INSTEPP INTERNATIONAL SCIENCE & TECHNOLOGY PRACTICE & POLICY

WORKING PAPER

OCTOBER 2016

Graphical analysis of agricultural research spillover potential

by

Maxwell Mkondiwa

Department of Applied Economics, University of Minnesota-Twin Cities, 1994 Buford Ave, Falcon Heights, St-Paul, MN 55108, Minnesota, USA.

Correspondence: mkond001@umn.edu

The Version of Record of this manuscript has been published and is available in Economics of Innovation and New Technology 2016, http://www.tandfonline.com/doi/full/10.1080/10438599.2016.1238913. DOI: 10.1080/10438599.2016.1238913

Graphical analysis of agricultural research spillover potential

This paper introduces two important extensions to the uncentered correlation metric, the commonly used metric proposed by Jaffe (1986) for analyzing research spillovers across firms or countries. First, it is shown that the Jaffe metric can be displayed graphically using the biplot, a graphical display of a two-dimensional approximation to any multidimensional matrix. Second, it is illustrated that since the data used to produce the Jaffe metric is constrained within the simplex (i.e. shares add up to one), then a theoretically superior metric satisfying the basic axioms of technological proximity measures in this sample space is the Aitchison distance measure, a metric based on log-ratios of shares. The findings of the paper using agricultural research and development spillover potential for Southern African countries show that the Jaffe metric overestimates the technological proximity across countries as compared to the proposed Aitchison measure.

Keywords: Africa; Biplot; Compositional data; Cosine similarity.

1. Introduction

Research and Development (R&D) spillovers are prevalent and important both in the innovation processes and economic development (Griliches 1992). They have thus been a major topic in the growth, productivity and industrial organization literatures for many decades (Bloom, Schankerman, and Reenen 2013). Despite the popularity, the measurement of spillovers still remains a challenge. This challenge is exacerbated by the lack of a clear definition. There are three main types of spillovers that have been studied in the literature. These are; knowledge-related spillovers, technology related spillovers and price related spillovers. These types of spillovers are not easily identifiable in practice (Byerlee and Traxler 2001) and are difficult to distinguish using existing empirical strategies. In terms of measurement, Zvi Griliches who was the first to recognize the importance of measuring spillovers in his seminal paper (Griliches 1979) subsequently summarized the future of R&D spillover analyses as follows, "progress here (in measuring spillovers) awaits appearance of better data and the development of better econometric techniques for tracing the interaction between firms (countries) and industries(regional blocs) over time in an ill-defined and changing multi-dimensional space of technological opportunities". (Griliches 1992, 44). In other words, measuring spillovers is not only an econometric or statistical problem; it is more to do with understanding the mechanism through which spillovers occur and the kind of precise data that can be collected to measure them.

In most cases, researchers circumvent the challenge of measuring spillovers by constructing proxy statistics. Such proxies essentially provide estimates of the spillover potential where potential refers to the physical, economic or biological performance of introduced technologies. According to Byerlee and Traxler (2001), actual spillovers are bounded by spillover potential but usually are smaller because of institutional and policy barriers that govern the transfer

of technologies across political boundaries or firms. The literature on research spillover employs simple to overly complex approaches to estimate R&D spillovers. Most of these measures originate from what Jaffe(1986) suggested to be the measure of "technological" similarity which is based on a vector space model of cosine similarity. This measure as used by Jaffe is essentially an un-centered correlation between different firm's shares of patents in different clusters of patents. The key logic in its use is that firms innovating in the same cluster are more likely to share knowledge and thus inter-firm research spillover are expected to be higher than firms innovating in disparate clusters. The list of papers and books in various fields employing the cosine similarity formula in measuring research spillovers is long and growing. The contributions to this literature include; Jaffe(1989a), Jaffe (1989b), Adams (1990), Branstetter (2000), Cincera (2006), Parent and LeSage (2008), Alston et al. (2010), Chyi et al. (2012) and Bloom, Schankerman, and Reenen (2013). Jaffe's insights have been developed extensively in various fields including industrial organization, manufacturing and services innovation (Kaiser 2002) and agricultural research (Byerlee and Traxler 2001; Alston et al. 2010; Ros-Freixedes and Estany 2013; Johnson et al. 2011). Coincidentally, this measure is also the mostly used distance measure in data mining and information retrieval.

The Jaffe measure has several limitations which are discussed throughout the paper. The two limitations this paper is concerned with are: (i) it is difficult to detect the pattern when there is a huge number of technological categories and firms/countries; (ii)it lacks a clear economic and statistical justification (Bloom, Schankerman, and Reenen 2013). Based on a review of most of the literature in the Jaffe tradition, there is an apparent presentational challenge that researchers face when reporting R&D spillover matrices. In most cases, a substantial number of pages are wasted to reporting these matrices, whereas in some cases the researchers aggregate at a higher level or report selected results. Even Jaffe himself limited the number of firms to only a sample of 10 to demonstrate the use of the similarity measures in his publications. In addition, he aggregated all the patent shares from 1969-79 in his paper (Jaffe 1989a). Other examples include Alston et al.(2010) and Johnson et al.(2014). Specifically, Alston et al.(2010) analyzed the inter-state spillover potential for the 49 states in United States of America using the Jaffe measure calculated using agricultural output value shares in each of the states. In their analyses they considered agricultural output shares from 1949 to 2002 and generated 1128 spillover values for each of these years. Because of space limitations, temporal average indices covering eight pages of spillover matrices were reported in the book. Johnson et al.(2014) provides a further proof of this challenge as they used the Jaffe measure to study the agricultural spillover potential in the Southern African Development Community (SADC) consisting of 15 member countries (analyzed only for 13 of the countries) with tables covering about 17 pages of the paper. This is a lot of pages wasted and deterrence to understanding details of spatial spillover potential.

The main objective of this paper is to explore two extensions of the Jaffe metric that address these two limitations. The two extensions are; graphically displaying Jaffe metric using the biplot and proposing a new technological proximity based on the Aitchison distance that

recognizes the unity-constrained sample space for datasets used in constructing the proximity measures. This paper makes two main contributions in line with these extensions. First, we show that a multidimensional matrix of Jaffe indices can be presented in a two-dimensional graphical display, the biplot. This paper is first to suggest such a theoretical equivalence. Biplot graphical analysis has been extensively used by breeders and agronomists to evaluate environmental similarity in terms of crop variety (genotype) performance using genotype by environment matrices (Yan and Kang 2003). This study shows that the Jaffe measure can be presented in a particular type of a biplot called environment preserving biplot. Therefore, economists working in this area can leverage this equivalence by visualizing their indices using biplots. Second, we propose extensions to the Jaffe measure based on this equivalence. In particular, we propose that compositional data analysis tools particularly, the Aitchison distance, should be considered in analyzing spillovers since in all cases the data used to calculate the cosine indices are compositional (have a constant sum) in nature. We place the Aitchison measure within an axiomatic analysis suggested by Bloom, Schankerman, and Reenen (2013) and demonstrate that it strictly dominates any other distance measure. This paper is again the first to establish the compositional nature of the datasets used in constructing spillovers and thus a pointer to a large body of literature that researchers working on R&D can leverage.

The remainder of the paper is organized as follows. Section 2 presents the Jaffe measure as used in the agricultural R&D spillover measurement and the subsequent Jaffe-Griliches knowledge production function. Section 3 links biplot analysis to the Jaffe measure. Section 4 presents the data and software used to demonstrate the equivalence. In section 5, the results and discussion are presented. The various spillover measures and Aitchison distance measure that this paper proposes are summarized in section 6. Finally, section 7 presents the concluding remarks.

2. Cosine Similarity in Economic Analysis of Spill overs

Jaffe(1986) proposed a measure that reflects the similarity of research focus of firms to measure closeness among firms in the technological space. The measure is simply the angular separation of the vectors- equal to the cosine of the angle between them- defining the "technological proximity" of any two firms. According to Jaffe (1989a), if spillovers are important then firms doing research in areas where much work is done by other firms should be more successful, all else equal. We use the adaptation of Jaffe's measure to international agricultural R&D as suggested by Alston et al. (2010). The Jaffe measure of technology spill over potential between country i and j is defined as an uncentered correlation between y_i and y_j

$$\omega_{ij} = \frac{\sum_{n=1}^{N} y_{in} y_{jn}}{(\sum_{n=1}^{N} {y_{in}}^2)^{0.5} \left(\sum_{n=1}^{N} {y_{jn}}^2\right)^{0.5}} = \cos\theta_{ij} \qquad \textbf{(1)}$$

where y_{in} is the value of production of output *n* as a share of the total values of agricultural output (traded volume) in country *i*, such that these shares fall between zero and one. Just as a correlation,

this measure is symmetric (Jaffe 1989a). According to Alston et al.(2010), the measure can be interpreted as follows; (i) ω_{ij} measures the degree of overlap of y_i and y_j, (ii) the numerator will be large when country *i* and *j* have very similar output mixes and (iii) the denominator normalizes the measure to be one when y_i and y_j are identical. Clearly, $\omega_{ij} = 1$ if output values-up to a factor of proportionality- are identical and will approach zero the more dissimilar the output mix is between any regions. If the output bundles of two countries are dominated by maize, for example, then the research portfolios of both countries will be similarly biased toward maize. In this case, the spillover potential would be higher than if the comparison were between a predominantly maize-producing country and say, coconut-producing country (Alston, Norton and Pardey 1998).

Several economic analyses are done incorporating this measure as the point of departure. The dominant and workhorse economic model for estimating R&D spillovers is the Jaffe-Griliches knowledge production function after pioneering development by Griliches (1979) and (Jaffe 1986). It is traditonally presented in modified Cobb-Douglas form as

$$\log Y_{it} = \beta_0 + \beta_1 \log(R_{it}) + \beta_2 \log(S_{it}) + \beta_3 \log(X_{it}) + \varepsilon_{it} \quad \textbf{(2)}$$

where i indexes unit of observation (countries in this case) and t indexes time. Y is any proxy for economically useful knowledge (this may include: patents, registered varieties, profits et.c), R is the R&D performed by each country in each technological area and time period, S is the R&D performed by other countries, X is a vector of all attributes that need to be controlled for in estimating the knowledge production function including the level of economic activity, labor e.t.c.. ε_{int} is the error term that is assumed to be identically and independently distributed.

The Jaffe's cosine similarity measure is used as a weight in constructing S_{it}

$$S_{it} = \sum_{j \neq i} \omega_{ij} R_{jt} \quad \textbf{(3)}$$

According to Jaffe (1986), S_{it} formulated in this way has significant power in explaining variations to "innovative success" across firms/countries. Alston, Norton and Pardey (1998) observed that this approach is not fruitful instead suggested that S_{it} should be calculated as, $S_{it} = \prod_{i \neq j} (R_{jt})^{\omega_{ij}}$. They argued that Jaffe's approach does not address the question of how to decide what constitutes relevant research, and in so doing, Jaffe's approach inappropriately treats all research done by other countries or firms as being equally relevant from the perspective of its spill-in. There are several specifications that have appeared in literature to incorporate several other attributes of interest to S_{it}. For instance, those more concerned with spatial externalities combine the Jaffe measure with the spatial contiguity matrix in a spatial econometric framework. Those interested in a simple model also only use the linear version of this model. In addition, due to the simultaneous nature of R&D, researchers have used equation 2 with other equations including one that has R&D perfomed as the dependent variable (e.g. Jaffe 1989b; Alene and Coulibaly 2009).

Both Bayesian and frequentist specifications have also been tried out in practice depending on context.

The underlying aspect in each of these specifications is that the Jaffe's index acts as a proxy for determining the levels of spillovers. In this study, we investigate how researchers using the measure can graphically visualize the spillover potential. We will demonstrate the different extensions that may be considered in this exploratory analysis and the suggested modifications required to this index consistent with the nature of the datasets used in practice to construct it. It suffices to mention some of the limitations of the Jaffe measure. It is an unconditional such that such issues as geographical proximity are ignored. In terms of presentation, the major disadvantage of this measure is that it is difficult for one to see the pattern overtime and across the countries/firms. In addition, it is impossible to use the measure if there are huge disparities in the commodities the countries or firms produce or if there is huge heterogeneity among the countries or firms. In this paper we do not try to address many of these limitations of the Jaffe measure but rather show how those who use it may gain more through the use of graphics. The biplot analysis provides a comprehensive and theoretically equivalent solution to the problem of presentation.

3. Biplot-Jaffe Measure Equivalence

There are several advantages to establishing the equivalence between Biplot and Jaffe measure. Firstly, the biplot allows us to extend the use of Jaffe measure to considering more than two countries or firms at a single point in time (year) or over a long period of time. With the biplot, we can also assess the stability of the spillover potential and identify a research policy integration zone for a particular commodity. The biplot was introduced by Gabriel (1971) as a visual display of a rank-two approximation matrix of any two-way table through plotting its two component matrices while also showing the inner product property as in equation 4 below. It thus allows visualization from all perspectives. Assume we have N commodities/output values and P countries. If we consider the share of each output value to the total value and denote it as Y, we will have a measure of relative importance of each output value to that country and thus a measure of agricultural technology spillover potential. It can be shown using trigonometric identities that each element of $\boldsymbol{Y}$, $\boldsymbol{Y_{np}}$ is given by the product of a row of a row matrix($\boldsymbol{N}$), column of a column matrix ($\boldsymbol{P}$) and the cosine of the angle separating the two vectors. Thus;

$$Y_{np} = |N_n|\left|P_p\right| \cos\theta_{np} \qquad \textbf{(4)}$$

where $|\boldsymbol{N_n}|$ is the vector length for row n, $\left|\boldsymbol{P_p}\right|$ is the vector length for column p, θ_{np} is the angle between the vectors of row n and column p . This equation is referred to as the inner-product property of the biplot. It is this property that allows us to estimate the elements of $\boldsymbol{Y}$, visualize the patterns in the matrix $\boldsymbol{Y}$, and compare any two columns (countries) relative to rows (output categories). We need to develop a general approach for identifying the row and column matrices that will demonstrate the inner product property. The process of decomposing a two-way matrix

Y into two component matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ is called singular value decomposition (SVD) as developed by Eckart and Young (1939) which is essentially the reverse process of matrix multiplication (Yan and Kang 2003). The biplot uses a least-squares approximation and relies analytically on SVD (Gower, Lubbe, and Roux 2011; Greenacre 2012) as below;

$$Y: N \times P = U\Sigma V' \tag{5}$$

where, U is an $\boldsymbol{N} \times \boldsymbol{Z}$ orthogonal matrix with columns known as the left singular vectors of $\boldsymbol{Y}$, the matrix V' is an $\boldsymbol{Z} \times \boldsymbol{P}$ orthogonal matrix with columns known as the right singular vectors of $\boldsymbol{Y}$, and matrix $\sum$ is a $\boldsymbol{Z} \times \boldsymbol{Z}$ diagonal matrix containing Z singular values with $\boldsymbol{Z} \leq min\ (\boldsymbol{N}, \boldsymbol{P})$. In summation notation, SVD decomposes $\boldsymbol{Y}$ into $\boldsymbol{Z}$ principal components, each containing a set of row or output share vectors (ξ_n) and column or country vectors (η_p) and a singular value (λ). Thus;

$$Y_{np} = \sum_{z=1}^{Z} \lambda_z \xi_{nz}\, \eta_{zp} \tag{6}$$

where $n = 1 \ldots N$ (output categories), $p = 1 \ldots P$ (Countries), $\boldsymbol{Y_{np}}$ is the output share for output category n in country p, z is the rank of a principal component, $z = 1 \ldots Z$, λ_z is the singular value of the zth principal component, with $\lambda_1 > \lambda_2 > \cdots > \lambda_Z$. The square of λ_z is the sum of squares explained by the zth principal component. Only the first two principal components are mostly used in biplot analysis. Additionally, η_{zp} is the eigenvector or singular vector of country p for the zth principal component, and ξ_{nz} is the eigenvector or singular vector of output n for the zth principal component. All Z principal components are orthogonal and orthonormal to one another for both the rows (output shares) and the columns (countries) thus satisfying the following restrictions;

$$\begin{aligned} \sum_{n=1}^{N} \xi_{nz}\, \xi_{nz'} &= 0 \\ \sum_{p=1}^{P} \eta_{zp} \eta_{zp'} &= 0 \\ \sum_{n=1}^{N} \xi_{nz} &= 1 \\ \sum_{p=1}^{P} \eta_{zp} &= 1 \end{aligned} \tag{7}$$

In addition to these restrictions, Gabriel (1971) showed that the singular value, singular column and singular row should be chosen to satisfy basic rules of eigen values and eigenvectors. The singular values must be partitioned into the row (output categories) and column (country)

scores before a biplot can be constructed to approximate the two-way data (Yan and Tinker 2006). This singular partitioning is given by;

$$Y_{np} = \sum_{z=1}^{Z} (\xi_{nz}\, \lambda^{f}{}_{z})\ (\lambda_{z}{}^{1-f}\, \eta_{zp}\ ) \qquad \textbf{(8)}$$

where f is the partitioning factor and can be anything between 0 and 1 resulting in unlimited number of ways of singular value partitioning. The main ones are; column-metric preserving (f=0), row-metric preserving $(\mathrm{f} = 1)$ and symmetrical partitioning $(\mathrm{f} = 0.5)$. For the purposes of studying the relations among columns (countries) say $i, j \epsilon\, P$, the column metric preserving is the most appropriate in which case $f = 0$ as it allows one to visualize similarity or dissimilarity among countries. Since the Pearson correlation between two columns (countries) is estimated by:

$$\omega^{*}{}_{ij} = \frac{\sum_{n=1}^{N}(y_{ni} - \bar{\mathrm{y}}_i)(y_{nj} - \bar{\mathrm{y}}_\mathrm{j})}{(\sum_{n=1}^{N}(y_{ni} - \bar{\mathrm{y}}_i)^2)^{0.5}\left(\sum_{n=1}^{N}(y_{nj} - \bar{\mathrm{y}}_\mathrm{j})^2\right)^{0.5}} \qquad \textbf{(9)}$$

Kroonenberg(1995) showed that when two way data are column (country) centered i.e. when $\overline{y}_j = \overline{y}_i = 0$, then the cosine of θ_{ij}, the angle between two columns (countries) is equal to their correlation, that is:

$$\omega^{*}{}_{ij} = \cos\theta_{ij}\ = \frac{\sum_{n=1}^{N} y_{ni}\, y_{nj}}{(\sum_{n=1}^{N} y_{ni}{}^{2})^{0.5}\left(\sum_{n=1}^{N} y_{nj}{}^{2}\right)^{0.5}} = \frac{\sum_{n=1}^{N} y_{ni} y_{nj}}{|P_i||P_j|} \qquad \textbf{(10)}$$

$$\sum_{i=1}^{N} y_{ni} y_{nj} = |P_i||P_j| \cos\theta_{ij} \qquad \textbf{(11)}$$

It is apparent that equation (11) has the same principle as equation (4) with the difference that in equation 11 the approximated correlation is between columns while in equation (4), the approximated correlation is between a particular row and column. This establishes the relationship between Pearson correlation and cosine similarity within the context of biplot analysis. Furthermore, equations (1) and (10) are the same. These equivalences are important in showing that the Jaffe measure and environment or country preserving biplot are equivalent. Thus, the Jaffe measure can be conceived as a particular distance measure in biplot analysis in which the inner products are over the columns.

Nevertheless, we need to add a disclaimer to this comparison. The cosine of the angle between the vectors of two columns, i, j is determined solely by the values in matrix $\boldsymbol{V}$ and has nothing to do with the values in matrix $\boldsymbol{U}$, whereas the correlation coefficient calculated based on matrix $\boldsymbol{Y}$ is dependent on both $\boldsymbol{U}$ and $\boldsymbol{V}$. Consequently, angles between columns of $\boldsymbol{U}$ in the biplot should be somewhat related to the correlation coefficients among the columns in $\boldsymbol{Y}$ but no strict

correspondence should be expected. Closer or near perfect correspondence is expected if matrix ***U*** has many rows and the rows are randomly scattered on the biplot (Yan and Kang 2003). In the context of column environment or country comparisons, the cosine similarity as developed by Jaffe is exactly the same as cosine similarity in environment preserving biplot. Most importantly they take values from 0 to 1, have the same interpretations and approximate the Pearson correlation (Yan 2014).

The column or country preserving biplot has the following basic properties, the first which is essentially similar to Jaffe measure; (i) the cosine of the angle between any two columns (countries) approximates their correlation, with equality if the fit is perfect; (ii) the lengths of the country vectors are approximately proportional to their standard deviations, with exact proportionality if the fit is perfect; and (iii) the inner product between two countries approximates their covariance, with equality if the fit is perfect (Yan and Tinker 2006; Greenacre 2010). The other important assumption that is implicitly made in using the Jaffe measure or country-centered biplot is that countries are homogenous in the way they can utilize spill-ins. This is not in any way true as we know that there is variation across countries in important factors that may affect its technology use including population, level of development and other social-cultural factors. It is an impossible task to incorporate all these in the approximation of the spillover potential measure. We can incorporate this heterogeneity by using a country standard deviation weighted/standardized measure when constructing the biplot. This standardization also helps in the posititioning of the different vectors and points in the biplot for easy visualization.

4. Data and Software

The paper uses the Food and Agriculture Organization of the United Nations (FAOSTAT) production estimates data (http://faostat3.fao.org/home/E). The production data were collected for all member countries of the Southern Africa Development Community (SADC). The countries included; Angola, Botswana, Democratic Republic of Congo, Lesotho, Madagascar, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania, Zambia, Zimbabwe. The data consisted of production values for 141 commodities from 1961 to 2011. In the analyses we aggregated the commodities into 13 groups for expository purposes. The commodity groups are; cereals, roots and tubers, sugarcane, pulses, nuts, oil crops, vegetables, fruits, fibres, spices, stimulants, tobacco, and livestock. In terms of software, there are several commercial and non-commercial software that can implement the various versions of the biplot that have been proposed in this paper. All the biplots in the paper were done in R (R Core Team 2014), a free statistical computing environment. Specifically, the biplot and canonical variate analysis (CVA) biplots were drawn using the UBbipl package (Roux and Lubbe 2013). The robust compositional biplots were drawn using Robcompositions package (Templ, Hron, and Filzmoser 2011). The pseudo R code in the appendix provides the key functions for the interested reader. Other software especially known to social scientists and economists that can also be used for the graphical approach includes, Matlab and Stata.

5. Results and Discussion

5.1 Spatial agricultural R&D spill over potential

In this expository analysis we have considered the Jaffe measure for the latest data (2011) as we assume that the latest output mix clearly reflects a resultant effect of past R&D effects in the performing countries as well as R&D spillovers. The Table 1 shows the Jaffe's similarity measures across the different countries calculated using 141 commodity shares. It is apparent that Seychelles and Mauritius are the least similar to all other countries with a similarity index to the average SADC region of 0.36 and 0.49 respectively. In terms of country to country comparisons; Mauritius has the least similarity (0.03) to Angola and Botswana implying that the potential for spillovers is minimal. On the other hand, Namibia and Botswana have the highest similarity index.

Table 2 shows the Jaffe measure calculated using aggregated commodity groups. It is apparent that the cosine similarity figures are higher in Table 2 than Table 1. This is because of an important property of correlation based measures. The correlation of aggregates is always higher than the correlation of individual elements. This is the case because the numerator (covariance) is the same for aggregates and individual elements yet the denominator (square of the variances) is smaller for aggregates (since the variability is lower) than for individual elements. The challenge of using highly disaggregated shares is that there is a high chance of having zero shares which again distorts the index. The use of Jaffe measure can therefore be ambiguous because the optimal way of constructing the classes is subjective. In the biplot comparisons we use the aggregated shares for expository purposes but disaggregated shares would also be used depending on one's preference.

Table 1: Agricultural spill over potential (using 141 commodities) across countries in SADC region, 2011

	ANG	BT	D.R.C	LES	MD	MW	MAU	MOZ	NAM	SEY	RSA	SWA	TZ	ZA	ZIM	All
ANG	1.00	0.16	0.87	0.21	0.34	0.59	0.03	0.88	0.16	0.17	0.18	0.07	0.60	0.37	0.24	0.61
BT		1.00	0.07	0.83	0.33	0.11	0.03	0.06	0.91	0.06	0.62	0.24	0.45	0.39	0.63	0.63
D.R.C			1.00	0.09	0.34	0.56	0.04	0.91	0.07	0.05	0.08	0.06	0.41	0.38	0.17	0.53
LES				1.00	0.32	0.47	0.04	0.15	0.78	0.12	0.69	0.22	0.56	0.51	0.67	0.68
MD					1.00	0.26	0.09	0.36	0.32	0.09	0.28	0.17	0.63	0.28	0.29	0.53
MW						1.00	0.11	0.67	0.15	0.11	0.38	0.16	0.61	0.70	0.53	0.63
MAU							1.00	0.12	0.03	0.33	0.40	0.88	0.08	0.25	0.30	0.49
MOZ								1.00	0.09	0.14	0.21	0.13	0.51	0.53	0.32	0.63
NAM									1.00	0.08	0.64	0.25	0.44	0.38	0.59	0.62
SEY										1.00	0.48	0.04	0.23	0.18	0.23	0.36
RSA											1.00	0.35	0.51	0.64	0.75	0.76
SWA												1.00	0.19	0.35	0.41	0.57
TZ													1.00	0.59	0.61	0.74
ZA														1.00	0.82	0.75
ZIM															1.00	0.78
All																1.00

Notes: Country codes; ANG=Angola, BT=Botswana, Les=Lesotho, Md=Madagascar, Mw=Malawi, Mau=Mauritius, Moz=Mozambique, Nam=Namibia, Sey=Seychelles, RSA=South Africa, SWA=Swaziland, TZ=Tanzania, ZA=Zambia, ZIM=Zimbabwe, All= All countries.

Table 2: Agricultural spill over potential (using 13 commodity groups) across countries in SADC region, 2011

	ANG	BT	D.R.C	LES	MD	MW	MAU	MOZ	NAM	SEY	RSA	SWA	TZ	ZA	ZIM	All
ANG	1.00	0.32	0.99	0.50	0.61	0.88	0.21	0.90	0.48	0.39	0.45	0.27	0.72	0.45	0.36	0.67
BT		1.00	0.29	0.97	0.56	0.30	0.51	0.35	0.98	0.95	0.92	0.38	0.57	0.66	0.87	0.85
D.R.C			1.00	0.47	0.62	0.90	0.19	0.93	0.45	0.35	0.42	0.25	0.73	0.49	0.35	0.66
LES				1.00	0.72	0.51	0.50	0.55	0.99	0.92	0.94	0.40	0.72	0.77	0.91	0.93
MD					1.00	0.77	0.35	0.71	0.65	0.58	0.77	0.37	0.89	0.91	0.75	0.83
MW						1.00	0.21	0.97	0.45	0.31	0.45	0.27	0.82	0.70	0.50	0.70
MAU							1.00	0.25	0.51	0.51	0.55	0.98	0.36	0.47	0.57	0.63
MOZ								1.00	0.50	0.34	0.45	0.28	0.78	0.66	0.48	0.71
NAM									1.00	0.94	0.94	0.40	0.67	0.71	0.88	0.91
SEY										1.00	0.96	0.40	0.67	0.64	0.86	0.86
RSA											1.00	0.48	0.79	0.80	0.93	0.93
SWA												1.00	0.39	0.45	0.49	0.59
TZ													1.00	0.86	0.76	0.86
ZA														1.00	0.89	0.85
ZIM															1.00	0.90
All																1.00

Notes: Country codes; ANG=Angola, BT=Botswana, Les=Lesotho, Md=Madagascar, Mw=Malawi, Mau=Mauritius, Moz=Mozambique, Nam=Namibia, Sey=Seychelles, RSA=South Africa, SWA=Swaziland, TZ=Tanzania, ZA=Zambia, ZIM=Zimbabwe, All= All countries.

An equivalent presentation of the Table 2 is shown by the biplot in Figure 1. In the figure, the lengths of the country vectors indicate how well the countries are represented by the graph with a perfect fit all vectors have equal lengths. If the country vector is close to the origin, it means that the country has little variability or does not fit well in two dimensions. The inner product between two countries (and the cosine of the angle between them) approximates their correlation with equality if the fit is perfect (Kroonenberg 1995). The smaller the angle, the greater the positive correlation between the two countries. It is apparent in the Figure 1 that the estimation for Mauritius, Swaziland and Madagascar is almost imperfect with respect to the lengths of the other countries. With a matrix of Jaffe measures, it is impossible to discern the quality of estimations. In the case of the biplot in Figure 1; the two principal components explained about 63% of the variation. These characteristics of a biplot provide a better assessment than the matrix of Jaffe measure.

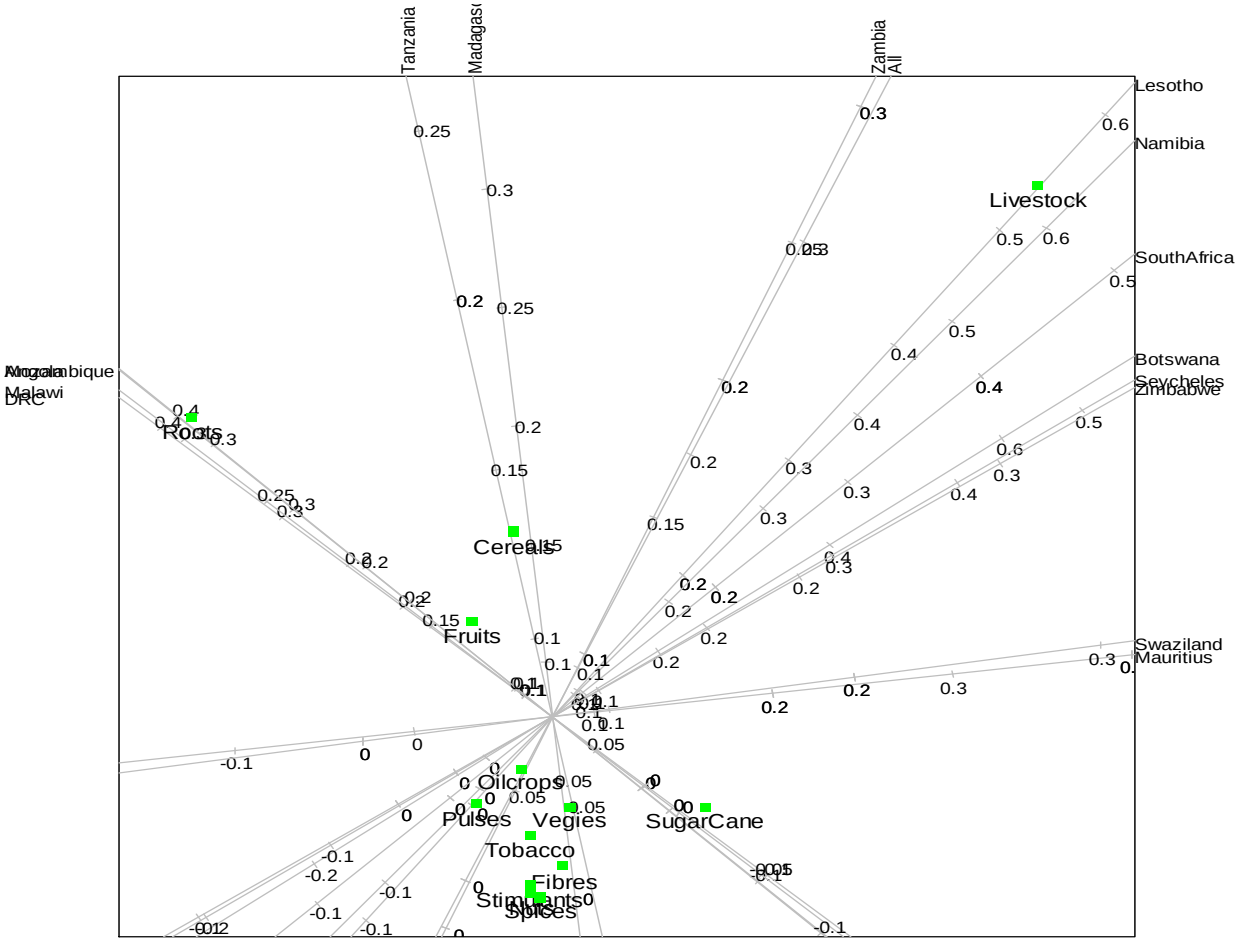

Figure 1: Biplot of output value shares and Southern African countries-2011

In order to understand the similarity among all the countries studied, we can simply check the different axis sides of the biplot. The countries on the same axis side have the smallest angle (i.e. high similarity) between themselves as compared to countries on the other axis side. For instance; Malawi, Mozambique, Democratic Republic of Congo and Angola are all on the left axis implying that they are agro-ecologically similar. Similarly, Madagascar, Tanzania and Zambia form their own group on the top. In the right axis, countries with large spillover potential include; Lesotho, Namibia, Botswana, South Africa, Swaziland, Zimbabwe, Swaziland, Seychelles and Mauritius. For this group, one striking factor is also the geographical proximity among them implying that neighborhood spatial effects may be important and can be further investigated. In general, these groupings can be verified in the Jaffe measure tables. Within each group, the countries have similarity indices above 0.5 and it is below this threshold against countries in the other groups. These groups can then be used in investigating further the potential for agricultural R&D policy integration zones. With matrices of Jaffe measures, it would have been difficult to discover these groups of countries.

In addition to the advantages already mentioned, a biplot also helps in visualizing the spillover potential for specific types of crops. For instance, though we have bundled up all cereals in the Figure 1, we may have a disaggregated plot that shows the similarity of the countries with respect to disaggregated values of cereals and other commodity groups (Figure 2). It is noticeable that this change has slightly affected the nature of the plot just as aggregating commodity groups affected the Jaffe measures in Table 1 and Table 2. This may be attributed to sub-composition incoherence though we can argue that the changes have not affected the groups of countries evident in Figure 1.

Though the purpose of this paper was to make a methodological contribution, the findings from the biplot analysis in Figure 1 and Figure 2 are consistent with empirical findings in the literature regarding the spillover potential among these countries. The interpretation of relationship between country vectors and commodity points is that, country vectors with a projection closer to a particular commodity are more similar in that commodity. And commodities closer to the center of the biplot are common among all the countries. Firstly, consider livestock which clearly discriminates over countries. It is apparent that countries with high R&D spillover potential in livestock include; Lesotho, Namibia, South Africa, Seychelles, Zimbabwe, Botswana and Swaziland. Johnson et al. (2014) reported that livestock R&D spillover potential is evident for Botswana, South Africa, Namibia and Swaziland. Sugarcane seems to the key commodity for Mauritius. Countries similar in roots and tubers include Mozambique, Malawi, Angola and Democratic Republic of Congo. These R&D policy relevant results can be discerned in a single biplot and one can easily change different aspects of the biplot to visualize the patterns of interest.

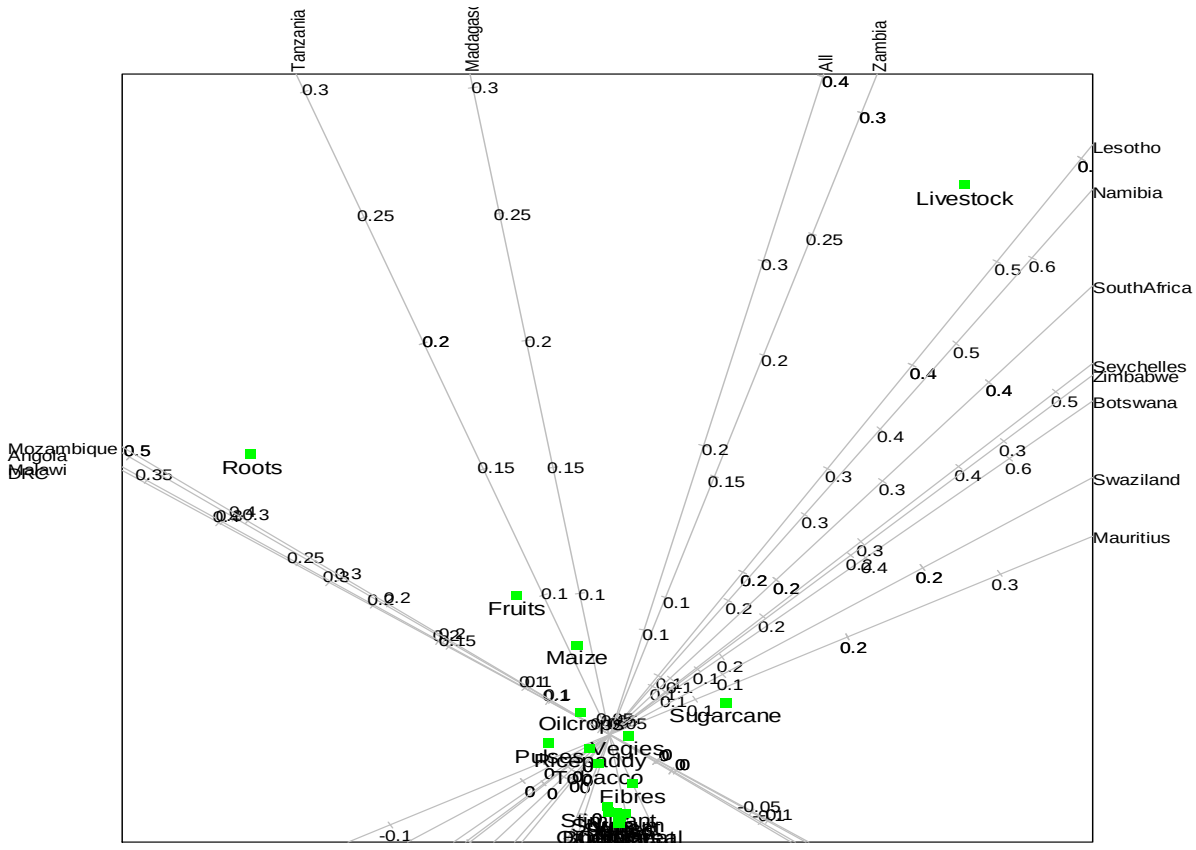

Figure 2: Biplot of output value shares (with disaggregated cereals)- 2011

5.2 Spatio-temporal patterns of spillover potential

Parent and LeSage (2008) observed that the structural assumption of symmetry in technological distance between two regions is untrue and argued that model based indices that are able to account for differences in economic activity and other variables of interest are the most appropriate in modelling spill overs. A graphical representation of model based indices can be accomplished using a multivariate extension to the biplot called the Canonical Variate Analysis (CVA) biplot. CVA is a useful method for describing and assessing the differences between means of groups or classes using the Mahalanobis Distance (formula shown in Table 4) to define inter-group distance (Gower, Roux, and Gardner-Lubbe 2014). CVA is simply a two stage rotation. The first stage involves an eigen analysis of the original variables. The second stage involves an eigen analysis of the variation between the group means for the variables from the first stage principal component analysis (Campbell and Atchley 1981). The reader is referred to Gower, Lubbe, and Roux (2011) for details.

An important aspect to characterizing spill overs is the temporal evolution of the spill over potential. The spill over matrix with temporally aggregated data or multiple spill over matrices can

be created to understand this temporal aspect. Evidently, it would be difficult to discern the patterns of spill overs in such matrices. The CVA biplot instead can be used to discern these temporal patterns of spill over potential as in Figure 3.

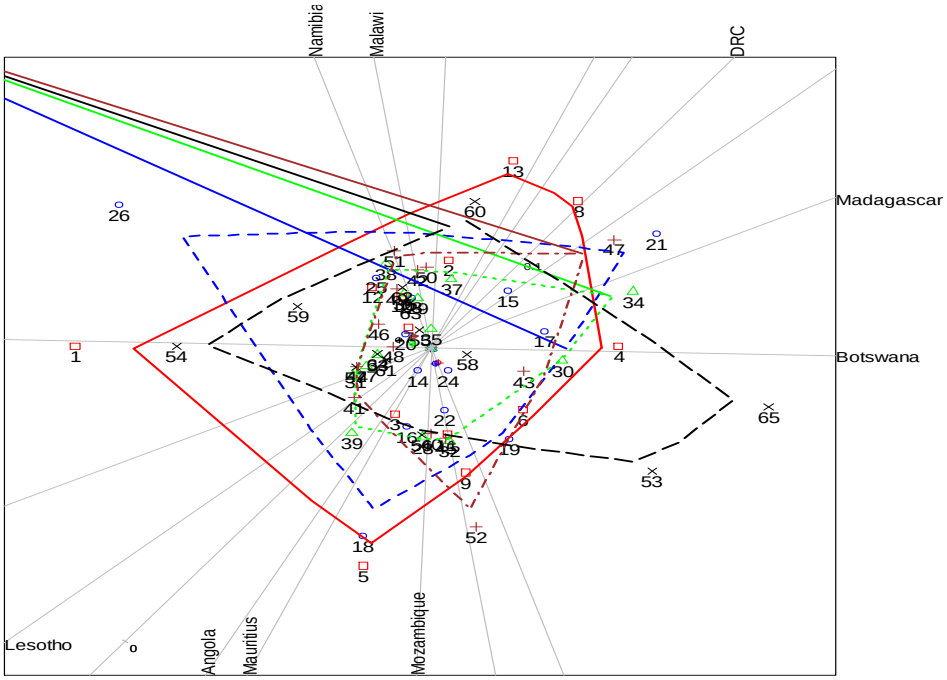

Figure 3: CVA biplot for output shares in nine countries, 1961-2011.

Notes: The numbers represent the crop categories from 1 to 13 for 1961-70 (**red**), 14-26 for 1971-80 (**blue**), 27-39 for 1981-90 (**green**), 40-52 for 1991-2000 **(brown**) and 53-65 for 2001-2011(**black**). Each category of crops corresponds to a corresponding list of commodities (1=Cereals, 2=Roots and tubers, 3=Sugarcane, 4=Pulses, 5=Nuts, 6=Oil crops, 7=Vegetables, 8=Fruits, 9=Fibres, 10=Spices, 11=Stimulants, 12=Tobacco, 13= Livestock).

Figure 3 is however not comparable to Figures 1 and 2 because only 2011 data were used for Figures 1 and 2. The focus in Figure 3 should not be on the country vectors but the groups of 95% alpha bags of output shares shown for each decade. It's apparent that during the 1961-1970 period, the countries were quite specialized in the commodities that were produced. This was the period when most of these countries were gaining independence from colonial governments and the commodities grown in each country were dependent on the preferences of the colonial governments. There is evidence of convergence in 1970s to 2000s with the alpha bags shrinking which means that the composition of commodities was getting similar. In 2000-2011, it is apparent that there is a slight divergence again which can be further investigated using formal econometric tools. The shapes of the alpha bags are different implying that the changes are heterogeneous across years and countries.

5.3 Compositional analysis of spatial R&D spillover potential

The cosine similarity has desirable properties in most spillover analyses as will be summarized in section 6. There is however a subtle property that is often ignored when this measure is used as a proxy for spillover potential. It is that, output or patent shares are compositional data. Compositional datasets have constant sums and only positive values e.g. row proportions of contingency tables and shares of patents. The standard data analysis techniques when used on compositional data usually tend to be misleading i.e. standard covariances and correlations are wrong (Kohler and Luniak 2005) .

McNamee (2013) also argued that the fundamental problem with Jaffe measure is that it assumes that groupings at an aggregation level chosen are unrelated and independent of one another. This may not usually be the case. The concept of a subcomposition, and the requirement that any form of analysis should have what is called subcompositional coherence (Aitchison and Greenacre 2002) is violated by most conventional technological proximity measures like Jaffe metric, Euclidean distance and Mahalanobis distance. Subcompositional coherence states that inferences about subcompositions should be consistent, regardless of whether the inference is based on the subcomposition or the full composition (Aitchison 1992). This property is illustrated using a simple example below. In addition, the results should be scale invariant, that is, the information in a composition should not depend on the particular units in which the composition is expressed (Hron and Filzmoser 2015). The Jaffe measure satisfies scale invariance but fails on subcompositional coherence (Aitchison 1992).

With compositional data, the concern should be on relative rather than absolute magnitudes. Thus, we should be concerned with ratios of the R&D/ output shares rather than the absolute shares themselves. One common transformation relevant for such data is the log-transformation. This is important because the variance of samples i and j, var $(^{y_i}/_{y_j})$ is not precisely related to var $(^{y_j}/_{y_i})$ such that one would need a large number of such descriptive summaries. With logarithms of ratios, a simple property that, $\ln(^1/_a) = -\ln a$ is utilized, so that $E\left\{\ln\left(^{y_j}/_{y_i}\right)\right\} = -E\{\ln(^{y_i}/_{y_j})\}$ and $\text{var}\left\{\ln\left(^{y_j}/_{y_i}\right)\right\} = \text{var}\{\ln(^{y_i}/_{y_j})\}$ (Aitchison 1990).

According to Aitchison (1990), the first requirement for any analysis of compositional variability must be the provision of a simple and effective way of summarizing the relative variability of components within a compositional dataset. The plotting should also be within the simplex and follow what is called Atchision Geometry since the simplex is only a part of the Euclidian Space that is normally used in standard statistical analyses. The other advantage to using compositional data analysis methods is, unlike Jaffe measure, the statistical distribution of compositional measures are well proven. With Jaffe measure, it is difficult to make any inferential statements. For instance, there is no any literature to our knowledge that explains how one may do formal hypothesis testing with the Jaffe measure. Though, we can speculate that bootsrapping can be considered for this purpose, the other limitations make it worthwhile to consider compositional

statistical methods. For the sake of brevity, the reader is refered to Aitchison and Greenacre (2002) for details on the derivations of the compositional biplot.

The distance measure corresponding to the Jaffe measure is the Aitchison distance,

$$d_{ij} = \sqrt{\frac{1}{2N}\sum_{n=1}^{N}\sum_{n'=1}^{N}(\ln\frac{y_{in}}{y_{in'}} - \ln\frac{y_{jn}}{y_{jn'}})^2}$$

where, $y_{in}$and $y_{in'}$ are elements of the vector of technology categories representing the share of a particular technology in a country $i$and N is the number of technology categories as before. The technology proximity measures are then constructed following Aldieri and Cincera (2009) for geographic proximity measures as the negative exponential function of d_{ij} so if the technological Aitchison distance is zero, the technological Aitchison proximity is 1, i.e. the maximum possible value:

$$w_{ij}^{A} = \frac{1}{e^{d_{ij}}}$$

This transformation allows the comparison between the Jaffe measure and the proposed Aitchison proximity measure. Precisely, $w_{ij}^{A} = 1$ whenever $i = j$ just as the jaffe measure and $w_{ij}^{A} \approx 0$ whenever the distance measure approaches infinity. Therefore w_{ij}^{A} can be used instead of w_{ij} in the equation (3) and the analysis would proceed as traditionally done with the Jaffe metric. In order to expound on the theoretical superiority of the Aitchison measure, we demonstrate using simple examples the consequences of using this measure instead of the Jaffe measure. A more complete comparison would require an analytical and mathematical analysis of the two measures or a Monte Carlo experiment to analyze the numerical properties of the two measures. This has never been shown in the literature to the knowledge of the author and is beyond the objectives of this paper but is an object for future research. It is however shown using the data for this paper that the Jaffe measure overestimates the technological proximity measures and whenever there are nonessential classes, the measure changes.

The key distinguishing factor for the superiority for the Aitchison proximity measure is the property of subcompositional coherence. We can illustrate this property with a simple example. Consider, two hypothetical firms or countries (A and B) patenting or producing some technologies in three technological classes. Let the shares for A in each of the classes be $A = (\frac{2}{4}, \frac{1}{4}, \frac{1}{4})$ and for B be $B = (\frac{1}{4}, \frac{2}{4}, \frac{1}{4})$. Now assume that the researcher analyzing technological proximity measures has access to data only for the two technological classes and reconstitutes the shares to add to one as before as follows: $A' = (\frac{2}{3}, \frac{1}{3})$ and $B' = (\frac{1}{3}, \frac{2}{3})$. Since the share of patenting in the third technological class was the same one would expect that the technological proximity measure

between A and B would not change when we consider A' and B'. When we calculate the Jaffe metric for A and B, it is found that $w_{AB} = 0.833$ while for A' and B' it is $w_{A'B'} = 0.8$. When we use the Aitchison technological proximity measure, using the two subcomposions is coherent with using the full composition and we get the same value of $w_{AB}^{A} = w_{A'B'}^{A} = 0.375$. The reason for the coherence in the Aitchison proximity measure is that the ratios between the two classes are maintained i.e. $\frac{2}{4} \div \frac{1}{4} = 2 = \frac{2}{3} \div \frac{1}{3}$ when the third class is removed. Similar examples are shown by Aitchison (1992). According to Simon and Sick (2016), the cosine/Jaffe measure increases with increasing number of irrelevant patent classes even though there is no change in the overlap. It is apparent from this example that the Aitchison distance measure is lower than the Jaffe measure. This is confirmed using the agricultural R&D spillover for Southern African countries as shown in the figure 4.

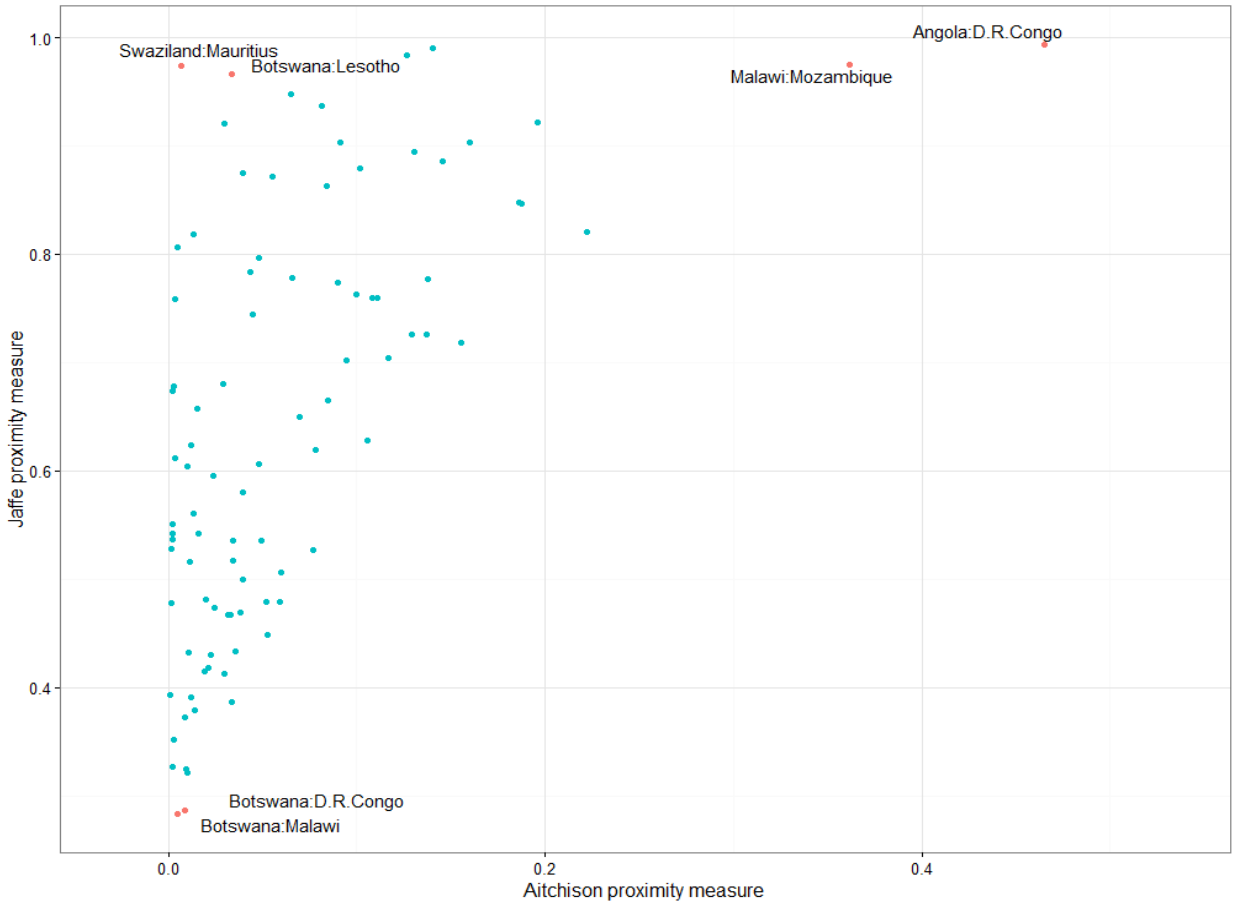

Figure 4: Scatterplot of Jaffe proximity and Aitchison proximity measures for Southern African countries [2011]

The figure 4 shows the scatterplot of 91 observations of Jaffe and Aitchison proximity measures corresponding to the upper triangular matrix of these measures for 14 countries. It is

apparent in the figure that the Jaffe measure is always higher than the Aitchison proximity measure. The correlation between the two measures is about 0.56. The differences in the two measures for the same underlying data are due to the differences in the way the measures are calculated. While the Jaffe measure increases with increasing covariation in the absolute shares of any particular pair of countries, the Aitchison measure increases on the basis of the differences in the log-ratios of shares. Therefore, the Jaffe measure will be high whenever the distribution of shares is similar across a pair of countries (i.e.there is concordance in the ordering of the classes) and higher than the Aitchison measure if some of the shares are very small or large relative to the other shares. Conversely, the Aitchison measure will be high whenever the distribution of the shares is similar and most importantly the distribution of the ratios of the respective shares is similar. The Aitchison measure is therefore more stringent than the Jaffe measure and is likely to be lower than the Jaffe measure.

The figure 4 illustrates the contrasts for some of the country pairs. The table 3 shows the distribution of the underlying shares data for each of the highlighted countries. In the table, the classes with the highest and lowest shares are highlighted to illustrate what drives the differences observed in figure 4. The countries with high Jaffe and high Aitchison measure (i.e. Malawi: Mozambique and Angola: Democratic Republic of Congo) have similar data structure with each having the same classes with highest (roots) and lowest shares (vegetables for Malawi:Mozambique and stimulants for Angola: Democratic Republic of Congo). In addition, the ordering of the other shares is also the same. The countries with high Jaffe and low Aitchison measure (i.e. Swaziland:Mauritius and Botswana:Lesotho) have data structures in which the class with the highest share is the same but is greater than 0.5 (for example, livestock share is 0.8654 for Botswana and 0.5763 for Lesotho) thereby having the other shares with very small shares. Therefore, though the share distributions co-vary implying a higher Jaffe measure, the differences in the log-ratios are huge implying lower Aitchison proximity measure. A similar reason explains the low Jaffe and low Aitchison measures for the pairs- Botswana:Malawi and Botswana: Democratic Republic of Congo.

Table 3: Underlying shares data [year 2011] for highlighted countries in figure 4

Share classes	**Countries**							
	ANG	BT	D.R.C	LES	MW	MAU	MOZ	SWA
Cereals	0.0497	0.0329	0.0707	0.1132	0.1832	0.0005	0.1438	0.0410
Roots	0.4385	0.0516	0.4386	0.1413	0.3102	0.0150	0.3606	0.0377
Sugar & Pulses	0.0925	0.0187	0.1165	0.0329	0.1697	0.5480	0.2110	0.5736
Vegetables	0.0187	0.0236	0.0292	0.0435	0.0245	0.0680	0.0396	0.0109
Fruits	0.2452	0.0069	0.2011	0.0399	0.1027	0.0310	0.0649	0.1041
Stimulants & Tobacco	0.0161	0.0010	0.0291	0.0529	0.1159	0.0123	0.0609	0.0034
Livestock	0.1393	0.8654	0.1148	0.5763	0.0937	0.3251	0.1193	0.2293

Notes: ANG=Angola, BT=Botswana, D.R.C=Democratic Republic of Congo, LES=Lesotho, MW=Malawi, MAU=Mauritius, MOZ=Mozambique, SWA=Swaziland.

The resulting biplot from using the proposed distance measure is called the relative variation biplot because it represents variation in all the component ratios. A robust version of compositional biplot proposed by (Boogaart and Tolosana-Delgado 2013) allows angles and distances in the simplex to be associated with angles and distances in real space. The problem with log ratio transformations in compositional data analysis is the preponderance of zeros due to measurement error or a structural missing category in one of the countries/ firms. For instance, with output shares; it is likely that some commodities are country specific and not produced by some countries. The log of zero is undefined such that a dilemma arises when using these methods. This is the key limitation of using the Aitchison distance measure and empirical economists need to consider the tradeoff between theoretical superiority of the Aitchison distance measure and the computational convenience of the Jaffe measure. The challenge of zero shares can however be dealt with by following several zero imputation algorithms proposed by Fry, Fry, and McLaren (2000) or calculate the distance measures using the positive shares across any two countries (i.e. depend on the independence of irrelevant technological classes property of the Aitchison distance) Among the proposed zero value imputation algorithms, almagamation is the simplest and involves just reducing the number of components by grouping them in such a way that zero shares disappear. This inevitably leads to informational losses. In the analysis, we eliminated zeros through almagamation of components with most zeros and deletion of one country (Seychelles) which had zeros for most of the commodities.

The interpretation of a resulting compositional biplot in Figure 5 is slightly different from the interpretation of the conventional biplots in Figure 1 and Figure 2. Firstly, the distances between row points approximate the technological/agro-ecological distance between countries in the Figure 5. For instance, Angola (1) and Democratic Republic of Congo (3) are closer in agricultural R&D potential than Angola (1) and Botswana (2). Secondly, the distances between column points are approximations of the standard deviation of the corresponding log-ratios (Aitchison and Greenacre 2002). There is therefore high relative variation between fibres and livestock as shown by the long distance between them. Finally, the cosine of the angles approximates the correlation of log-ratios. It is apparent that there are correlations among commodity groups that can distort the way similarities among countries can be interpreted when one ignores inter-technology/commodity group distances. This is exacerbated by the fact that in all applications, aggregations of technology areas are rather arbitrary and subjective.

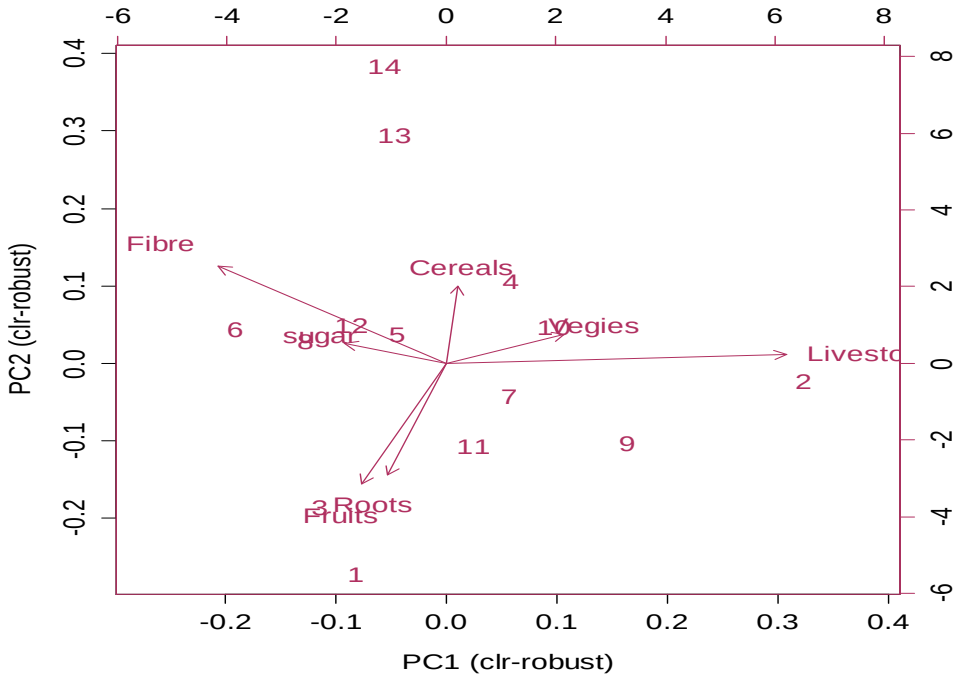

Figure 54: Robust compositional biplot with amalgamated commodity shares

Notes: Numbers represent countries. 1= Angola, 2=Botswana, 3=Democratic Republic of Congo, 4=Lesotho, 5=Madagascar, 6=Malawi, 7=Mauritius, 8=Mozambique, 9=Namibia, 10=South Africa, 11= Swaziland, 12=Tanzania, 13=Zambia, 14=Zimbabwe.

6. Summary of axioms for comparing spill over measures

Bloom, Schankerman, and Reenen (2013) proposed the first ever series of desirable properties (axioms) of distance measures that can be used to decide which measure to use based on the research question. The axioms are shown in the first column of Table 4 below. Table X in Bloom, Schankerman, and Reenen (2013) also had economic microfoundations as one of the axioms. This remains the most important axiom but for which all the distance measures considered fail to satisfy. They concluded from a list of distance measures that , Jaffe's measure and the Mahalanobis distance satisfy the most axioms. Nevertheless, the Jaffe measure, which has been the benchmark for empirical spillover research for almost two decades, is strictly dominated by the Mahalanobis measure. Precisely, the Mahalanobis measure allows for inter-group distance, which is important due to the arbitrary groupings of the technology classes and the fact that crop management research may apply across the different commodity groups. Both Mahalanobis and Jaffe measure fail to satisfy the non-overlapping fields property which requires that the index be invariant to the allocation of R&D by country *i* in fields/technology classes where country *j* does no R&D and which are not technology related to those in which country *j* is active. This property says that technological proximity between two countries should depend only on the extent to which their R&D overlaps. This axiom is related to the subcompositional coherence in compositional data

analysis and the well known axiom in economics-the independence of irrelevant alternatives (IIA). Bar and Leiponen (2012) argued that the independence of irrelevant (patent) classes is a desired property of technological measures in that technological distance between two firms/countries should depend on the shares of patents these firms/countries have in classes in which they both actively patent, but should not depend on how patents for one firm/country are distributed between classes in which the other firm does not patent at all.

The proposed Aitchison distance/proximity measure strictly dominates the Jaffe measure and Mahalanobis distance measure as it satisfies the non-overlapping fields axiom. There is therefore a case for the R&D spillover literature to consider Aitchison measure as a better alternative to Jaffe measure. The seminal monograph (Aitchison 1986) on this distance measure and the other compositional data analysis literature following it means that Aitchison distance measure is on a better footing interms of its theoretical properties than the Jaffe measure. For instance, the Aitchison distance measure has well defined statistical distributions, inferential apparatus, and thus can allow a complete characterization of parameter uncertainty of the weights into the economic analysis of R&D spillovers.

Table 4:
Desirable properties and axioms of distance measures

Properties	Jaffe	Mahalanobis	Aitchison
Definition/ Formula	$\frac{Y_i' Y_j}{\sqrt{Y_i}\sqrt{Y_j}}$	$\frac{Y_i' \Omega Y_j}{\sqrt{Y_i}\sqrt{Y_j}}$	$\sqrt{\frac{1}{2N}\sum_{n=1}^{N}\sum_{n'=1}^{N}(\ln\frac{y_{in}}{y_{in'}} - \ln\frac{y_{jn}}{y_{jn'}})^2}$
Biplot type	Column centered	CVA biplot	Compositional biplot
Scale invariance	X	X	X
Within Group Overlap	X	X	X
Between Group Overlap		X	X
Non overlapping fields (Sub-compositional coherence/ Independence of irrelevant classes)			X
Invariance to aggregation over non-active fields/crops	X	X	X
Robustness to aggregation of active fields/crops	X	X	X

Source: adapted from Bloom, Schankerman, and Reenen (2013).

Notes: Y_i denotes the vector of output shares of country i in different crop categories, y_{in} and $y_{in'}$ are elements of the vector, N is the number of technology categories, and Ω is the mahalanobis matrix summarizing the co-location of crop categories. An "X" denotes that the distance measure has the indicated property, whereas a blank indicates that it does not.

7. Conclusions and implications

This paper has introduced two important extensions to the literature on technological proximity measures. These are: the biplot, as a graphical display of technological proximity measures and the Aitchison proximity measure, as a numerical metric that captures the adding up constraint of the sample space for calculating technological proximity measures. The analysis presented in the paper is an attempt to offer the richness of a graphical approach in R&D spillover exploratory analysis. In terms of agricultural R&D policy, there are several interesting findings that are elucidated by a graphical analysis. The first is that, the findings from the graphical analysis are consistent with the literature on agricultural R&D spillovers in sub-Saharan Africa. Second, the spatial and temporal convergence in the spill over potential based the findings of this study reflects the emerging globalization trends across all Africa and developing countries. The convergence results can be explained by the fact that spill over potential of crop varietal technologies has increased overtime because environments have been "homogenized" owing to the widespread adoption of modern technology such as irrigation and fertilizer and because tastes and preferences have begun to converge (Byerlee and Traxler 2001). Further research is needed to explain the slight divergence in the period from 2001 to 2011. Though regional agricultural research institutions like Forum for Agricultural Research in Africa (FARA) and Centre for Coordination of Agricultural Research and Development for Southern Africa (CCARDESA) were developed on the basis of geopolitical reasons; the graphical analysis has shown that further assessment of whether there is research policy integration potential within these regions is required. This study has shown that even within a regional group like CCARDESA, there are similarities and dissimilarities that suggest country groupings in terms of agricultural R&D policy.

In terms of the numerical metric, the paper has shown that it is important to recognize the sample space of the data that is mostly used to calculate the technological proximity measures. Almost all studies using the technological proximity measures use shares of patents or citations etc. that add up to a constant number, 1. In this sample space, Aitchison (1986) and the compositonal data analysis literature have proved that proximity measures should satisfy certain properties including: scale invariance and subcompositional coherence. It has been shown in this paper and also in the compositional data literature e.g. Aitchison (1992), that the Jaffe measure fails on the subcompositional coherence property. This property is equivalent to the non-overlapping fields axiom and the independence of irrelevant patent classes property introduced in the economics of innovation literature by Bloom, Schankerman, and Reenen (2013) and Bar and Leiponen (2012) respectively, who both showed that the Jaffe measure fails on this axiom. The Aitchison proximity measure satisfies all the axioms suggested in the literature. The limitation of using this measure is that it cannot be used when the data has zero shares. In that scenario, researchers can almagamate shares to eliminate zero shares or use positive shares for overlapping classes only and rely on subcompositional coherence.

It follows then from study's findings that biplot analysis can be a visualization tool for exploratory analysis of R&D spillovers. The data transformation and software requirements are

also currently available in many traditional statistical computing environments. In addition, biplot analysis can be extended to multivariate cases as demonstrated in the paper thereby providing a good exploratory data analysis approach that can help explore the variables to include in the knowledge production function. The realization that Jaffe's measure is constructed using compositional data can also open much deeper analysis of spillovers using the well-developed compositional data analysis methods like the Aitchison distance measure. In particular, the Aitchison distance measure has well defined statistical distributions, inferential apparatus and thus can allow a complete characterization of the parameter uncertainty of the weights into the economic analysis of R&D spillovers. These extensions can therefore help researchers analyzing R&D spillover potential to consider multiple dimensions of visualizing and understanding spillovers.

Acknowledgements

This research was conducted when the author was being supported by the McKnight Foundation under the Collaborative Crop Research Program for his PhD studies in the Department of Applied Economics, University of Minnesota-Twin Cities. I thank Philip Pardey, late Jason Beddow and colleagues at the International Science & Technology Policy & Practice (INSTEPP) Center for comments on an earlier draft of the paper. I also thank the editor and the anonymous referees for suggestions that significantly improved the paper. The paper is dedicated to the memory of Jason Beddow who inspired me through his lectures to work on agricultural research spillovers. The usual disclaimer applies to institutions and people mentioned.

References

Adams, James D. 1990. "Fundamental Stocks of Knowledge and Productivity Growth." *Journal of Political Economy* 98 (4): 673. doi:10.1086/261702.

Aitchison, John. 1990. "Relative Variation Diagrams for Describing Patterns of Compositional Variability." *Mathematical Geology* 22 (4): 487–511. doi:10.1007/BF00890330.

Aitchison, John. 1986. *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

———. 1992. "On Criteria for Measures of Compositional Difference." *Mathematical Geology* 24 (4): 365–79. doi:10.1007/BF00891269.

Aitchison, John, and Michael Greenacre. 2002. "Biplots of Compositional Data." *Applied Statistics* 51 (4): 375–92. doi:10.1111/1467-9876.00275.

Aldieri, Luigi, and Michele Cincera. 2009. "Geographic and Technological R&D Spillovers within the Triad: Micro Evidence from US Patents." *Journal of Technology Transfer* 34 (2): 196–211. doi:10.1007/s10961-007-9065-8.

Alene, Arega D., and Ousmane Coulibaly. 2009. "The Impact of Agricultural Research on

Productivity and Poverty in Sub-Saharan Africa." *Food Policy* 34 (2). Elsevier Ltd: 198–209. doi:10.1016/j.foodpol.2008.10.014.

Alston, Julian M., Jennifer S. James, Matthew A. Andersen, and Philip G. Pardey. 2010. *Persistence Pays: U.S. Agricultural Productivity Growth and the Benefits from Public R&D Spending*. doi:10.1007/978-1-4419-0658-8.

Alston, Julian M., Norton, George. W., and Pardey, Philip. G. (1998). *Science under scarcity: principles and practice of agricultural research evaluation and priority setting.* New York: Cab International.

Bar, Talia, and Aija Leiponen. 2012. "A Measure of Technological Distance." *Economics Letters* 116 (3). Elsevier B.V.: 457–59. doi:10.1016/j.econlet.2012.04.030.

Bloom, Nicholas, Mark Schankerman, and John Van Reenen. 2013. "Identifying Technology Spillovers and Product Market Rivalry." *Econometrica* 81 (4): 1347–93. doi:10.3982/ECTA9466.

Boogaart, K. Gerald Van Den, and Raimon Tolosana-Delgado. 2013. *Analyzing Compositional Data with R*. doi:10.1007/978-3-642-36809-7.

Branstetter, Lee G. 2000. "Looking for International Knowledge Spillovers A Review of the Literature with Suggestions for New Approaches." In *The Economics and Econometrics of Innovation*, 495–518. Boston, MA: Springer US. doi:10.1007/978-1-4757-3194-1_20.

Byerlee, Derek, and Greg Traxler. 2001. "The Role of Technology Spillovers and Economies of Size in the Efficient Design of Agricultural Research Systems." In *Agricultural Science Policy: Chaning Global Agendas*, edited by Julian M. Alston, Philip G. Pardey, and Michael J. Taylor, 161–86.

Campbell, N. A., and William R. Atchley. 1981. "The Geometry of Canonical Variate Analysis." *Systematic Zoology* 30 (3): 268. doi:10.2307/2413249.

Chyi, Yih-Luan, Yee-Man Lai, and Wen-Hsien Liu. 2012. "Knowledge Spillovers and Firm Performance in the High-Technology Industrial Cluster." *Research Policy* 41 (3). Elsevier B.V.: 556–64. doi:10.1016/j.respol.2011.12.010.

Cincera, Michele. 2006. "Firms' Productivity Growth and R&D Spillovers: An Analysis of Alternative Technological Proximity Measures." *Economics of Innovation and New Technology* 14 (8): 657–82. doi:10.1080/10438590500056768.

Eckart, C, and G Young. 1939. "A Principal Axis Transform for Non-Hermitian Matrices." *Bulletin of the American Mathematical Society* 45 (1939): 118–21.

Fry, Jane M., Tim R. L. Fry, and Keith R. McLaren. 2000. "Compositional Data Analysis and Zeros in Micro Data." *Applied Economics* 32 (8): 953–59. doi:10.1080/000368400322002.

Gabriel, K. R. 1971. "The Biplot Graphic Display of Matrices with Application to Principal Component Analysis." *Biometrika* 58 (3): 453. doi:10.2307/2334381.

Gower, John C, Niel J. le Roux, and Sugnet Gardner-Lubbe. 2014. "The Canonical Analysis of Distance." *Journal of Classification* 31 (April): 107–28. doi:10.1007/s00357-014-9149-8.

Gower, John, Sugnet Lubbe, and Niël Le Roux. 2011a. *Understanding Biplots*. Chichester, UK: John Wiley & Sons. doi:10.1002/9780470973196.

Greenacre, Michael J. 2012. "Biplots: The Joy of Singular Value Decomposition." *Wiley Interdisciplinary Reviews: Computational Statistics* 4 (4): 399–406. doi:10.1002/wics.1200.

Grilliches, Zvi. 1979. "Issues in Assessing the Contribution of Research and Development to Productivity Growth." *The Bell Journal of Economics* 10 (1): 92–116.

Griliches, Zvi. 1992. "The Search for R&D Spillovers." *The Scandinavian Journal of Economics* 94: S29–47. doi:10.2307/3440244.

Hron, Karel, and Peter Filzmoser. 2015. *Exploring Compositional Data with the Robust Compositional Biplot*. *Advances in Latent Variables*. doi:10.1007/978-3-319-02967-2.

Jaffe, Adam B. (1986). Technological Opportunity and Spillovers of R & D: Evidence from Firms' Patents, Profits, and Market Value. *The American Economic Review, 76*(5), 984–1001. http://doi.org/10.2307/1816464

Jaffe, Adam B. 1989a. "Characterizing the 'technological Position' of Firms, with Application to Quantifying Technological Opportunity and Research Spillovers." *Research Policy* 18: 87–97. doi:10.1016/0048-7333(89)90007-3.

———. 1989b. "Real Effects of Academic Research." *American Economic Association* 79 (5): 957–70.

Johnson, Michael, Sam Benin, Xinshen Diao, and Liangzhi You. 2011. "Prioritizing Regional Agricultural R&D Investments in Africa: Incorporating R&D Spillovers and Economywide Effects." *ASTI/IFPRI-FARA Conference*.Accra, Ghana. http://www.asti.cgiar.org/pdf/conference/Theme4/Johnson.pdf.

Johnson, Michael, Samuel Benin, Liangzhi You, Xinshen Diao, Pius Chilonda, and Adam Kennedy. 2014. "Exploring Strategic Priorities for Regional Agricultural Research and Development." 1318. *IFPRI Discussion Paper*. IFPRI Discussion Papers. Washington, D.C.

Kaiser, Ulrich. 2002. "Measuring Knowledge Spillovers in Manufacturing and Services: An Empirical Assessment of Alternative Approaches." *Research Policy* 31 (1): 125–44. doi:10.1016/S0048-7333(00)00159-1.

Kohler, Ulrich, and Magdalena Luniak. 2005. "Data Inspection Using Biplots." *Stata Journal* 5(2):208-223. http://www.stata-journal.com/article.html?article=gr0011.

Kroonenberg, Pieter M. 1995. "Introduction to Biplots for GXE Tables." https://openaccess.leidenuniv.nl/bitstream/handle/1887/11604/7_702_074.pdf?sequence=1.

McNamee, Robert C. 2013. "Can't See the Forest for the Leaves: Similarity and Distance Measures for Hierarchical Taxonomies with a Patent Classification Example." *Research*

Policy 42 (4). Elsevier B.V.: 855–73. doi:10.1016/j.respol.2013.01.006.

Parent, Olivier, and James P. LeSage. 2008. "Using the Variance Structure of the Conditional Autoregressive Spatial Specification to Model Knowledge Spillovers." *Journal of Applied Econometrics* 23 (2): 235–56. doi:10.1002/jae.981.

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Ros-Freixedes, Roger, and Joan Estany. 2013. "On the Compositional Analysis of Fatty Acids in Pork." *Journal of Agricultural, Biological, and Environmental Statistics* 19 (1): 136–55. doi:10.1007/s13253-013-0162-x.

Roux, N. l., and Lubbe, S. (2013). UBbipl: Understanding biplots: datasets and functions. R package version 3.0.4. Retrieved from http://www.wiley.com/go/biplots

Simon, H., and N. Sick. 2016. "Technological Distance Measures: New Perspectives on Nearby and Far Away." *Scientometrics* 107 (3). Springer Netherlands: 1–22. doi:10.1007/s11192-016-1888-3.

Templ, M., Hron, K., and Filzmoser, P. (2011). robCompositions: an R-package for robust statistical analysis of compositional data. In P.-G. a. Buccianti (Ed.), *Compositional Data Analysis: Theory and Applications* (pp. 341-355). Chichester, UK: John Wiley & Sons.

Yan, Weikai, and Manjit S Kang. 2003. *GGE Biplot Analysis: A Graphical Tool for Breeders, Geneticists, and Agronomists*. CRS Press.

Yan, Weikai. (2014). *Crop Variety Trials: Data Management and Analysis.* Sussex: John Wiley & Sons, Ltd.

Yan, Weikai, and Tinker, N. A. (2006). Biplot analysis of multi-environment trial data: principles and applications. *Canadian Journal of Plant Science*, 623-645.

Appendix: Pseudo R code for the tables and graphs

```
# Table 1
vshares2011=read.table("sadc_value_shares.csv",sep=",", header=TRUE)
attach(vshares2011)
vshares=vshares2011
# Cosine similarity/Jaffe Omega Function
jaffem=function (d, w = rep(1, nrow(d))/nrow(d))
{
  s <- sum(w)
  m1 <- sum(d[, i] * w)/s
  m2 <- sum(d[, j] * w)/s
  (sum(d[, i] * d[, j] * w)/s)/sqrt((sum(d[, i]^2 *
                                         w)/s) * (sum(d[, j]^2 * w)/s))
}
nr = ncol(vshares)
nc = ncol(vshares)
r=array(0,dim=c(nr,nc))
```

```
for (i in 1:nr) {
  for (j in 1:nc){
r[i,j]=jaffem(vshares)}}

# Table 2
shares=read.table("shares.txt")
jaffem=function (d, w = rep(1, nrow(d))/nrow(d))
{
  s <- sum(w)
  m1 <- sum(d[, i] * w)/s
  m2 <- sum(d[, j] * w)/s
  (sum(d[, i] * d[, j] * w)/s)/sqrt((sum(d[, i]^2 *
                                         w)/s) * (sum(d[, j]^2 * w)/s))
}
nr = ncol(shares)
nc = ncol(shares)
r=array(0,dim=c(nr,nc))
for (i in 1:nr) {
  for (j in 1:nc){
r[i,j]=jaffem(shares)}}

# Figure 1
library(UBbipl)
shares=read.table("shares.txt")
figure1=PCAbipl(shares,scaled.mat = TRUE, colours = "black", pch.samples =
15,pch.samples.size=1,rotate.degrees=70,offset = c(-0.2, 4, 0.1, 0),ax.name.size=0.6)

# Figure 2
sharesd=read.table("sharesd.txt")
b=PCAbipl(sharesd ,scaled.mat = TRUE,colours = "black", pch.samples =
15,pch.samples.size=1,rotate.degrees=-100,offset = c(-0.2, 4, 0.1, 0),ax.name.size=0.6)

# Figure 3
decadeshares=read.table("decadeshares.txt",header=TRUE)
CVAbipl(decadeshares[,3:11], G = indmat(decadeshares[,2]),colours =
c("red","blue","green","brown","black"),density.plot = "groups", legend.type =
c(TRUE,TRUE,TRUE,TRUE,TRUE),line.width = 2,alpha = 0.95, specify.bags = 1:5)

# Figure 5
amalgamatedsharet=read.table("amalgamatedsharet.txt",header=TRUE)
library(dplyr)
amalgamatedsharet=rename(amalgamatedsharet,sugar=SugarPulsesNutsOil)
amalgamatedsharet=rename(amalgamatedsharet,Fibre=FibreSpicesStimulantTobacco)
attach(amalgamatedsharet)
library(compositions)
library(robCompositions)
PrinCompRob <- pcaCoDa(amalgamatedsharet, method="robust")
plot(PrinCompRob,col="black")

# Subcompositional coherence example
library(compositions)
library(robCompositions)
A=c(2/4,1/4,1/4)
B=c(1/4,2/4,1/4)
shares=cbind(A,B)
dij=aDist(A,B)
wijA=1/exp(dij)
wijA
cor(A,B)

jaffem=function (d, w = rep(1, nrow(d))/nrow(d))
{
 s <- sum(w)
 m1 <- sum(d[, i] * w)/s
 m2 <- sum(d[, j] * w)/s
 (sum(d[, i] * d[, j] * w)/s)/sqrt((sum(d[, i]^2 *
 w)/s) * (sum(d[, j]^2 * w)/s))
}
nr = ncol(shares)
nc = ncol(shares)
```

```
r=array(0,dim=c(nr,nc))
for (i in 1:nr) {
 for (j in 1:nc){
r[i,j]=jaffem(shares)}}
r

C=c(2/3,1/3)
D=c(1/3,2/3)
sharesub=cbind(C,D)
dij=aDist(C,D)
wijA=1/exp(dij)
wijA
cor(C,D)

jaffem=function (d, w = rep(1, nrow(d))/nrow(d))
{
 s <- sum(w)
 m1 <- sum(d[, i] * w)/s
 m2 <- sum(d[, j] * w)/s
 (sum(d[, i] * d[, j] * w)/s)/sqrt((sum(d[, i]^2 *
 w)/s) * (sum(d[, j]^2 * w)/s))
}
nr = ncol(shares)
nc = ncol(shares)
r=array(0,dim=c(nr,nc))
for (i in 1:nr) {
 for (j in 1:nc){
r[i,j]=jaffem(sharesub)}}
r
```