The three-dimensional activity index – an alternative transformation to logarithmic calculus of the activity index

Joel Emanuel Fuchs (✉ jfuchs@uni-wuppertal.de)
Bergische Universität Wuppertal https://orcid.org/0000-0002-9951-3492

Research Article

Keywords: Activity Index, AI, logarithmic AI, LAI, three-dimensional Activity Index, Normalized AI, NAI, log-normal distribution, normal distribution, z-standardisation, European Patent Office, EPO, patent analysis

Posted Date: September 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-939228/v1

The three-dimensional activity index – an alternative transformation to logarithmic calculus of the activity index

Joel Emanuel Fuchs[1]

[1] jfuchs@uni-wuppertal.de
University of Wuppertal, Institute of Sociology, Gaußstraße 20, D-42119 Wuppertal (Germany)

Abstract

The Activity Index (AI) is a well-known index for comparing the contribution of different entities on various fields, for example scientific articles from different countries on various subjects, but it lacks important properties. We will show, that the AI is a log-normal distribution and that it is common in literature to transform the AI by the logarithm to a normal distribution. Last, we will present an alternative transformation special for longitudinal data, that transforms the AI to a normal distribution, too, without the negative properties of the logarithm.

Keywords: Activity Index; AI; logarithmic AI; LAI; three-dimensional Activity Index; Normalized AI; NAI; log-normal distribution; normal distribution; z-standardisation; European Patent Office; EPO; patent analysis;

Introduction

In sociological data analysis the comparison of data from different countries or institutions occurs all the time. The comparison of absolute values is often problematic due to the different sizes of the observed entities. A country like Germany, that has eleven times the landmass of Belgium and seven times of its population, will unavoidable be leading in producing output as scientific publications or other. It is common to use relative data if comparing different sized entities like countries, as for example publications per resident or publication per institution. One solution for comparing different sized entities is the activity index (AI). It enables the normalized international or inter-institutional contrasting of various fields and has the advantage, that no extra variable (as residents or institutions in the examples before) is needed. As we will see below, the AI sets the focal variable in relation to sums over subgroups of the focal variable, making it especially interesting for data sets, that have only one measuring variable per observation.

The AI lacks some essential qualities. As we will see below, its range is $[0, +\infty)$ and it is log-normal distributed, while many analytic techniques work better with a normal distribution of the data, for example Pearson's correlation, z-transformation, or with a two-sided open range, like some regression techniques. Most common solution to this is, to apply the natural logarithm to the data to receive a normal distribution. This will lead us to an independent index called LAI (logarithmic AI).

As mentioned above the LAI has a normal distribution, and so we can apply to it the z-standardization. If we do this on the macro level, i.e. all observations and once, we will state no major difference to the non-standardized LAI values, because they already tend to a z-standardization (even though we can show, that LAI values never will reach a z-standardization without normalizing them). On the micro level we have the longitudinal observation for a certain country/field combination. Even though in our example there are to less observation points (we observe five years) to decide if this data is normal distributed, because all shares are distributed randomly, we should find a normal distribution on each level and thereby the possibility to z-standardize observations for each country/field combination over our five years. The result is a comparison of performance on the micro level to determine, in which years a country perform over or under its mean especially for the focal field.

The second part of this paper will introduce a new index derived from the AI: the three-dimensional Activity Index or Normalized Activity Index (NAI). It is a transformation of the Activity Index specifically for use on longitudinal data – that immediately explains the third additional "dimension". Not only will we motivate this new index, but also show its superior qualities over the logarithmic AI, its inferior qualities, and mainly the similarities of both transformations. In the end it will replace the z-standardization of the LAI on the micro level and additionally to the AI will characterize the behaviour or performance (depending on the observation variable) of a country in a given field over the observed years.

We admit that this paper covers a very specific field, the analysis of single-variable data sets identifiable over two categorial variables and containing longitudinal data on their micro level. We hope that the presented information of AI and LAI extent the potential audience. Nevertheless, the additional analysis on micro level can help to differentiate data furthermore as we will see below. Hereby the answer to the question, how much information can be derived by a single variable and where starts just the fishing for new indices, is left to the reader.

European Patent Office

To demonstrate the important facts about AI, logarithmic AI (LAI) and NAI, we do not depend on data and could just discuss the pure theoretical approach. Alternatively, we could use fictional data, constructed especially for our showcase, like country A is performing on field 1. But we decided to use a real-world example for comprehensible reasons. As mentioned before, the NAI transformation needs longitudinal data. So, we decided to use data from the European Patent Office (EPO), rather to use the granted patents per field of technology and per country of residence for 2011-2015. The data is open and can be gained at https://www.epo.org. It is easy to understand, traceable and – most important for our case – longitudinal. Later, we will select special countries and fields to demonstrate some developments in the data. All these choices will be arbitrary to help understanding the formulas presented later. Please consider no aims behind the selections!

But first, let us analyse the given data set. We have 8,225 observations reflecting the number of patents that were granted by the EPO. Each observation is uniquely identifiable by the country the patent came from, the patents field, and the year it was granted, and it reflects the amount of granted patents for each triple country/field/year. So, our data set has four variables. Because we are focussing on the years 2011-2015 we have – of course – five years to differ. Also, 46 individual countries and areas are mentioned and one residual category to sum up all remaining countries (called "other countries"). Last but not least, the patents are divided into 35 different fields. The three identifiers make 8,225 ($= 5 \cdot 47 \cdot 35$) observations possible, that is the number of observations of our data set. We conclude, that we have a comprehensive survey.

Table 1:Key data of the EPO data set.

Observations	Years	Countries	Fields
8,225	5	47	35

The analysis of the absolute data can give us some interesting information, but not all. However, let us first focus on this part of analysis. As we can see from the histogram in Figure 1, 36.8% of the observations are zero (not shown in Figure 1 due to the logarithmic scale), 30.1% are between 1 and 10, 24.1% are between 11 and 100, and 9.0% are over 100. It is no wonder, that the data is right skewed in this way. Patents are not generated by countries, but by institutions and companies of these countries, meaning larger countries generate more patents – beside other country specific differences as how many resources are invested into research and development and other factors not mentioned here. Table 2 shows the five countries with the most granted patents in our time period.

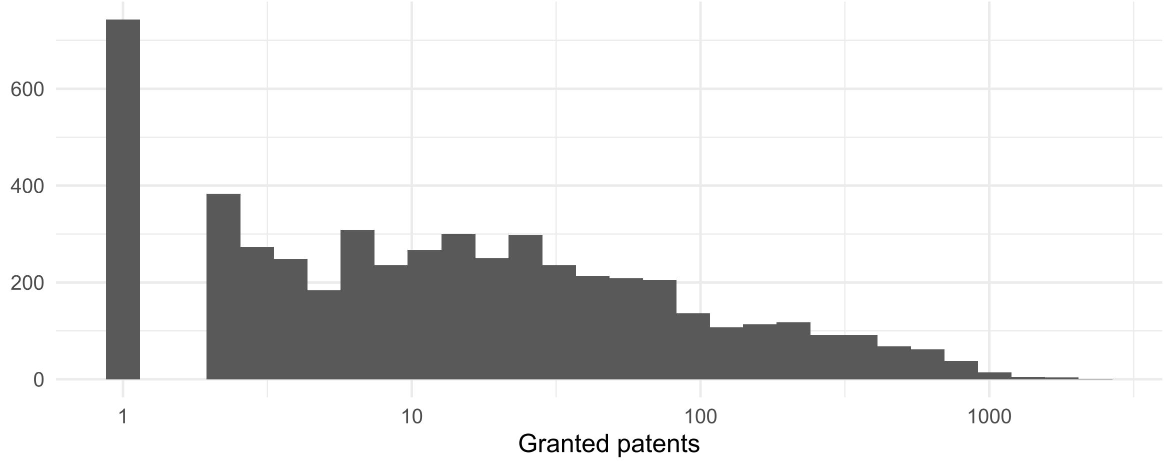

Figure 1: Histogram of number of patents granted per country, field and year (zero excluded; logarithmic scale on x axis).

Table 2: Five highest values of granted patents of our data set aggregated by country.

Country	United States	Germany	Japan	France	Switzerland
Patents	72,293	67,518	58,339	24,673	13,629

Already between these five countries, there are huge differences in the amount of granted patents, so we get a first feeling for the problem of analysing absolute data of our set. We have some few big players (United States, Germany, and Japan), dominating our observations, and making it problematic to find country specific characteristics. In the next example, we make this last point clear, why their domination disguises the characteristics of other countries.

Let us have a look at the two countries Germany and the United Kingdom (UK), that has 10,200 patents between 2011 and 2015, and the two fields 'Food chemistry' (FC) and 'Semiconductors' (SC). On purpose we chose one of the big players, but the remaining choices of country and fields are arbitrary. Figure 2 visualizes the observations of these two countries and fields over the years 2011-2015. As expected, Germany is dominating both fields in absolute numbers of granted patents. Table 3 complements the graphical information with the mean over the years.

Table 3:Arithmetic Mean of granted patents of the years 2011-2015.

	Germany	UK
Food chemistry	68.4	34.4
Semiconductors	189.8	19.8

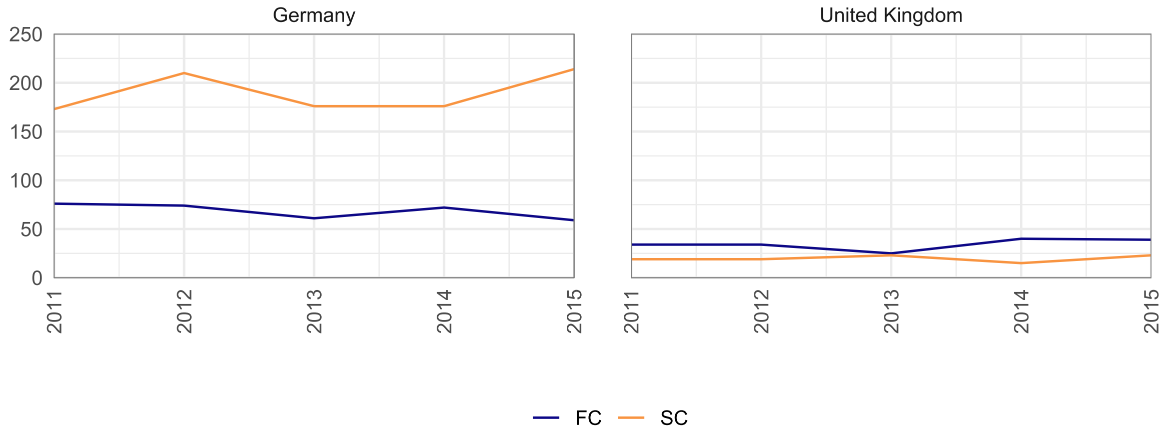

Figure 2: Granted patents.

Even though Germany is clearly dominating (just Germany's FC has a higher mean of granted patents than UK's FC and SC means together), the UK reflects a contrary situation to Germany regarding the two fields FC and SC. While in Germany patents of SC are dominating over the other field, in the UK patents of FC are dominating the granted patents of SC. We can derive further information. For example, it seems that patents of SC in Germany are dominating patents of FC more intensive than patents of FC in the UK are dominating patents of SC. Of course, this here used dominance term has to be filled with concrete numbers and criteria. If we want to describe this dominance further, one way would be to normalize the observations by the highest mean of each country.

At this point we constitute three facts. Fact 1: Even though we have some dominating countries, the proportions of fields can vary from country to country not only in their amount but especially in their relation to other proportions. Fact 2: Even analysing two countries and two fields raise problems. Absolut numbers between countries are difficult to compare because countries are differently active in granted patents ("big" versus "small" countries). Fact 3: Normalizing absolute numbers of two countries is problematic, because there is no objective choice which country and which field should represent the normal level. And even doing so, looking at 47 countries means having 1,081 pairs of countries, and means in the worst case 1,081 different normalizations of data (without considering the different fields).

Below we will present the Activity Index (AI), that can be calculated for each country and field combination at once, so that we have only one normalization for all observations. Additionally, it evens the differences in size between the countries out. Therefore, it concentrates on the proportions of the fields in one country and make it easy to compare these values with another one. It seems to be the response to our facts/problems, even though it has its disadvantages, too.

The Activity Index

We will have a look at a relative index, that is suitable for revealing the presented differences between countries regarding the underlining fields. We will call it the activity index (AI), as denominated by Narin et al. (1987). But it is also known under the revealed technological advantage (Soette & Wyatt, 1983), revealed comparative advantage (Balassa, 1965) or the Balassa index (lbid.).

Let D be a data set with a summable variable v, i.e. each observation is a non-negative integer, and each observation of D is identifiable by the unique combination of two categorial variables

I and *J,* e.g. *I* for a country and *J* for a field. The observation of the variable *v* identified by *(i,j)*, hereby *i* is a level of the variable *I* and *j* of the variable *J,* can be denoted by v_{ij}. Further, let *D* be a comprehensive data set regarding both categorial variables *I* and *J*, meaning, each observation v_{ij} exists in our data set. Than we define the AI as

Equation 1: Activity Index (AI)

$$AI_{ij} \coloneqq AI(v_{ij}) \coloneqq \frac{v_{ij} / \sum_{j \in J} v_{ij}}{\sum_{i \in I} v_{ij} \,/\, \sum_{i \in I, j \in J} v_{ij}}.$$

The AI often operates on single-value data sets like our EPO data, that allows to use the expressions AI_{ij} and $AI(v_{ij})$ equally. We conclude, that the AI is a relation of shares, as which we want to formulate it below to gain a deeper understanding. Below we will sometimes use the word share as synonym for a number $x \in [0,1]$, that is the relevant property of a share. It follows a more verbal version of the AI.

Equation 2: verbal version of AI.

$$AI_{ij} \coloneqq AI(v_{ij}) \coloneqq \frac{\text{The share of } v_{ij} \text{ on all accumulated obs. of } i.}{\text{The share of all accumulated obs. of } j \text{ on all accumulated obs..}}$$

That means, the AI compares the share of field *j* in a country *i* with the average share of field *j*.[1] We can conclude some properties of the AI directly from the formula. If we have $AI_{ij} > 1$, then the share of field *j* in country *i* is higher than the average share of field *j*. Contrary, if $AI_{ij} < 1$, then the share of field *j* in country *i* is lower than the average share of field *j.*

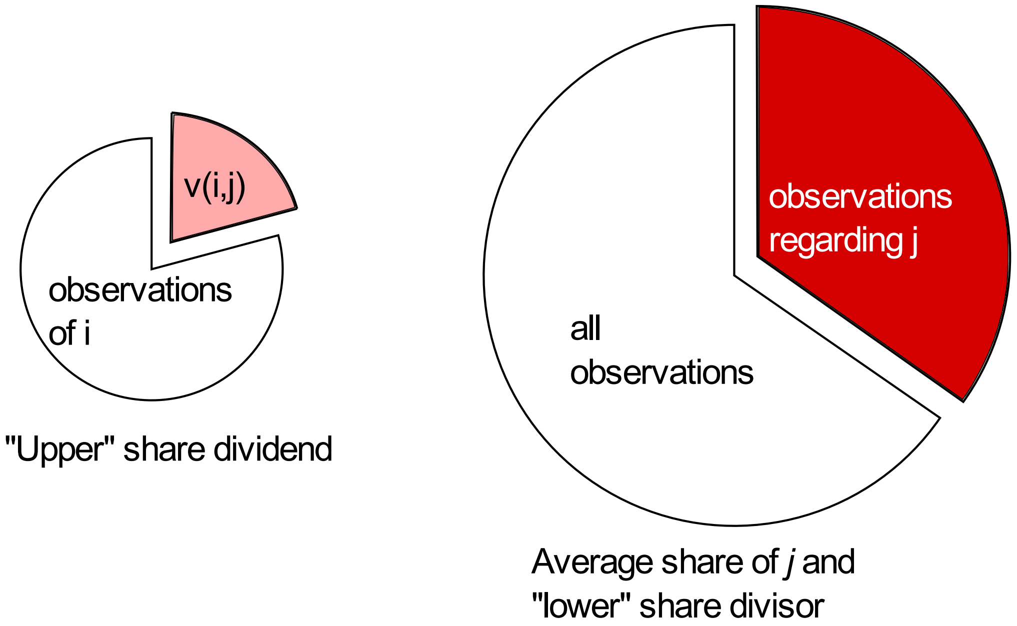

Figure 3: Visualization of the shares of the AI with AI < 1.

[1] In most literature, average is a synonym of the arithmetic mean. The arithmetic mean of the share of field *j* is $\frac{1}{\#I}\sum_{i \in I} \frac{v_{ij}}{\sum_{j \in J} v_{ij}}$, where *#I* is here and everywhere else the number of categories in *I* (e.g. countries). If we talk of the average share of field *j*, we mean the overall share of *j* in the data set, i.e. $\frac{\sum_{i \in I} v_{ij}}{\sum_{i \in I, j \in J} v_{ij}}$.

The AI has some properties and advantages with respect to our mentioned facts or problems above. Fact 1: although we have 47 countries in our set, the reference for each country is the same, that is the average share of *j*. Fact 2: Size of countries does not matter. We compare two shares, that are size independent. Fact 3: The computation of the AI can be done for all countries and fields equally. There is no choice done depending neither on countries nor on fields.

We computed the AI for all of our 8,225 observations by year as seen in Figure 4 (country and field only identify an observation uniquely by year, so any other calculation is not possible). Unfortunately, we lose 350 observations, that means if one of the sums of Equation 1 is zero, the AI cannot be computed. This happens for example at North Macedonia, that did not have one granted patent in all years and fields. Missing AIs are note really a constraint, because if a country has no contribution at all, why compare it with the others. We see a log-normal distribution at once, to which we will refer below in more detail.

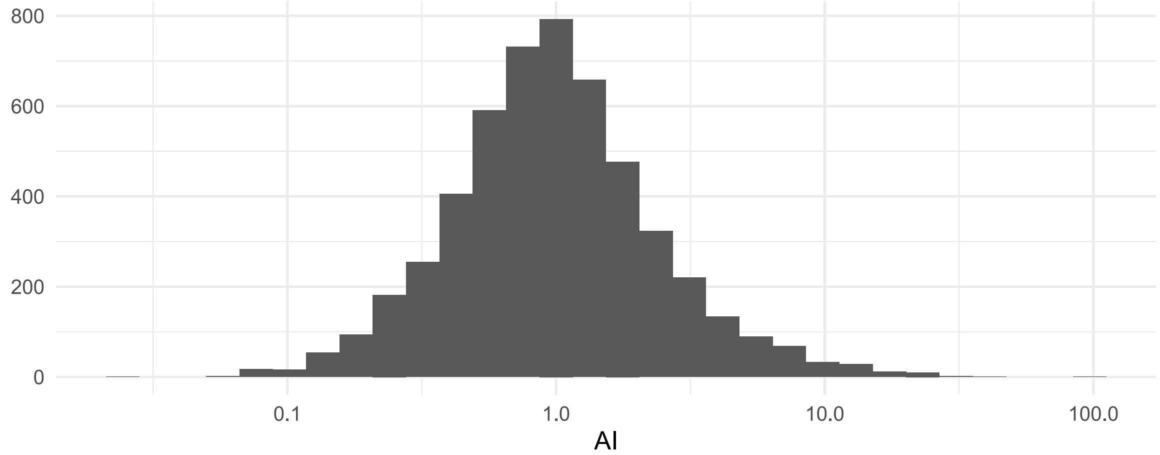

Figure 4: Activity Index for granted patents in 2011-2015 (zero excluded; logarithmic scale on x axis).

What does this mean for our example? Figure 5 shows the AI for the countries Germany and UK regarding both fields FC and SC. This figure is complemented by Table 4 showing the average AI values of these country field combinations.

We derive some expectable and some surprising information from the AI. The AI of Germany regarding our two fields FC and SC performs like expected. Even though both AI time lines are below 1 (meaning the share of FC respectively SC on all German patents is less then the

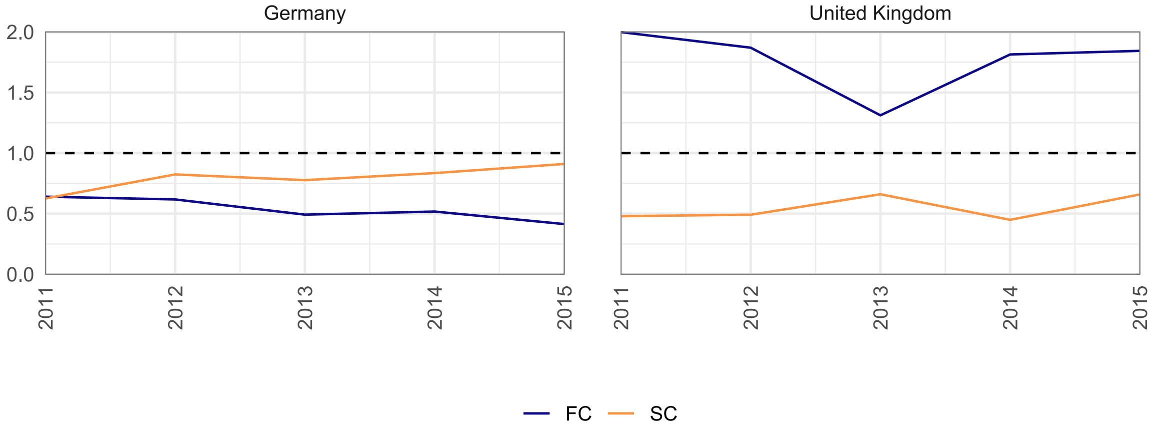

Figure 5: AI of granted patents.

average), they are correlated to Germany's absolute number of granted patents (Spearman's correlation coefficients for absolute numbers and their AI are 1 for FC and 0.8 for SC). Having a look at the UK, we derive a new information. Table 3 indicated, that both fields FC and SC are treated nearly equally by the UK, and this may be true in absolute numbers. In international comparison, we see, that the share of granted patents in FC is 1.77 times the average share of FC, so the UK places emphasis on this field. On the other hand, the share of granted patents in SC is only 0.55 times the average share of SC. So, the international community puts more effort on this field than the UK. To mark this point, we have to fields of a country, that are not really distinguishable by absolute numbers, but are very differentiable by their AI.

Table 4: Arithmetic mean of the AI of granted patents 2011-2015.

	Germany	UK
Food chemistry	0.536	1.767
Semiconductors	0.794	0.548

Having a relative index as the AI does not make a statement, in which field someone is successful. Especially the situation of Germany shows a great disadvantage by the AI. In absolute numbers it outnumbers the UK on both fields FC and SC, probably meaning, that it is leading on both fields in comparison with the UK. Looking at the AI, UK has a way higher AI in FC than the "big" Germany – but from this, we cannot conclude, that the UK is superior on the field FC. It only means, the UK is more engaged in the field FC with its possible resources than Germany, even if Germany has this huge amount of granted patents on a field, that it seems to put no effort on. How can we explain this phenomenon, having high absolute numbers but very low AI values? There can be many reasons, but always remember, examining the AI we examine shares, that means if a country has a high share in one field (in comparison to the average share), it has to have a lower share in any other field without the necessity to underperform on a certain field, just because all shares of a country have to sum up to 1. This means, every country will possess higher and lower shares and higher and lower AIs consequentially.

In short, we want to examine some properties of the AI besides the above mentioned. We already talked about the domain of the AI, a non-negative integer variable uniquely discriminable by two categorial variables (even though we could expand the domain to non-negative numeric variables). The range of the AI is clearly $[0, +\infty)$, because a share is located in [0,1], so a relation of shares (using Equation 2) is located in $\frac{[0,1]}{[0,1]} = [0, +\infty)$.[2] More interesting is the distribution of the AI. The AI as a transformation of a variable or of a set of observations is always finite and, in this sense, it cannot possess a continuous distribution. Nevertheless, we can see in Figure 4 that AI values tend to be log-normal distributed with an expected value of 1. This case is not an exception but a rule, although it would go beyond the scope of this paper to show this.[3]

[2] Here we use interval calculation. The Division of two intervals is defined by: $\frac{[a,b]}{[c,d]} := \left[\frac{a}{d}, \frac{b}{c}\right]$. The division through 0 is solved by reformulating the interval [0,1] by $\lim_{n\to\infty} \left[\frac{1}{n}, 1\right]$. We get $\frac{[0,1]}{\lim_{n\to\infty}\left[\frac{1}{n}, 1\right]} = \lim_{n\to\infty} \left[\frac{0}{1}, \frac{1}{1/n}\right] = \lim_{n\to\infty} [0, n] = [0, \infty)$.

[3] One way to show this is interval calculation again. We divide the interval $[0,1]$ into n equal sized subintervals $\left[\frac{i}{n}, \frac{i+1}{n}\right]$ for $0 \leq i < n$. Then we presume, that every subinterval has the same probability to contain a share. If we have two shares $x, y \in [0,1]$, we get a certain probability for $\left[\frac{i}{j+1}, \frac{i+1}{j}\right]$ to contain the share $\frac{x}{y}$, denoted by $P\left(\frac{x}{y}; \left[\frac{i}{j+1}, \frac{i+1}{j}\right]\right)$. For an arbitrary interval $[n_1, n_2] \subset [0,1]$ we have $P\left(\frac{x}{y}; [n_1, n_2]\right) = \sum_{0 \leq i,j < n} P\left(\frac{x}{y}; [n_1, n_2] \cap\right.$

The logarithmic AI

Having log-normal distributed data brings the idea close to apply the natural logarithm to it. We will call this the Logarithmic AI (LAI). In literature, the LAI is also known by the term revealed patent advantage (Grupp, 1994). It is defined as

Equation 3: The logarithmic Activity Index (LAI).

$$LAI_{ij} \coloneqq \ln\left(AI_{ij}\right).$$

Even though it transforms the log-normal distributed data of the AI into normal-distributed data preparing them for further analysis, it has two disadvantages. But first let us characterise the newly achieved data. The range of the LAI is $(-\infty, +\infty)$, that is the logarithm of the range of the AI. Its expected value tends to 0 $(= \ln(E[AI]))$ if the number of observations[4] tend to infinity. See Figure 6 for a histogram of the LAI of granted patents (please observe the similarity to Figure 4).

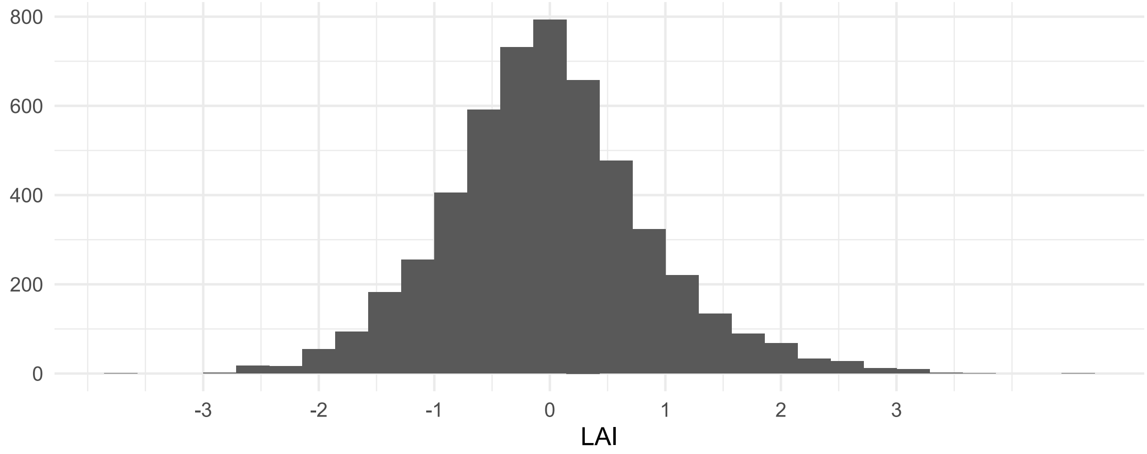

Figure 6: Logarithmic Activity Index for granted patents in 2011-2015.

Even though a normal distributed index has some advantages right concerning the use of some further analysis techniques like the z-standardisation, it has some disadvantages, too. As mentioned before, computing the AI reduced our data set from 8,225 observations to 7,875 observations. This is further reduced to 5,197 observations, because 2,678 observations have an AI = 0 and cannot be transformed by the logarithm. This corresponds to a loss of data of 37%.

Another disadvantage is the interpretation of the LAI. As well as before, LAI values above 0 (instead of 1) indicate, that the share of a field in a country is larger than the field's average share and a LAI value below 0 indicates the opposite. But while we could say about an AI of

$\left[\frac{i}{j+1}, \frac{i+1}{j}\right]\Big)$. If $[n_1, n_2] \to \frac{x}{y}$ and $n \to \infty$, we get the log-normal distribution for P. If the AI values are independent and identically distributed, the AI approximates a log-normal distribution, if the number of observations tend to infinity. We can assume, that AI values are in most cases randomly distributed; e.g. in our example there is no higher control that determines which country has to engage in which field.

[4] If we demand the observations tending to infinity, we here and below mean a uniform extension of the categories country, field and year. Because every observation is uniquely identifiable by this three, they have to gain more factors while the number of observations is growing. Also, we demand, that only finitely many observations belong to each country and the same for fields and years.

1.5, that its related share is 1.5 times larger than the average share, a LAI of 1.5 will not tell us anything else about the relation between related share and average share besides that the related share is larger, larger even than any other share with a LAI less than 1.5. So, a level of interpretation is lost that we have if we consider the AI.

We claim without further proof, that the z-standardization of the LAI does not lead to further information, because with a rising sample, the LAI is nearly normal distributed with an expected value of *0* and a deviation of *1*. Indeed, we have an expected value of *-0.02* and a deviation of *0.89* for our data set. Furthermore, comparing the AI with its derivation the LAI does not expose any new information as we can see in Figure 7 for the countries Germany and UK and the fields FC and SC. We conclude, that the LAI as independent index has no significance and is only important for further analytic techniques.

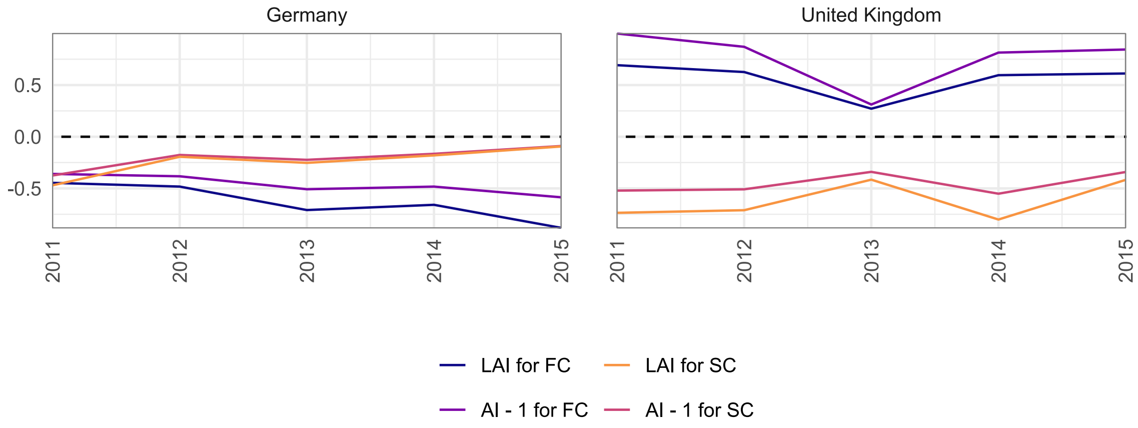

Figure 7: Comparison of AI (shifted down by 1 to scale to LAI) and LAI of granted patents.

It seems, that all we did in this section to this point has no additional value to us. To counter this appearance, we want to change the level of our examination. So far, our focus lied on all AI values respectively all LAI values at once, thus the macro level. But of course, we have different levels of focus: the population, the countries, the fields and the years. Our example is a focus on the micro level, thus on special country/field combinations over the years. From a stochastically point of view, we should have a normal distribution on each level, namely the LAI values of each country and field combination should be normal distributed. In our data set this is not true, but this can be explained by the low number of cases (years) for each country/field combination and by a certain degree of autocorrelation. Anyway, we take this theoretical approach to motivate the z-standardization of the LAI values for each country/field combination across the years 2011-2015. Therefore, we have to expand the index of our observation by a year notation (although we already calculated our AI values by year). Let v_{ijt} be the granted patents of country *i* in field *j* and year *t,* and *T* the set of all years. The AI changes slightly to $AI_{ijt} := AI\left(v_{ijt}\right) := \frac{v_{ijt}/\sum_{j\in J} v_{ijt}}{\sum_{i\in I} v_{ijt}/\sum_{i\in I, j\in J} v_{ijt}}$, while the LAI changes to $LAI_{ijt} := LAI_{ijt}\left(AI_{ijt}\right) := \ln\left(AI_{ijt}\right)$. Finally, we get the country/field specific z-transformation by

Equation 4: z-standardisation for each country x field combination.

$$z\left(LAI_{ijt}\right) := \frac{LAI_{ijt} - \frac{1}{\#T}\sum_{t\in T} LAI_{ijt}}{\sqrt{\frac{1}{\#T}\sum_{t\in T}\left(LAI_{ijt} - \frac{1}{\#T}\sum_{t\in T} LAI_{ijt}\right)^{2}}}.$$

We call the z-standardisation of the LAI for each country/field combination *the over years z-standardized LAI*. Remember, that for our 47 countries and 35 fields this actually means to compute 1,645 different z-standardizations. Of course, this operation is only reasonable, if we have longitudinal data, marking a critical point in this paper. The calculation of AI and LAI can be done on two-dimensional data (using a numerical variable uniquely identifiable by two categorial variables – lacking a time variable). But from now on – working with the z-standardization over years – we presume that our data is longitudinal. And with the z-standardization over years, we do get new information out of our data as Figure 8 shows.

We want to analyse a single country/field combination, that will be Germany and FC, and leave the other three combinations to the reader. Germany's FC values drop from 76 to 59 with a mean of 68.4 and a standard deviation of 7.8. The average standard deviation of all country/field combinations is 35.2, so we claim that Germany's FC values tend to a constant value between 59 and 76. We get a totally other picture of Germany's FC performance, if we look at the AI values. They are falling from 0.64 to 0.41, that means the FC share on granted patents in Germany is less then the average share, loosing over time, and it seems that we have a trend here. To distinguish this further we have a look at our z-standardised values. Here we can see (Figure 8) that Germany performs in the years 2011 and 2012 over its own average. Even though absolute numbers and AI indicate a comeback of Germany's FC granted patents in 2014, the z-standardisation clearly shows that the average performance (in the sense of the FC share in Germany) is not reached after 2012. We can distinguish a Germany before 2012 (year included) and after 2012 (year excluded). This short example shows that the z-standardisation over the years enables us to connect important points of trends with certain years.

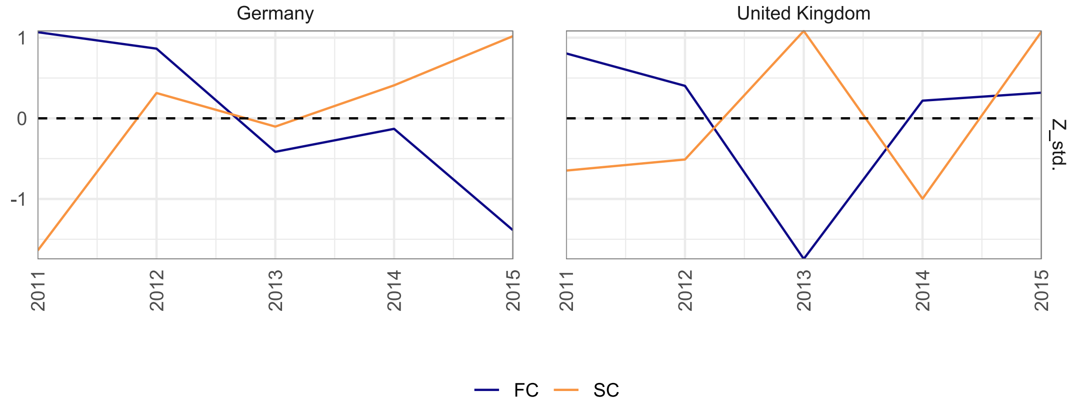

Figure 8: Granted patents as over years z-standardised LAI.

A new index: the three-dimensional AI

Calculating the z-standardisation for each country/field combination over the years for the logarithm of the Activity Index of our original data set and losing 37% of the data by the logarithm only for the information of relating trend points to certain years does not sound very offering. We have some computational objections. There is the loss of data by applying the logarithm, but also the effect of round-off errors when generating lots of different z-standardisations (we remember the 1,645 country/field combinations of our data set). We have some theoretical objections, too. How comparable are the z-standardised values of two country/field combinations? After all we put them together in Figure 8 indicating without proof that they should be compared. And while the idea of relation of shares thus the idea of the AI to compare different countries on the same field seems kind of natural (like in Figure 3), the z-standardisation of every country/field combination seems to be arbitrary. Why not z-standardise just every country or every field?

Below we proclaim a new index called the Normalized Activity Index (NAI) for longitudinal data. We will derive it from the AI hopefully by a way that is comprehensible, turning out this derivate to be a kind of natural development. We will show, that the NAI has the same advantages as the over years z-standardised LAI, but less disadvantages. And last but not least we will proclaim, that in a certain sense the NAI generates the same values as the over years z-standardised LAI making the last one overdue for the interpretation of granted patent numbers.

Using longitudinal data, it is obvious to consider the mean of each country/field combination as we have done before, for example when we talked about the absolute values of Germany's granted patents in FC. The arithmetic mean of absolute values on the micro level (distinguished by country/field) is $\overline{v_{ij}} = \frac{1}{\#T}\sum_{t\in T} v_{ijt}$ (we are using the notation $\overline{v_{ij}}$ without the index *t* to stress out that the mean over the time dismissed the dimension *T*). For the AI there are at least two methods to calculate the mean on the micro level. The first one is the arithmetic mean of the AI, i.e. $\overline{AI_{ij}} = \frac{1}{\#T}\sum_{t\in T} AI_{ijt}$. This formula is only true, if the AI_{ijt} of the selected *i* and *j* exists for all years. As we have seen before, this is only true for 7,875 of our 8,225 cases. The second one is the AI of the arithmetic means, in formula

Equation 5: The AI of arithmetic means or the Temporal Activity Index (TAI).

$$TAI_{ij} \coloneqq TAI(v_{ijt}) \coloneqq \frac{\overline{v_{ij}}/\sum_{j\in J}\overline{v_{ij}}}{\sum_{i\in I}\overline{v_{ij}}/\sum_{i\in I,j\in J}\overline{v_{ij}}} = \frac{\sum_{t\in T} v_{ijt}/\sum_{j\in J,t\in T} v_{ijt}}{\sum_{i\in I,t\in T} v_{ijt}/\sum_{i\in I,j\in J,t\in T} v_{ijt}}.$$

We call the AI of the arithmetic means also the Temporal Activity Index (TAI) to express, that we get the mean of the AI over the years (in this sense temporal), even though by dismissing the years the TAI is independent of them. It exists as long as at least one observation of all fields and all years of country *i* is positive and one observation of all countries and all years of field *j* is positive. This is a weaker demand than the one regarding the arithmetic mean of the AI. A result of this is, that we have 1,610 TAI values for the 1,645 possible country/field combinations (or 8,050 TAI values for our 8,225 observations), but only 1,540 arithmetic means calculated on all five years. The TAI_{ij} seems to be more robust and – more important – delivers to every AI value a corresponding TAI value.

Table 5: Comparison of Arithmetic Mean of AI and TAI, both calculated for the years 2011-2015.

	Mean of AI		TAI	
	Germany	UK	Germany	UK
Food chemistry	0.536	1.767	0.530	1.766
Semiconductors	0.794	0.548	0.790	0.545

Table 5 shows that the mean of AI and the TAI deliver quiet similar results. Indeed, the root mean square error (RMSE), i.e. $\sqrt{\frac{1}{\#I\cdot\#J}\sum_{i\in I,j\in J}\left(\overline{AI_{ij}} - TAI_{ij}\right)^2}$, is 0.25 and we claim, that this number tends to zero, if the number of observations tend to infinity.

We define the TAI as the average of the AI for one specific country/field combination over the years and consequentially we can normalize our AI values by dividing them through the TAI values. Often the average is subtracted, so why normalize by division? Three facts argue for normalization by division rather than for normalization by subtraction. Fact 1: The AI is a relation of shares and the normalization should keep the properties of a relation. Subtraction would not do this, but division does. Fact 2: We z-standardized the LAI, meaning we normalized it by subtracting its mean. Because the LAI is a logarithm, we get: $LAI(AI_{ijt}) - \frac{1}{\#T}\sum_{t\in T} LAI(AI_{ijt}) \; = \ln\left(\frac{AI_{ijt}}{\sqrt[\#T]{\prod_{t\in T} AI_{ijt}}}\right)$. We conclude, that the z-standardization is achieved by dividing the AI through another from the AI derived index, here $\sqrt[\#T]{\prod_{t\in T} AI_{ijt}}$. This indicates to prefer division over subtraction. Fact 3: By division we will gain symmetry of the indices *i*, *j* and *t*. This would not be the case if we would normalize by subtraction.

All after all, we define our new index, the Normalized AI (NAI), by the normalization of the AI through division by the TAI, in formula

Equation 6:The Normalized Activity Index (NAI)

$$NAI_{ijt} \coloneqq NAI(v_{ijt}) \coloneqq \frac{AI(v_{ijt})}{TAI(v_{ijt})} = \frac{v_{ijt} \cdot \sum_{i\in I, j\in J} v_{ijt} \cdot \sum_{i\in I, t\in T} v_{ijt} \cdot \sum_{j\in J, t\in T} v_{ijt}}{\sum_{i\in I, j\in J, t\in T} v_{ijt} \cdot \sum_{t\in T} v_{ijt} \cdot \sum_{j\in J} v_{ijt} \cdot \sum_{i\in I} v_{ijt}}.$$

We refer to this index by two different names: as three-dimensional Activity Index, because the observations have to be distinguishable by three different indices or dimensions i, j and *t*. In our example these dimensions are the country, the field and the year of each granted patent. But the name three-dimensional Activity Index shall also express a property of this index. It is symmetric in its dimensions. The classical AI is also symmetric in its dimensions, i.e. $AI(v_{ijt}) = AI(v_{jit})$ if we would alter the first two indices. For the new index we have the equality $NAI(v_{ijt}) = NAI(v_{jti}) = NAI(v_{tij}) = \cdots$. So, we can alter all three indices as we wish with the same result. On the other hand, the name three-dimensional AI – probably with the abbreviation tdAI or 3dAI – is kind of unhandy. In the end we decided us for the name Normalized Activity Index (NAI), because it is more pleasant and still describes the new index precisely.

The quality of interchangeable indices, that the NAI possesses, seems to be rather irrelevant, but it is a very strong one. Till now we always considered the share of a field *j* in a country *i* in relation to the average share or – what is the same – the AI$_{ij}$. We calculated the AI by year and so we gained AI$_{ijt}$. After that we applied the logarithm to AI$_{ijt}$ gaining LAI$_{ijt}$ or we normalized by TAI$_{ijt}$ gaining NAI$_{ijt}$. What happens, if we change our focus? We could consider the share of a field *j* in a certain year *t* in relation to the average share or AI$_{it}$. We than would calculate the AI for each country gaining AI$_{itj}$. These values would significantly differ from AI$_{ijt}$ but after normalization we would gain NAI$_{itj}$ = NAI$_{ijt}$. The normalized AI is always the same, whether the AI values where calculated by year, field or country or in any other order.

But this is not the only advantage of the NAI. As told before, applying the LAI we had a loss of data of 37%. Calculating the NAI for our data set we get 6,480 observations. Indeed, we get for every AI value a corresponding TAI value, but TAI is equal zero in 1,395 cases. Overall, we have a loss of data of 21%, that is 16 percent points better than the LAI. Astonishing is the distribution of the NAI values. The NAI has like AI and TAI a range of $[0, +\infty)$ and so it follows, that it is log-normal distributed as we can see in Figure 9, that shows a normal distribution by applying the NAI values to a log scaled x-axis. But focussing on the range $[0{,}2]$ as we see in Figure 10, where we used a non-transformed x-axis, we see a nearly normal distribution of the NAI values. We claim, that for a raising number of observations, the distribution of NAI values tend to a normal distribution.

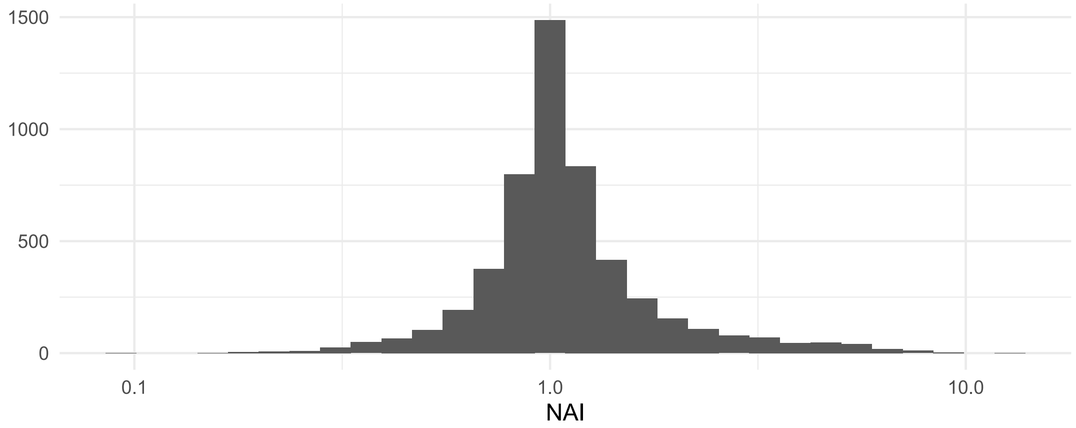

Figure 9: Normalized Activity Index for granted patents in 2011-2015 (zero excluded; logarithmic scale on x-axis).

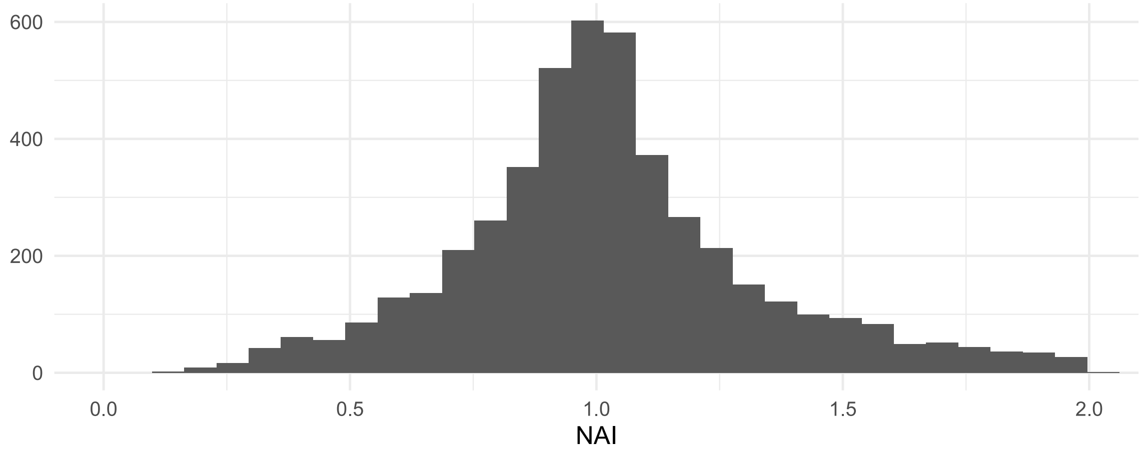

Figure 10: Normalized Activity Index for granted patents in 2011-2015 (values above 2 excluded; normal scale on x-axis).

But let us turn to our example. Figure 11 shows the NAI values of Germany and the UK for the fields FC and SC (lower two plots) in comparison to the z-standardized LAI values of the two countries and two fields (upper two plots). Besides a shift of 1 in direction of the y-axis and strongly from each other differing standard deviations, the plots seem to be very similar in their basic structure. Indeed, the RMSE of the over years z-standardized LAI values and the shifted NAI values, i.e. NAI – 1, for all observations for which both values exist is 0.77. We claim again that the RMSE tends to zero if the number of observations raises.

As a result, the NAI approximates the over years z-standardized LAI, so the interpretation of the NAI results in the same statements as of the over years z-standardized LAI. We can derive from the NAI in which years a field in a country performs over or under its own mean performance, where performance is measured as a higher or lower share of that field in its country compared to the average share. Because the interpretation of Figure 11 would be the same as of Figure 8, we skip it.

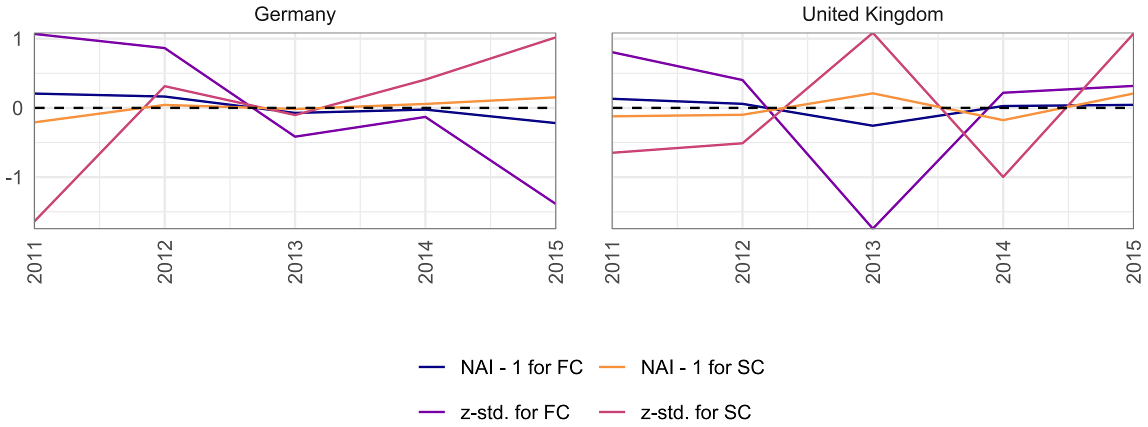

Figure 11: Granted patents as over years z-standardized LAI and as NAI (shifted down by 1 to scale to z-std.).

Till to this point, the argumentation has perhaps taken a wrong direction. We introduced a new index, the NAI, and proclaimed, that it approximates the z-standardisation of the well-known LAI executed for every country/field combination. We reproduced an already existing index. But the NAI is an original index and it is superior to the over years z-standardized LAI in some points. First, as we have seen, we have less data loss by transformation if we use the NAI instead of the over years z-standardized LAI. Second, after applying the logarithm to the AI, the index LAI losses the correspondence to the absolute data. The AI is the relation of two shares and an AI coefficient tells us, if the country specific share is upper or lower the average. The LAI does not have a correspondence like this. The NAI also has no direct translation to absolute data as the AI, but it has more interpretative content than the LAI. Remember, that $NAI(v_{ijt}) = \frac{AI_{ijt}}{TAI_{ijt}} = \frac{v_{ijt}}{\overline{v_{ijt}}} \cdot C_{ijt}$ with $C_{ijt} = C(v_{ijt}) = \frac{\sum_{i\in I, j\in J} v_{ijt}}{\sum_{i\in I, j\in J} \overline{v_{ijt}}} \cdot \frac{\sum_{i\in I} \overline{v_{ijt}}}{\sum_{i\in I} v_{ijt}} \cdot \frac{\sum_{j\in J} \overline{v_{ijt}}}{\sum_{j\in J} v_{ijt}}$. This means that the NAI is the absolute value of an observation divided through its mean over the years multiplicated by the coefficient C_{ijt}. The NAI values correspondence to the absolute values in the way that they are a special normalization of them – differently from the over years z-standardized LAI, that was a normalization of the LAI.

Conclusion

By the NAI we get a well-motivated new index, that is a natural expansion of the AI. It helps to identify performance over and under a country/fields specific mean. It approximates the over years standardized LAI and therefore has nearly a normal distribution. Thereby it omits some of the disadvantages of the z-standardized LAI – it has lower loss of data, is easier to calculate and easier to interpret. We advise the reader, to dismiss the triple for interpreting data of AI, LAI and over years z-standardized LAI and instead use the new triple AI, LAI and NAI, as the AI is a size-adjusted index to compare the field activity of countries; LAI is the normal distributed transformation of the AI that is useful for further analyse techniques as regression; the NAI is the normalized AI and while LAI is useful for analysing data on the macro level, the NAI is strongest in analysing certain country/field combinations to highlight characteristic points in longitudinal data.

For the future we claimed several statements that should be proven. The NAI approximates the over years z-standardized LAI and therefore is approximately normal distributed, and the TAI and the mean of the AI tend to the same value. These statements can be proven analytically or by a Monte-Carlo experiment. Further, there are more interpretation levels than we presented. As mentioned before, the NAI cannot only be presented for a certain country/field combination, but for example also for a certain year/field combination, because the NAI is symmetrical in its indices. And last but not least, if the division of the AI by the TAI corresponds to the normalization by the expected value, is there a technique or operation corresponding to the normalization by the standard deviation, too, to draw the NAI even nearer to a distribution with expected value of 1 and standard deviation of 1?

References

Balassa, B. (1965). Trade liberalization and "revealed" comparative advantage. *The Manchester School of Economic and Social Studies* 32, 99–123.

Grupp, H. (1994): The measurement of technical performance of innovations by technometrics and its impact on established technology indicators. *Research Policy* 23, S. 175-193.

Narin, F., Carpenter, M.P. & Woolf, P. (1987). Technological assessments based on patents and patent citations. In Grupp, H. (Ed.), *Problems of measuring technological change*, Cologne, 107-119.

Soete, L.G. & Wyatt, S.M.E. (1983). The use of foreign patenting as an internationally comparable science and technology output indicator. *Scientometrics* 5, 31-54.