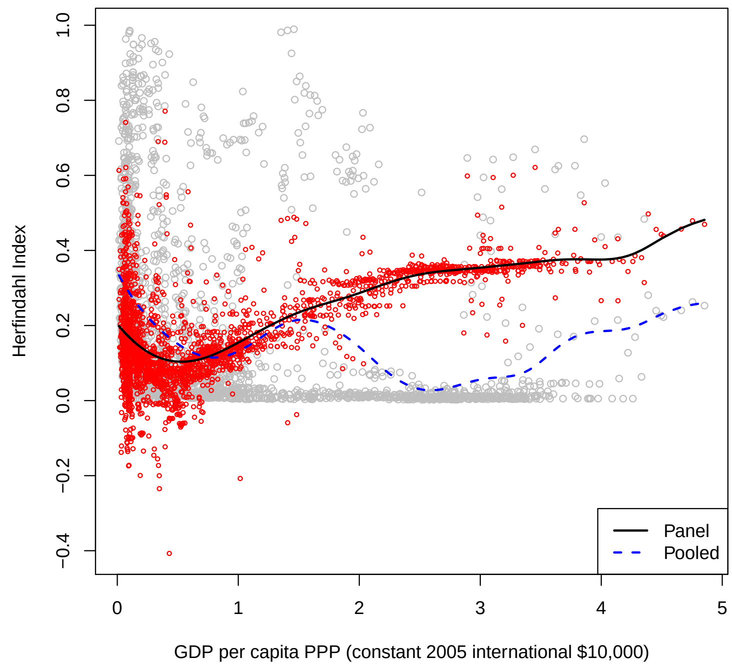


(a) Conditional Mean

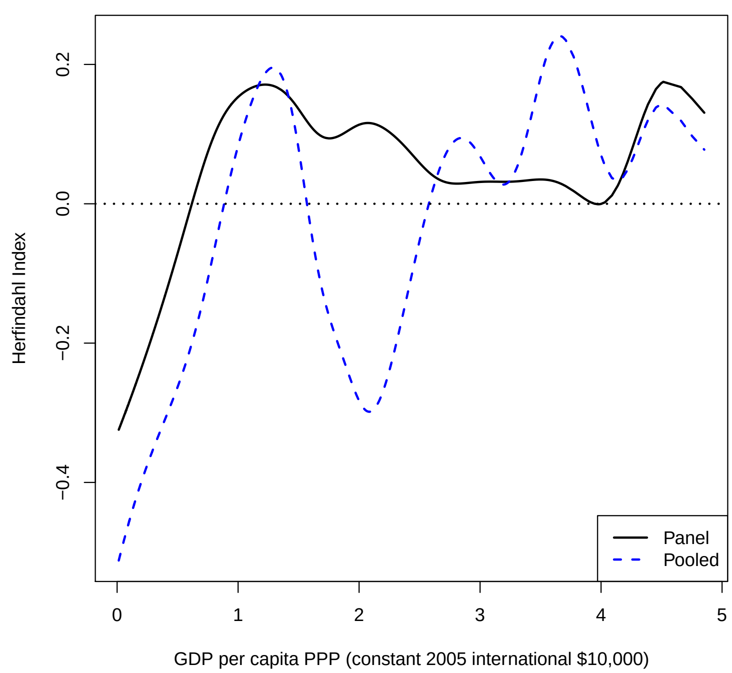


(b) Gradients

FIGURE 4. Cross-section and panel estimates of the diversification-development nexus using the Herfindahl index.

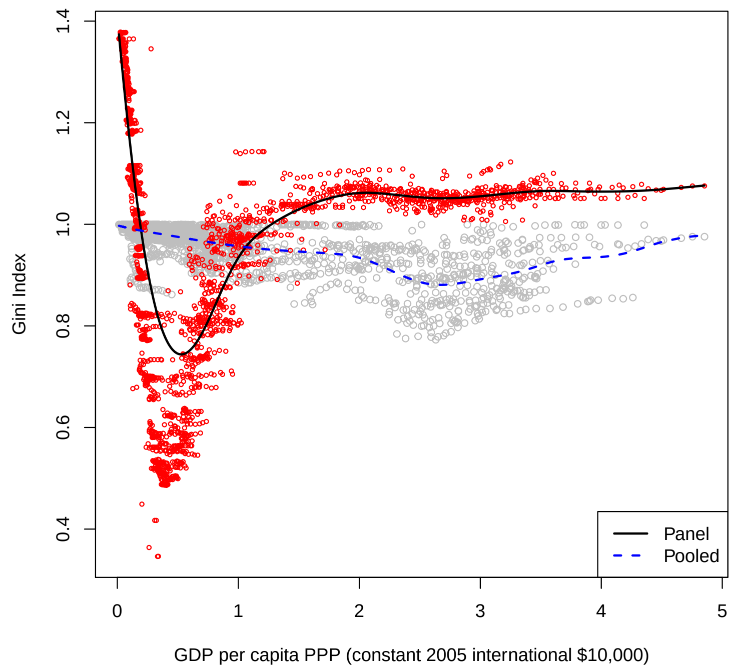


(a) Conditional Mean

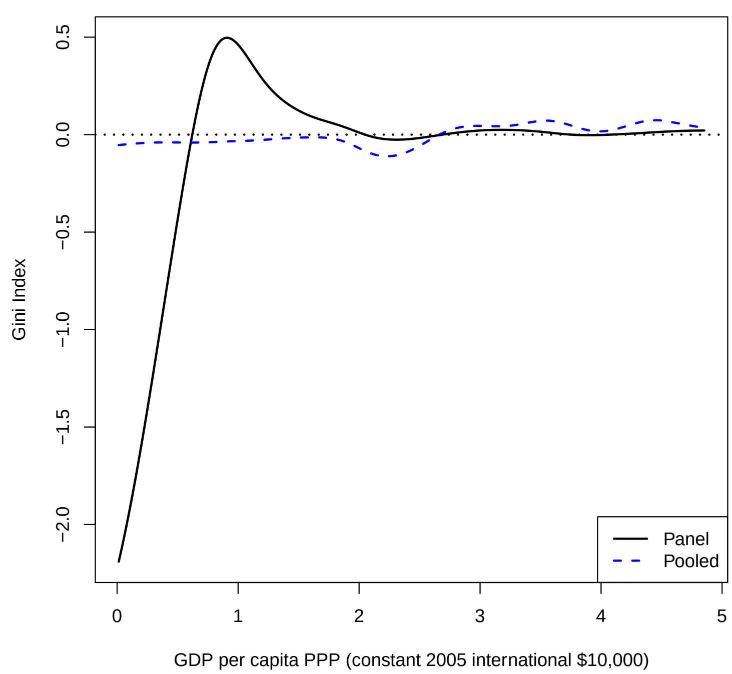


(b) Gradients

Figure 5. Cross-section and panel estimates of the diversification-development nexus using the Gini index.

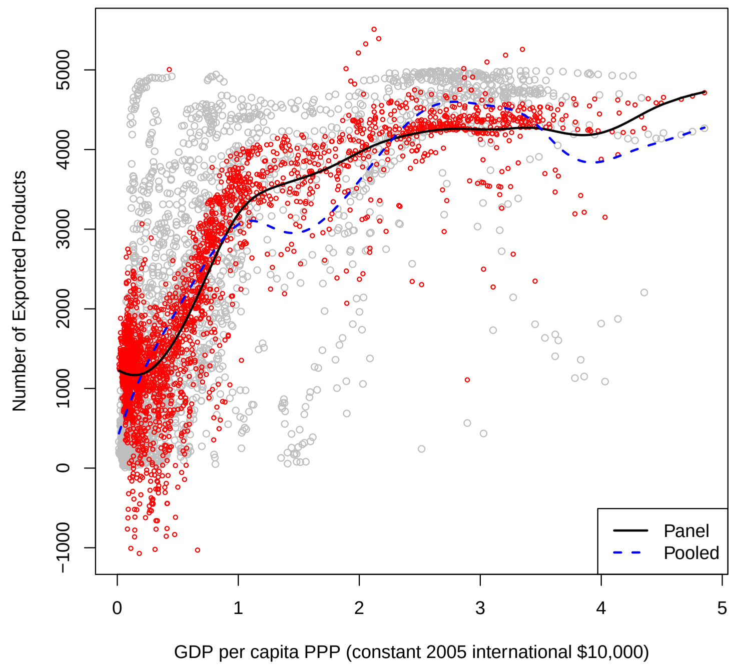


(a) Conditional Mean

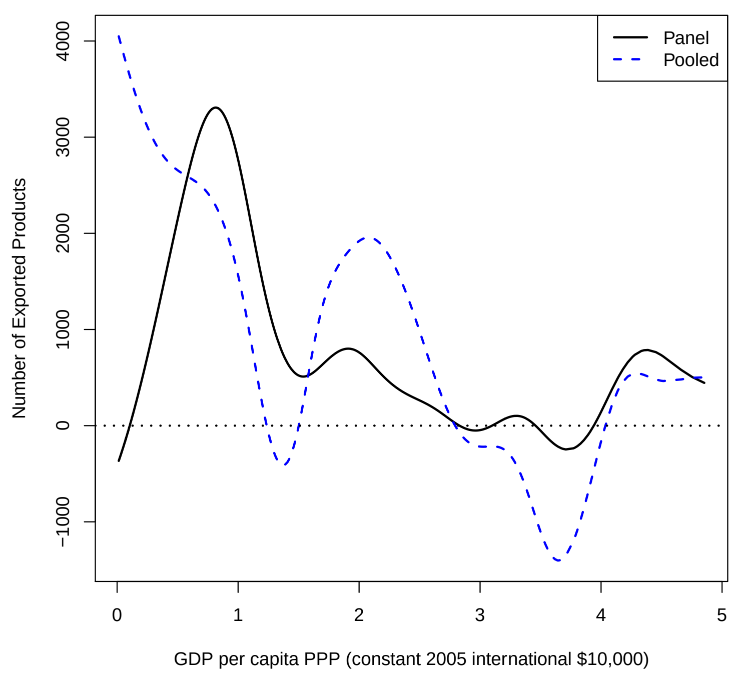


(b) Gradients

Figure 6. Cross-section and panel estimates of the diversification-development nexus using active product lines.



(a) Conditional Mean



(b) Gradients

References

Cadot, O., Carrère, C. & Strauss-Kahn, V. (2007), Export diversification: What's behind the hump? Expert Meeting on Participation of Developing Countries in New and Dynamic Sectors of World Trade: The South-South Dimension, Geneva, Switzerland.
URL: *http://r0.unctad.org/ditc/tab/events/emstrade/Speakers/Cadot_doc.pdf*

Cadot, O., Carrère, C. & Strauss-Kahn, V. (2011), 'Export diversification: What's behind the hump?', *The Review of Economics and Statistics* **93**, 590–605.

Gao, Y. & Li, K. (2013), 'Nonparametric estimation of fixed effects panel data models', *Journal of Nonparametric Statistics* **25**(3), 679–693.

Härdle, W. & Mammen, E. (1993), 'Comparing nonparamtric versus parametric regression fits', *Annals of Statistics* **21**, 1926–1947.

Hayfield, T. & Racine, J. S. (2008), 'Nonparametric econometrics: The np package', *Journal of Statistical Software* **27**(5).
URL: *http://www.jstatsoft.org/v27/i05/*

Henderson, D. J. & Parmeter, C. F. (2015), *Applied Nonparametric Econometrics*, Cambridge University Press.

Imbs, J. & Wacziarg, R. (2003), 'Stages of diversification', *American Economic Review* **93**, 63–86.

Klinger, B. & Lederman, D. (2006), Diversification, innovation, and imitation inside the global technology frontier. World Bank Policy Research Working Paper # 3872, The World Bank.

Li, G., Peng, H. & Tong, T. (2013), 'Simultaneous conifdence band for nonparametric fixed effects panel data models', *Economics Letters* **119**(2), 229–232.

Li, Q. & Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Li, Q. & Wang, S. (1998), 'A simple consistent bootstrap test for a parametric regression function', *Journal of Econometrics* **87**(2), 145–165.

Lin, Z., Li, Q. & Sun, Y. (2014), 'A consistent nonparametric test of parameteric regression functional form in fixed effects panel data models', *Journal of Econometrics* **178**(1), 167–179.

Lu, X. & Su, L. (2017), Determining individual or time effects in panel data models. Unpublished working paper.

Papageorgiou, C. & Spatafora, N. (2012), Economic diversification in lics: Stylized facts and macroeconomic implications. IMF Staff Discussion Note SDN/12/13.

Parmeter, C. F. & Racine, J. S. (2018), Nonparametric estimation and inference for panel data models. Department of Economics Working Papers 2018-02, McMaster University.

Parteka, A. (2007), Employment and export specialization patterns versus GDP per capita performance - unifying approach. Universita Politecnicá delle Marche Working Paper # 302.

Su, L. & Ullah, A. (2006), 'Profile likelihood estimation of partially linear panel data models with fixed effects', *Economics Letters* **92**(1), 75–81.

Sun, Y., Carroll, R. J. & Li, D. (2009), Semiparametric estimation of fixed effects panel data varying coefficient models, *in* Q. Li & J. S. Racine, eds, 'Advances in Econometrics: Nonparametric Econometric Methods', Vol. 25, Emerald, chapter 3, pp. 101–130.