Revealed Comparative Advantage. From binary to categorical.

Matias Iglesias[1]

Abstract— **In the context of international trade empirics, Revealed Comparative Advantage (RCA) is an index usually used to assess the competitivity of countries for exporting a given product.**
It has a strong limitation: RCA values are usually not comparable across countries, products, and time periods, with the exception of the level RCA = 1.

Here I propose a line of reasoning for distinguishing degrees of intermediate comparative advantage. I compute the probabilities of surpassing the threshold RCA = 1 within one time period. In this way, we have a metric of how far you are from RCA $\geq$ 1, i.e. how likely you are to observe RCA $\geq$ 1 in the next period.

I develop a framework in which RCA, the size factors and the exports of a country are disentangled, which makes the whole analysis clearer. We can see that smaller countries and products are more volatile in RCA because smaller export volumes are more volatile.

I. Introduction

The Balassa revealed comparative advantage (RCA) index is used in descriptions of specialization of economies, and technological innovation. It attempts to quantify the concept of competitivity put forward by Ricardo. Formally it is defined as [2]:

$$RCA_{cpy} = \frac{x_{cpy} \sum_c \sum_p x_{cpy}}{\sum_c x_{cpy} \sum_p x_{cpy}} \tag{1}$$

where x_{cpy} are country c exports of product p in year y.

This index is the ratio of actual exports x_{cp} to what could be thought of as an expectation for x_{cp}: the world export of the product times the country share in world exports. Or what is the same: the total exports of the country times the product share in world exports.

It has an interpretation especially as a boolean: is x_{cp} higher or lower than you would expect with this simple rule? i.e. is *RCA* $>$ *1* true?

Interpretations of the RCA values other than 1 have always been problematic. Comparison across countries, products and time periods is not consistent ([10], [3]). To get a grasp of the size effects in the RCA measure, consider one of the highest RCA observed for China [1], exporting raw silk (not thrown), i.e. HS 5002, in year 2003. It exports 240 m US$ out of 290 m US$ total world exports. RCA is 11.1. Even exporting the whole of the world exports, their RCA on this product can not go above 13.1, which is the inverse of the share of China in world exports. Small countries and products may on the other hand easily show RCA values of hundreds or thousands, as the denominator in eq. 1 shrinks.

Many alternative measures for RCA proposed in the literature are based on transformations of Balassa's index. One of the most disliked characteristics of this index is that its distribution is highly skewed (it goes roughly as 1 / RCA) and so the mean is usually high and has little meaning (cf. for example Laursen [1]). Also it does not go well with econometric regressions where gaussianity is a condition. Actually, export volumes lend themselves very naturally to a log scale. In such scale their distribution is essentially gaussian (see fig 7 in Appendix). Similarly, RCA values show a bell shape distribution in a log scale. However taking log of RCA is prematurely disregarded by many authors because of the issue of dealing with zero values.

The issue of comparability of RCA values across countries, products and years is also brought about often. Attempts to treat it include normalizations across some of these dimensions, or all of them, sometimes with rather cumbersome formulas ([10], [3]). On the other hand ignoring this artifact and using fixed thresholds (other than 1) for RCA is essentially incorrect, and conclusions thereof are not trustworthy. [2]

In this work I do not propose any tweak for obtaining a new index. For justifying my approach I would like to bring up the quote to Balassa in Hoen et. al.[8]: *One wonders, therefore, whether more could not be gained if, instead of enunciating general principles and trying*

[1]M. Iglesias is PhD candidate at the Scuola Superiore Sant'Anna, University of Pisa, Italy. `matuteiglesias@gmail.com`

[1]UN COMTRADE data. More details in section III

[2]This issue is for instance evident (although unnoticed by the authors) in Hinloopen et. al. [7] where Greece has higher than 4 comparative advantage in 40 industries and Germany in none (naturally, as it is a much larger economy)

to apply these to explain actual trade flows, one took the observed pattern of trade as a point of departure., with the difference that the subject is not the *trade flows* themselves but the *RCA index*: instead of applying ad-hoc transformations, why not see what the index is actually telling? The quality and quantity of exports data available nowadays allows us to see more directly what RCA values are and what they are not. Agnostic (data-first) observation, along with an appropriate analytical framework can help us understand better what RCA values imply, and how geometric factors (size of country, product and world market) have an influence.

In section II I show how taking log(RCA) can help interpret it as a difference between actual log exports of a country-product and a factor that depends on size of product, country and world exports. Nothing is new, but this framework makes the rest of the analysis much clearer (everything can now be mapped in 2D), and furthermore we will see that this size factor is key for visualizing the size effects in RCA.

I propose that two observations show *equivalent advantage scenarios* if they have equal *probability to show* $RCA > 1$ *in the following period* which we will call equal *pRCA*. I leverage on more than a million observations of RCA values in consecutive periods to uncover it empirically. This already shows us a diagnostic of what are the distortions inherent to Balassa's measure of RCA. Namely small economies show higher fluctuations of RCA and so they show the same chances of surpassing the $RCA = 1$ threshold with a lower RCA than big countries do. A similar situation holds for the chances of loosing the $RCA > 1$ status.

Remember we compute pRCA from the actual data. The variables that come into play are the current RCA and the size of country, product and world trade comprised in a single variable (cf next section). Some of this datapoints relate to events where RCA $>$ 1 the following year and some do not. To compute pRCA we need to describe a space where datapoints are layed out. We split the space in neighborhoods (I use quantiles to obtain equal population neighborhoods) and compute the fraction (within each of them) of successful events. This value is pRCA and is assigned to the whole neighborhood where it came from. Other algorithms suitable for this situation may be used instead, for example k-nearest neighbors.

Then I develop a model that explains the observed pattern solely by finding growth distributions of export volumes that depend only on export volumes themselves (x), i.e. $p(x, g)$, with a standard deviation that decreases with x.

The paper concludes with a brief account of how non binary RCA values would modify the definition of variables usually used in the literature. Essentially, the bipartite country-product network becomes a weighted one. The analysis by Hidalgo et al ([5], [6]), or that of Caldarelli et al.[9] can be extended naturally in this framework. Maybe the method presented here has more impact on the econometric analyses such as those of Boschma et. al. [4], as we provide a baseline for evaluating the performance of countries in exporting their products which takes account of the combination of the current RCA level and the non-linear interaction of the size factors which has so far been neglected.

II. Variables used throughout the analysis

Here I present the mathematical framework and notation. I define variables that are convenient for the analysis that follows.

It is very convenient to use the log RCA. [3]From equation 1:

$$\log(RCA_{cpy}) = \log(x_{cpy}) - \log(\sum_p x_{cpy}) -$$

$$- \log(\sum_c x_{cpy}) + \log(\sum_c \sum_p x_{cpy})$$

Just to ease the notation, define:

$$T_{wy} = \log \left(\sum_c \sum_p x_{cpy} \right)$$

$$T_{cy} = \log \left(\sum_p x_{cpy} \right)$$

$$T_{py} = \log \left(\sum_c x_{cpy} \right)$$

that is *log total trade volume* for the world, country and product respectively, for the year y.

$$log(RCA_{cp}) = log(x_{cp}) - (T_c + T_p - T_w)$$

[3]Many authors disregard the use of log(RCA) because in many cases RCA = 0. However we should keep in mind that the lowest export value recorded is 1000 US$. Hence for practical uses, 500 US$ is the same as 100, 10 or 0. When we take logs, anything that is below the thousand dollars threshold is just 'very low' and falls out of our range. Taking $\log(RCA+\epsilon)$ on the other hand would create important artificial distortions that we better avoid.

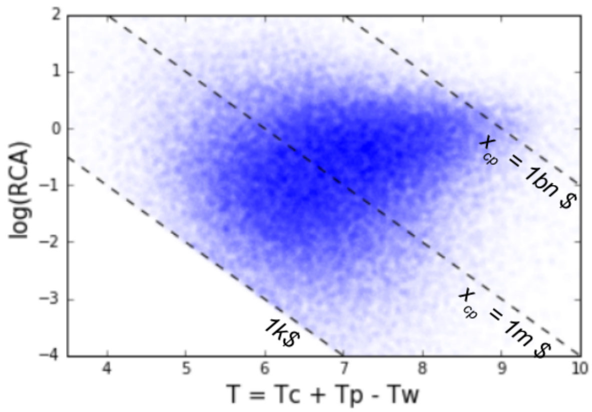

Fig. 1. Location of data points in the (T, log(RCA)) plane. $T = log(\sum_c x_{cp} \sum_p x_{cp} / \sum_c \sum_p x_{cp})$. This choice of variables simplifies the interpretation of how export volumes combine with size of country and product when computing RCA. Diagonal lines indicate levels of equal export value, hence, for example $log(x) = 6$ (i.e. one million US\$) implies a $log(RCA) > 0$ (i.e. comparative advantage) if $T < 6$ (i.e. country and product need to be small enough), and no advantage if they are too large (so that $T > 6$). The lower limit for export reports is $1000 US\$$, hence the empty region in the low left corner.

where we take for granted that observations belong to some year. In general:

$$log(RCA) = log(x) - T \tag{2}$$

where we naturally have dropped the indices for country and product. The subtlety here is that we are abstracting from the actual observations c, p, y and describing an always true relation between variables.

The only thing we did so far is to rewrite the definition of RCA in such a way that it is decomposed in exports $\log(x)$ and the size factor T. RCA can be interpreted as the comparison between $\log(x)$ and the size factor. The simple relation between the three terms means that we can plot two of them against each other and the third one will take constant values along diagonals of slope 1 (or minus 1). We can see this clearly in the plot of figure 1, where I included actual data points to provide a sense of the actual dimensions involved.

In the case of international trade data, we are lucky to have an additional handy property. T changes only negligibly when x changes[4]. For practical purposes we can consider the terms in equation 2 as independent, which means that x growth (additive in logs) is nearly equal to RCA growth. In plain language, if you duplicate your exports of some product, your RCA will duplicate too. While it is true increased x increases in the same amount country, product and world export market size, hence modifying T, this effect is negligible for all practical purposes. So growth of x can be taken to equal growth of RCA.

III. Description of data

I use data of export volume by country, product code (4 digit) and year (2003 - 2014) as provided in the Atlas of Economic Complexity. The data contains observations for 1200 SITC4 product classes, and 162 countries which presented observations in at least 800 products over the 12 year sample period.

The table below shows a sample of the raw data used:

entry no.	year	pcode	ccode	x_{cp} (US\$)
737241	2008	are	1001	64666264.85
253622	2004	aus	5109	1969466.61
1347202	2012	kor	3914	15178941.31
45748	2003	pak	3903	9921667.73
1614760	2014	tza	3002	1026272.00

Snippet of the raw data from which RCA values and all other variables in the present work are computed. This is UN COMTRADE data, downloaded through the Observatory of Economic Complexity

RCA values are computed with the 'calculateRCA function from *binet* python package.

From this data I pick up all possible pairs of observations, which belong to a given country, product and subsequent years. The focus is on relating RCA and export values to that of the following year.

IV. RCA. From binary to categorical

One of the most evident patterns is that there is a close relation between RCA > 1 at period t and the probabilities of RCA > at t+1. In other words, the main predictor

[4] From equation 2 for $\log(RCA)$, note that growth in export volume of product p and country c ($x_{cp,t}$), is not directly translated to growth in RCA because T_c, T_p and T_w will change with the new $x_{cp,t+1}$. This effect is however usually negligible. If $log(x_{t+1}) = g + log(x_t)$ and we define $\tilde{g}$ such that $log(RCA_{t+1}) = \tilde{g} + log(RCA_t)$, then to first order in g:

$$\tilde{g} \approx \left(1 - \frac{x_{cp}}{\sum_p x_{cp}} - \frac{x_{cp}}{\sum_c x_{cp}} + \frac{x_{cp}}{\sum_c \sum_p x_{cp}} \right) g$$

In the worst cases, the term in parentheses differs from 1 in 10^{-2}. Typically it will differ by 10^{-5} or less for having comparative advantage next year is to have it already this year. **The interesting part however, is that these probabilities take intermediate values from 0 to 1, then providing a natural interpolation between these two extremes**.

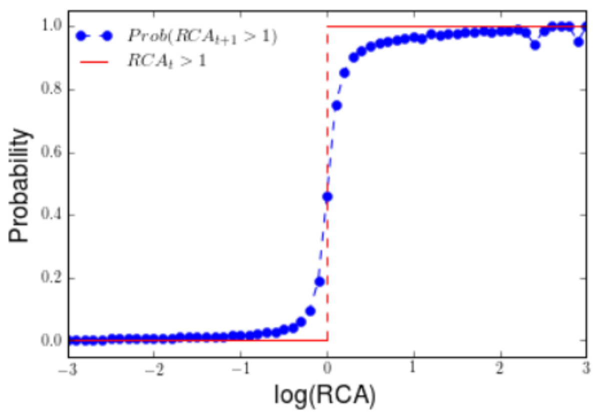

Fig. 2. The discrete variable RCA ≥ 1 (red), and the (theoretically continuous) probability of $RCA_{t+1} \geq 1$ (blue), given RCA_t.In the extremes both coincide, but near the threshold of RCA = 1, the latter one provides a natural interpolation between the two values. All observations aggregated.

I say natural, because $P(RCA_{t+1} \geq 1 \mid RCA_t)$ is giving us information on *how far* from the threshold RCA_t is. In other words, how hopeless an RCA < 1 is, or how secured an RCA > 1 is. Keep in mind this is empirical information that comes straight from the observations.

We showed that log(RCA) is composed of two factors: log(x) and the size factor T. What if we add them as a second axes and do the same computation of figure 2? We can just use T on the horizontal axis, and any dependence on log(x) would appear on the -45^o diagonals.

In figure 3 I plot the direct computation of the probability of observing $RCA \geq 1$ depending on both the current RCA (vertical axis) and the size factor T (horizontal axis). The result is striking: the size factor is key in determining the scale of RCA fluctuations. For obtaining this plot, neighboring data points have been arranged in equal sized groups (white dots signal a group center) and the probability within each group of observing RCA > 1 in the following year has been computed.

The horizontal line splitting the area in half is RCA = 1. We see near zero values in the lower half and near one values in the upper half. The most interesting feature is again the transition between these, which shows a clear dependence on T.

Here we see more clearly how RCA values are not comparable across countries. Consider, for example the value of RCA = 0.5 which is marked with a dotted line. For an average product, this means a higher than 15% chance of getting to $RCA > 1$ for small peripheral economies, a 5% chance for mid-sized developing countries, and between 1% and 2% for large European or Asian economies.

What is the meaning of the level curves in the plot of figure 3? If two different country-product-year export values (i.e. x_{cp}'s) have the same probability to show $RCA > 1$ the following year, can we say they are equally far from $RCA > 1$?

This is the concept that I want to put forward here, we can build a measure of distance to $RCA > 1$ (which can also be interpreted as a distance to $RCA < 1$: how likely you are to loose the developed product status).

V. Model

To understand better which factors come into play in shaping this dependence of the profile of RCA $>$ 1 probabilities on the size factor, we can develop a little analytical model. The approach to reconstruct the empirical observation in figure 3 will be to go one step behind: model the growth distributions and integrate them out to find probabilities of surpassing the RCA = 1 threshold. In the meantime we will uncover some of the explaining factors for what we observe empirically.

First note the close relation between the probabilities profile and the growth distributions (see figure 4). In the simplified case of a fixed growth distribution (PDF), the probabilities profile observed are the integral of the growth distribution (CDF). For example, if growth distribution (lower plot of figure 4) was a gaussian of fixed σ, the function in the upper plot will be the $Error function$. we will actually allow growth distributions to be different for each initial log(RCA) value, so the integration will be slightly more involved but the concept is the same.

The plot in figure 3 can be thought of a discrete counterpart $f_{disc}(T, log(RCA))$ of a continuous (modeled) probability function which we will call $f_{cont}(T, log(RCA))$, and takes values $p \in [0, 1]$. f_{disc} is computed simply as (python):

```
df.groupby(['log(RCA) bins','T
bins'])['RCA (t+1) > 1'].mean()
```

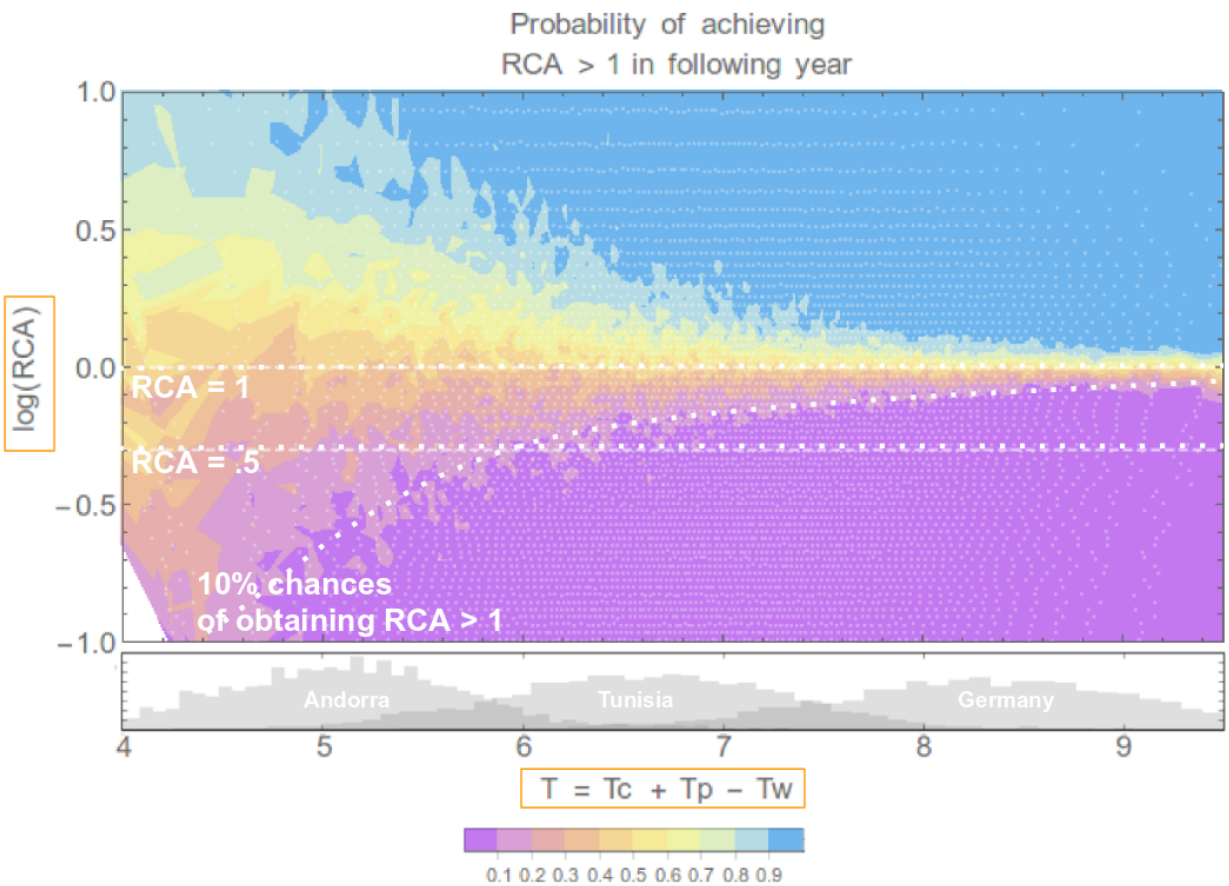

Fig. 3. Here I map the probabilities of surpassing the RCA = 1 threshold the following year, against the variables $log(RCA)$ and $T = log(\sum_c x_{cp} \sum_p x_{cp} / \sum_c \sum_p x_{cp})$. These probabilities are computed directly from more than a million observations of RCA in consecutive years, by grouping all observations by their rounded percentile in both variables. White dots indicate the centroid of each group. Below the horizontal axis, a reference of the distribution of T values for countries of different sizes. Dotted lines indicate fixed values of RCA, in this case RCA = 1 and RCA = 0.5, and a qualitative curve of constant chances to achieve RCA $>$ 1 the following year. If we say that the width of the transition between probabilities 0 and 1 is an indicator of the inherent scale of the RCA values involved, then we are observing in this figure how size of country and size of product affect the scale of the RCA index.

where df is a pandas dataframe containing the necessary data (including columns indicating binning of variables $\log(RCA)$ and T), `'RCA (t+1) > 1'` is a boolean variable telling if $RCA > 1$ the next year.

To get an analytical model of this, first consider an RCA growth distribution, that is, if $RCA_{y+1} = k\ RCA_y$ then $\log(RCA_{y+1}) = g + \log(RCA_y)$ (with $g = \log(k)$)

If we compute all g values for a group (remember groups are defined as neighborhoods in the T, $\log(RCA)$ plane), we would get a histogram where f_{disc} is the sum of all bins for which $g \geq log(1) - log(RCA_0)$ (see figure 4 for an illustration in the case we ignore T). In the continuous case, we should assume a probability distribution for growth $p(g)$ and then f_{cont} would be defined as:

$$f_{cont}(T, log(RCA)) = \int_{-log(RCA)}^{\infty} P_{T,log(RCA)}(g)dg \tag{3}$$

Of course it would be easier to go directly find a fit for the level curves of the plot in figure 3. However we would gain only a description with little foundation. Instead what I try to do with this model is to reconstruct how the shape observed in figure 3 comes about.

Model for the growth distributions P(g)

We can start by observing how the moments of growth distribution look in the (T, log(RCA)) plane. The standard deviation of g in the plane (T,

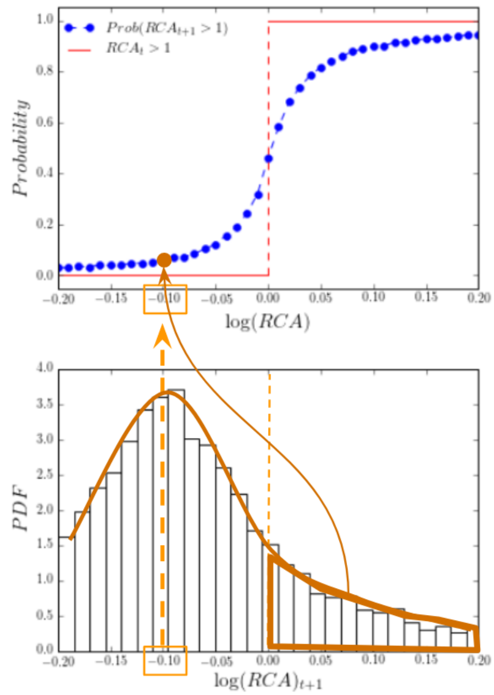

Fig. 4. Top: probabilities of reaching RCA $>$ 1 in the next period. Same plot of figure 2 but in a thinner horizontal range. Bottom: the histogram shows where all points of the bin 'log(RCA) = -0.1' ended up in the following period. Highlighted are those points which surpassed the RCA = 0 threshold. Their area is equal to the height of the corresponding dot in the plot above. The curve imposed suggests a continuous growth probability, which we should integrate to obtain the highlighted are and hence reconstruct the blue curve of above.

log(RCA)) is plotted in figure 8, Appendix I. It suggest that modeling growth distributions as dependent on $\log(x) = \log(RCA) + T$ is a decent first order approximation. I.e. standard deviation of growth decreases as export volumes grow.

If we make this assumption, we do not need both log(RCA) and X to precise the probability function, but we could express it as a function of $\log(x) = \log(RCA) + T$:

$$P_{X,lRCA}(g) = P_{log(x)}(g)$$

Now we are a step closer to modeling it. The dependence on log(x), written here as a subindex denotes that so far we do not have an expression for the dependence of P on log(x), i.e. $P(log(x), g)$.

At this point, we can take advantage of the fact that we can observe an empirical and discrete approximation to this $P_{log(x)}(g)$. We just have to compute it for all data points in a region and assign the computed value to that region. The result is plotted in the upper plot of figure 5. To proceed, I part this function into slices of $\log(x) \approx log(x_0)$ to obtain an empiric $P_{\log(x_0)}(g)$, for many different x_0 values (see inset 2 of figure 5). After that we will be in a condition to determine if there is such clear dependence of P on log(x).

In this analysis I do not pretend to discuss exactly what is the shape that these fixed-log(x) growth distributions have, but to capture their main characteristics, and verify that these characteristics alone can explain qualitatively the pattern observed in figure 3.

It can be seen that fixed-log(x) growth distributions, p(g), are compatible with Laplace distributions (inset 2 of figure 5). For simplicity, I assume mean growth $\langle g \rangle = 0$ in all cases. The normalized expression:

$$p(g) = \frac{1}{2b} Exp\left(-\frac{|g|}{b}\right)$$

where b is a parameter such that $\sqrt{2}b$ is the standard deviation of the distribution. Next we do a regression of p(g) for all slices, obtaining as a result a Laplace distribution fit characterized by a parameter $b(\log(x))$.

In this case we can also model the dependence on $\log(x)$. As expected, $b(\log(x))$ is decreasing in $\log(x)$ but what is more, if we take $\log(b)$ a roughly linear pattern is evident (see inset 3 of figure 5. It would be worth it to look further into this regularity, as well as explaining the residuals left, but for the moment we will just take the results as the stylized reconstruction of the growth probability function P(log(x), g) we set out for.

The conclusion: we have an analytical expression for the growth probability distribution P(x, g)

$$P(x, g) = \frac{1}{2b(x)} Exp\left(-\frac{|g|}{b(x)}\right) \tag{4}$$

$$b(x) = 10^{a_0} x^{-a_1} \tag{5}$$

with $a_0 = 1.02$ and $a_1 = 0.26$.

Finally, we can go back to the $(T, logRCA)$ plane, and integrate this $P(x, g)$ numerically (see eq 3) to obtain the modeled continuous version of the f_{disc} shown in figure 3 above. The result of comparing the model with the actual data is shown in figure 6.

We observe the same qualitative pattern, which suggests that indeed the higher variance of growth rates of smaller records explains why RCA values from smaller countries and products are inherently distinct from those of larger ones.

To recap, smaller countries and products need a smaller absolute export volume x_{cp} to show $RCA \approx 1$.

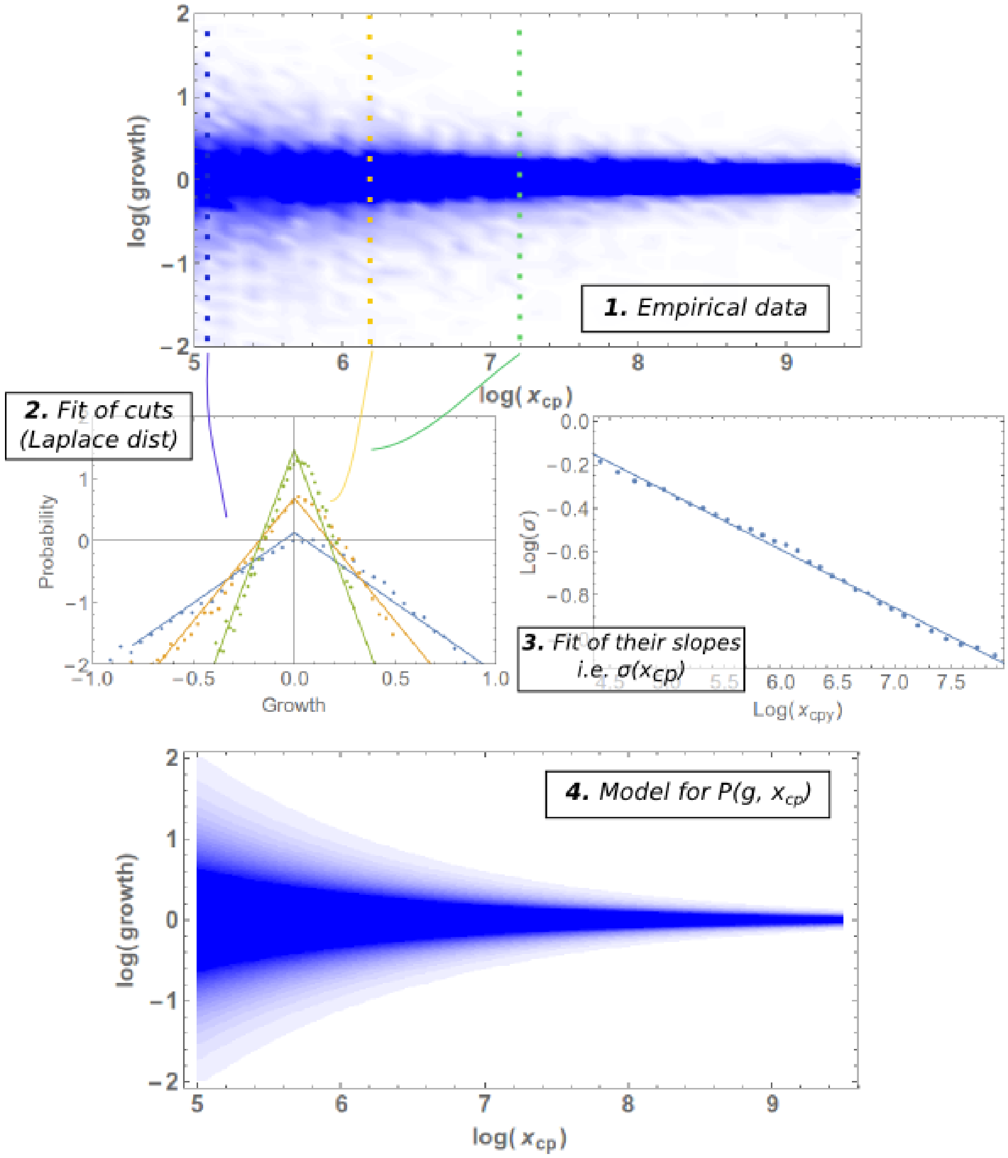

Fig. 5. Scheme of the process for modeling growth probability density function. 1. (top) actual distribution of growth rates (vertical axis) with log(x) on the horizontal axis. Intensity of color indicates probability. 2. (middle left) illustration of some cuts at fixed log(x) in log scale and their fit with Laplacian distribution functions. Note that the higher x is, the smaller the variance. The mean is constrained to be zero. 3. (middle right) plot of standard deviation of Laplace fits vs x, in log-log scale. It suggests a power law between the two (see eq. 5). 4. (bottom) model obtained for the growth probability function, $P(x, g)$ (see equation 4). It can be integrated to obtain the modeled level curves plotted in figure 6.

Together with the fact that smaller x_{cp} are more volatile, we get that RCA values equally lower than 1 describe a more advantageous situation in smaller countries than they do in large ones (in the case of higher than 1 RCA values it is a less advantageous one).

VI. Applications

Many recent works exploit the notion of RCA larger or smaller than 1 for deciding if a product is effectively exported by a country or not, hence building a binary country-product matrix (M_{cp}) and computing different metrics from there, such as product proximity, among many others. For extending this binary approach to a categorical or continuous one, we should rewrite M_{cp} in its multivalued form and redefine the variables usually computed from it. Without loss of generality, we can ask for the values of the new M_{cp} to range between 0 and 1. For all practical purposes, we have turned a binary bipartite network into a weighted one. We now have essentially the usual definitions for product ubiquity and country diversity:

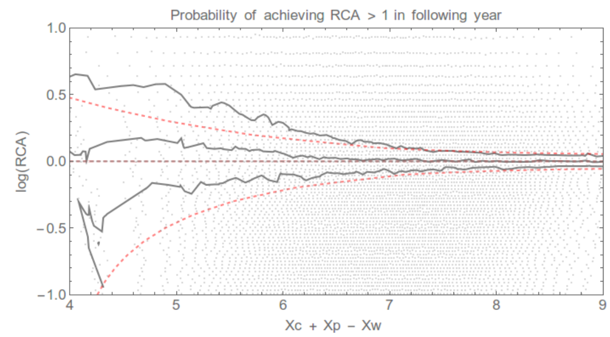

Fig. 6. Comparison of level curves (.25, .5, .75) for the actual data and the modeled one. We see the same qualitative pattern of higher chances of upgrading or downgrading at equal RCA and smaller T. For this model mean growth has been contrained to be 0 for simplicity, and so by definition the level of RCA = 1 will correspond to the .5 level curve (equal chances to be above or belor RCA = 1 next year).

$$k_{c,0} = \sum_p M_{c,p}$$

$$k_{p,0} = \sum_c M_{c,p}$$

Proximity cannot be defined from products' conditional probability of being exported by two countries. In plain language, countries no longer *'DO export'* or *'DON'T export'* a product. Rather they are competitive with different degrees in each of them. A natural choice for redefining proximity would be to use cosine similarity:

$$\tilde{\phi}_{i,j} = \frac{P_i \ . \ P_j}{||P_i||.||P_j||}$$

where P_i is a vector equal to the row of the M_{cp} matrix that corresponds to product i.

Now we can continue to naturally define the *weighted density* around good p, given basket of country c:

$$\tilde{\omega}_p^c = \frac{\sum_i M_{c,i}\tilde{\phi}_{p,i}}{\sum_i \tilde{\phi}_{p,i}}$$

APPENDIX I

APPENDIX II

REFERENCES

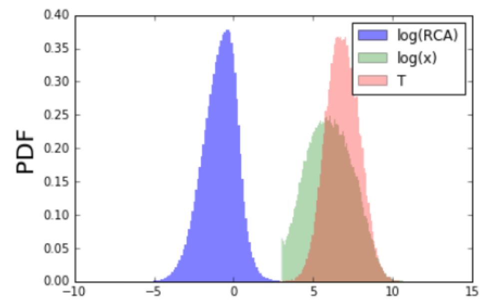

Fig. 7. Distribution of T, log(x) and log(RCA), which is by definition the difference between the previous two in each observation. The cut for log(x) ¡ 3 is noticeable. These distributions do not present the problem of high non-gaussianity, and their mean is closer the their median.

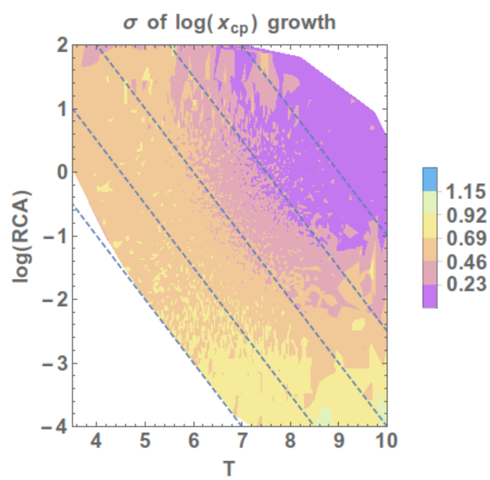

Fig. 8. Dependence of the standard deviation of growth distributions in the (T, log(RCA)) plane. Is it safe to model it as depending on log(x)?

[1] Technical report.
[2] B. Balassa. Trade liberalisation and revealed comparative advantage1. *The Manchester School*, 33(2):99–123, 1965.
[3] U. G. Bebek. Consistency of the proposed additive measures of revealed comparative advantage. *Economics Bulletin*, 31(3):2491–2499, 2011.
[4] R. Boschma and G. Capone. Institutions and Diversification: Related versus Unrelated Diversification in a Varieties of Capitalism framework. Papers in Evolutionary Economic Geography (PEEG) 1421, Utrecht University, Section of Economic Geography, Nov. 2014.
[5] C. Hidalgo and R. Hausmann. The building blocks of economic complexity. Papers, arXiv.org, 2009.
[6] C. A. Hidalgo, B. Klinger, A.-L. Barabsi, and R. Hausmann. The product space conditions the development of nations. *Science*, 317(5837):482–487, 2007.
[7] J. Hinloopen and C. Van Marrewijk. On the empirical distribution of the balassa index. *Weltwirtschaftliches Archiv*, 137(1):1–35, Mar 2001.
[8] A. Hoen and J. Oosterhaven. On the measurement of comparative advantage. *The Annals of Regional Science*, 40(3):677–691, 2006.
[9] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero. A new metrics for countries; fitness and products; complexity. Papers, 2012.
[10] R. Yu, J. Cai, and P. Leung. The normalized revealed comparative advantage index. *The Annals of Regional Science*, 43(1):267–282, Mar 2009.