Statistical Evaluation of Biofuel Role in Land Use Change

Kobi Abayomi, Valerie Thomas, Dexin Luo*

November 17, 2009

Abstract

We investigate the effect of biofuels on land use change through a case study of U.S. corn production. Currently, agricultural models are used to estimate the effect of biofuel production on crop production and, correspondingly, land use change. As biofuel production grows, the models can potentially be validated with statistical analysis of changes in crop production. Determination of the influence of biofuel production on other uses of the biofuel feedstock - such as for food or animal feed - cannot be evaluated with standard statistical methods because the uses of an agricultural crop are constrained: total crop use is always the sum of the constituents. We develop a general method for these *compositional distributions*, and apply this method to determine competition between biofuel feedstock production and other uses of the same feedstock. We find evidence of competition among corn yield constituents, particularly with respect to ethanol production.

Introduction

The Energy Independence and Security Act (EISA) of 2007 mandates an increase in ethanol production to 36 billion gallons per year by 2022.[1] In 2008 the United States produced 9 billion gallons of ethanol fuel — an increase of more than 5000 percent since 1980.[5] At the same time, the total U.S. corn yield has less than

*All authors: School of Industrial and Systems Engineering (ISYE), Georgia Institute of Technology.

doubled. Increases in the remaining constituents of corn use are similar — between 50 and 200 percent. See Figures 1 and 2.

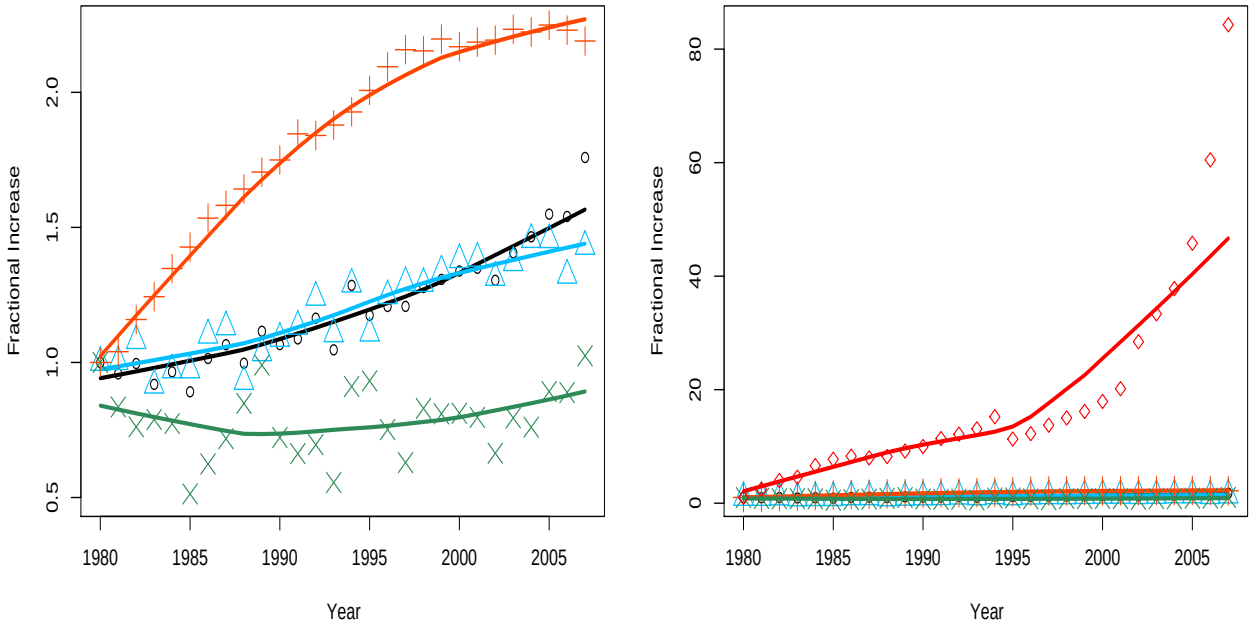


Figure 1: Fractional increase in corn production and constituents, 1980-2007, in color: Residual food, feed stocks, exports, ethanol, and total output. The smoothed curve suggests the trend. Yearly values are ratios to 1980 production of/by each constituent output. Total corn yield, food and feed stocks have increased similarly; residual food output has increased more dramatically. Exports have decreased relative to 1980. Left hand graph excludes ethanol fractional increase: fractional increases for residual food, feed stocks, exports and total output appear flat with respect to increase in ethanol production. All data in thousands of bushels.

The environmental impacts of this mandate (net energy budget, effect on corn based commodities, greenhouse emissions, etc.) are unresolved, significant[6], and addressed elsewhere [3]. In this brief paper we develop a straightforward method for

using agricultural data to investigate how increased biofuel production affects production of other agricultural crops. We illustrate the methods with a case study of dependency within the distribution of the constituents of total corn yield, specifically:

- Feed Stocks
- Exports
- Residual Food Stocks
- Ethanol

as a fraction of total output.

At one extreme, increasing biofuel production from corn could yield more corn production and no change in the production of corn for other uses. Oppositely, this increase could result in an immediate reduction in other uses of corn with the increased use of corn for biofuel. These phenomena can be examined from either a crop production, as is done here, or from land use; these results are parallel and mediated by yield.

Corn is not the only crop where these questions arise. Additional cases of interest include: the influence of sugarcane-derived ethanol production on land use in Brazil, the influence of soy biodiesel production on land used for soy production (note that soy and corn production in the US are not independent - many farms practice corn-soy rotations) and, more broadly, the influence of biofuel production even from non-food crops on food crop production.

Currently, models of land use change from biofuel production rely either on basic assumptions about the extent to which biofuel production displaces other uses of the feedstock, or use agricultural models including FAPRI and FASOM (references needed here). As biofuel production grows, there is potential to validate the assumptions of these models using data on crop production and biofuel production. We develop that methodology here.

Constituents of Corn Yield: Compositional Data

The data are 28 years — from 1980 to 2007 — of the allocations of total corn yield in thousands of bushels ([32]). Preparatory illustrations of the fractional increases

(Figure 1) and relative fractions (Figure 2) illustrate the nominal increase in ethanol output and the relative increase in corn allocation to ethanol production.

Figure 2 is an illustration of the statistic of interest: the joint distribution of the relative fractions, modulo the total yield, over time. This multivariate distribution — strictly positive, sum fixed and inferior or equal to one — on the simplex is called **a** *compositional distribution* and arises in many contexts (see [8], [9], [10], especially [7]).

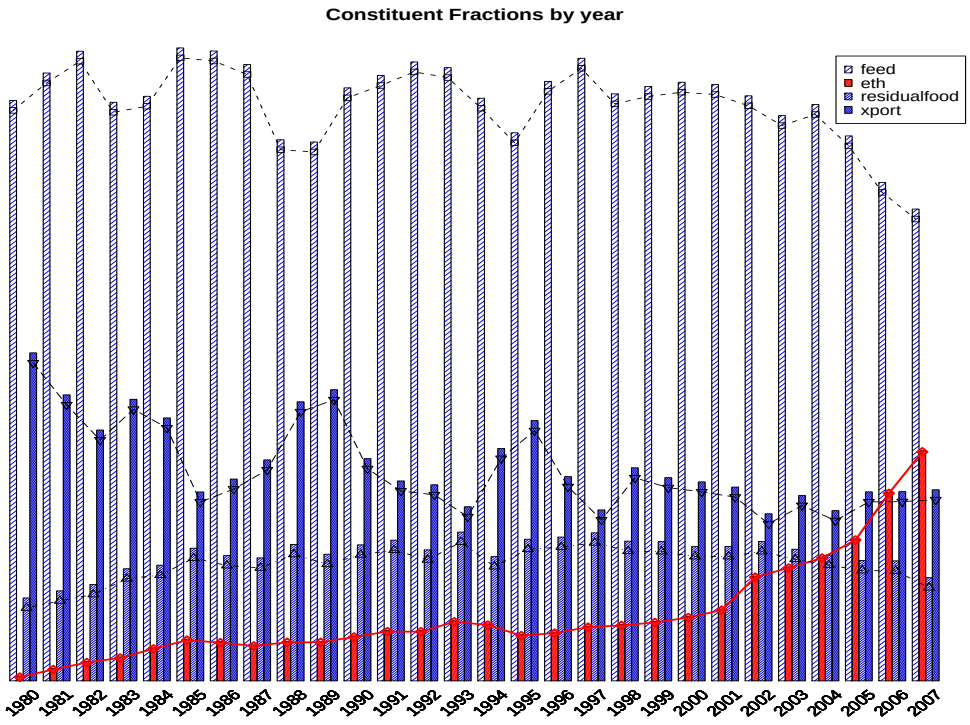


Figure 2: Joint distribution of fractional inputs of corn yield, ethanol in red. The heights of the bars are the ratio of the constituent fraction to the total corn yield (normalized to 1 for each year). The graph is a representation of a three-dimensional positive simplex, a four dimensional composition in Aitchison terminology: the heights are the joint distribution modulo the total yield (and variation).[4] Dependence in this *compositional distribution* is a function of the statistic represented by this illustration.

Isolating dependence in a compositional distribution is non-trivial: the restriction of the distributional shape to the simplex imposes dependency — in particular linear dependence — in the same way segments on a fixed interval are necessarily dependent. Standard inspection and testing of a correlation matrix is insufficient for tests of independence of the compositional distribution. Dependence metrics — like correlation — based upon Euclidean distance are in fact conclusively inappropriate for compositional data (see [11]). This characteristically constrains or wholly excludes standard methods and tests for multivariate independence: like tests of pairwise correlation or multivariate correlation (e.g. Fisher's Z), or multivariate tests relying on distributional assumptions on the covariance matrix (e.g. Wishart type tests).

There are two common, apposite methods for addressing dependency in distributions of proportions (distributions on the simplex). The first is to use a transformation on the compositional data, from the sample space of the simplex to the positive real hyperplane and investigate tests in resultant distribution. The log-ratio transformation is popular; tests of independence are on the constrained covariance matrix of the transformed data (see [12]). A second approach is to conduct testing on the simplex space via a necessary generalization of the Dirichlet distribution — which only admits independent or *neutral* components — to the Liouville distribution (see [2] and [13]).

We engage a blend of these approaches: we transform the sample data via the log-ratio; exploit the natural role of the Dirichlet distribution in neutrality/independence to generate replicates with identical marginal distributions; and conduct independence testing using the Kolmogorov-Smirnov probability measure **of** distance. This technique is computationally straightforward, can flexibly allow independence testing via a variety of similar metrics, and it is in the direction of the more complete *sub-compositional independence* suggested by [13]. This is means of testing independence for compositional data that is accessible to an array of practitioners.

In the body of the paper we illustrate the methodology in an accessible manner: we confine technical concerns to the Appendix where possible.

Methodology

Let

$$\mathbf{x} = (x_1, ..., x_k) \tag{1}$$

be a basis or open vector of positive quantities, $\mathbf{x} \in \mathbb{R}^{k^+}$ — the k dimensional positive hyperplane. In this example the positive quantities are the constituents of total corn yield, in order (in bushels): $\mathbf{x} = (x_{eth}, x_{rfood}, x_{feed}, x_{xport})$; corn to ethanol production, residual food stock, feed stock, and exported.

Let

$$y_j = x_j / \sum_j^k x_j \tag{2}$$

with $\mathbf{y} = (y_1, ..., y_k)$ the vector of fractions; in the Aitchison ([12]) terminology the composition of the basis $\mathbf{x}$. Here the y_j are the (relative to the total) fractions of ethanol, food stocks, feed stocks, and exported (bushels) of corn illustrated in Figure 2. See Appendix 1.

The log-ratio transformation sets

$$v_{\underset{m}{j}} = log(\frac{y_j}{y_m}) = log\, y_j - log\, y_m \tag{3}$$

in a slight modification of Aitchison's notation (where $v_j = log(y_j/y_{k+1})$).

Here, since the total is fixed and known, the residual is y_{k+1}=0 and Aitchison's v_j is undefined. This notation is a natural and useful affixation to Aitchison's; $v_{\underset{m}{j}}$ is the log of the relative fraction of constituent j to constituent m. See Appendix 1.

Distributional Models for Compositional Data

The Dirichlet model for $(y_1, ..., y_k)$; $\sum_j^{k+1} y_j = 1; y_j > 0\ \forall j$, with parameters $\alpha = (\alpha_1, ..., \alpha_{k+1})$ is:

$$dF(\mathbf{y}) \propto (1 - \sum_j y_j)^{\alpha_{k+1}-1} \cdot \prod_j y_j^{\alpha_j - 1} \tag{4}$$

with $\alpha_0 \equiv 0$ (see [18]), and $dF(\cdot)$ the density.

Connor and Mossiman characterize a vector of proportions $(y_1, ..., y_{k+1})$ as Dirichlet distributed in [2] and [16]. Dependency characterization, however, is insufficient for non-independent, or non-*completely neutral* proportions (see [12] and as well [13]): compositional data that are positively associated cannot be modeled via the Dirichlet distribution which precludes an immediate (parametric) test of independence.

Aitchison [17] advocates the use of the log-normal distribution for the vector of log transformed proportions $\mathbf{v}$ in tests of independence: a variance-covariance matrix (Σ) is sufficient for dependency in the log-normal distribution.

Under a composition the covariance matrix has this form:

$$\Sigma_{\mathbf{v}} \propto diag(\omega_1, ..., \omega_k) + \omega_{k+1}, \tag{5}$$

where $\omega_j > 0, \ \forall j$ — thus the variance-covariance matrix is non-diagonal, *even on an independent composition* and strictly positive. See Appendix 2.

Rayens and Srinivasan propose the generalized Liouville distribution — a generalization of the Dirichlet distribution — as a richer model for compositional data under dependence [13].

A Liouville distribution is

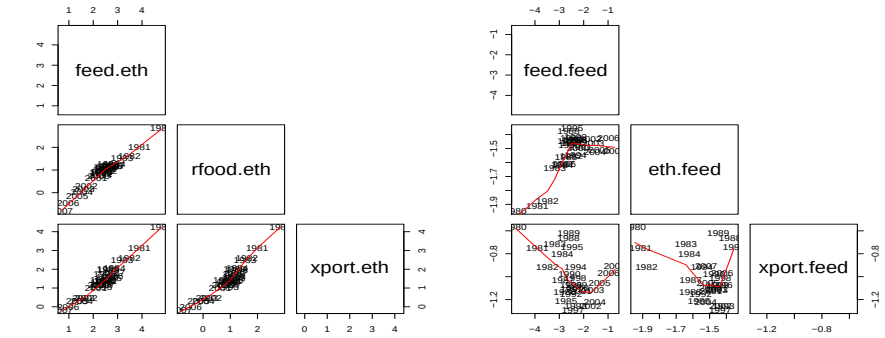
$$dF \propto h(\sum_j y_j) \prod y_j^{\alpha_j - 1} \tag{6}$$

with $\alpha_j > 0$ (as before) and h some function. Note that when $h(t) = 1 - t$ the Liouville distribution is the special case Dirichlet distribution with $\alpha_{k+1} = 1$.

We propose, in an alloy and extension of Aitchison's and Rayens' methods, to test independence using distance on probability measures:

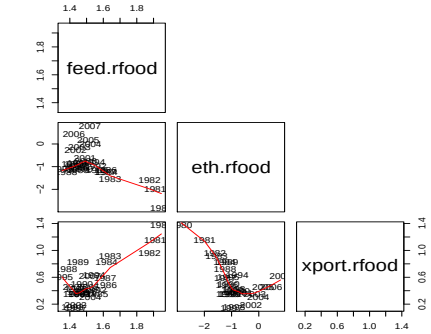
- We exploit the special role of the Dirichlet as the *neutral* distribution to generate marginally equivalent multivariate replicates.
- We apply Aitchison's log-ratio transform m-fold times, holding each component of the data as the residual singly
- We generate, via simulation, the distribution for a statistic on the distance between independence and dependence. This is a simulated distribution for a *measure of association.*

K-dimensional random, sum-constrained, positive replicates are marginally Dirichlet equivalent but jointly Liouville distributed — *as such they are not necessarily neutral* or jointly independent. This allows generation of measures of association or

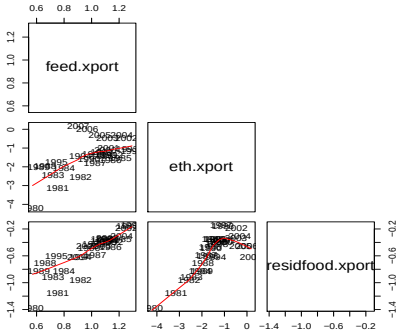


(a)

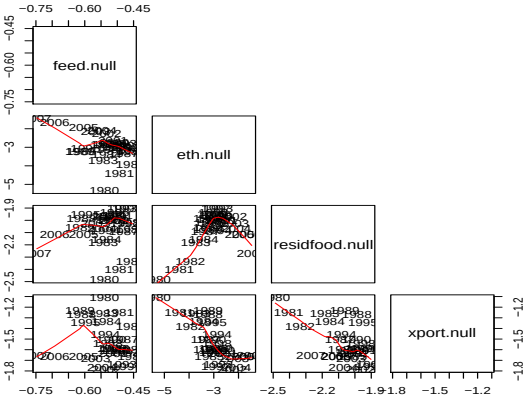
(b)



(c)



(d)



(e)

Figure 3: Scatterplots of log-ratios of constituents of total corn yield (v_{j_m}), labeled by years. $m = (eth,\ feed,\ residual\ food,\ export,\ total)$ in plots (a)-(e), in order. Observed log-ratios are labeled by year. Loess smoothed curve in red. The dependence among the log-ratios appears to vary by choice of 'residual' m.

distance measures — computed on these replicates — to envelop dependency beyond

mere neutrality. See Appendices 2 and 3.

Results

Figure 3 illustrates the data we test for independence. Each of the subplots (a)-(d) are the joint distributions of the compositions with one constituent as 'residual'. The plots illustrate the conditional joint dependence among the constituents, with respect to the 'residual' constituent.

The plots illustrate the pairwise dependence among the joint conditional distributions; a LOESS (smooth regression) curve is fit on each pair (see [30]). The plots and LOESS fits suggest dependency exists within the subcompositions; the pairwise plots, though, are imperfect illustrations for multivariate dependency. The data in the plots are labeled by year.

These data are the observed values of $\underset{m}{\mathbf{v}}_j$ — the log-ratios of the compositions. We *do not fit* a logistic-normal distribution ([31]) to the data; in fact, we make no distributional assumption in the calculation of the distance from independence, beyond the utilization of random, marginal Dirichlets as replicates for the compositional data.

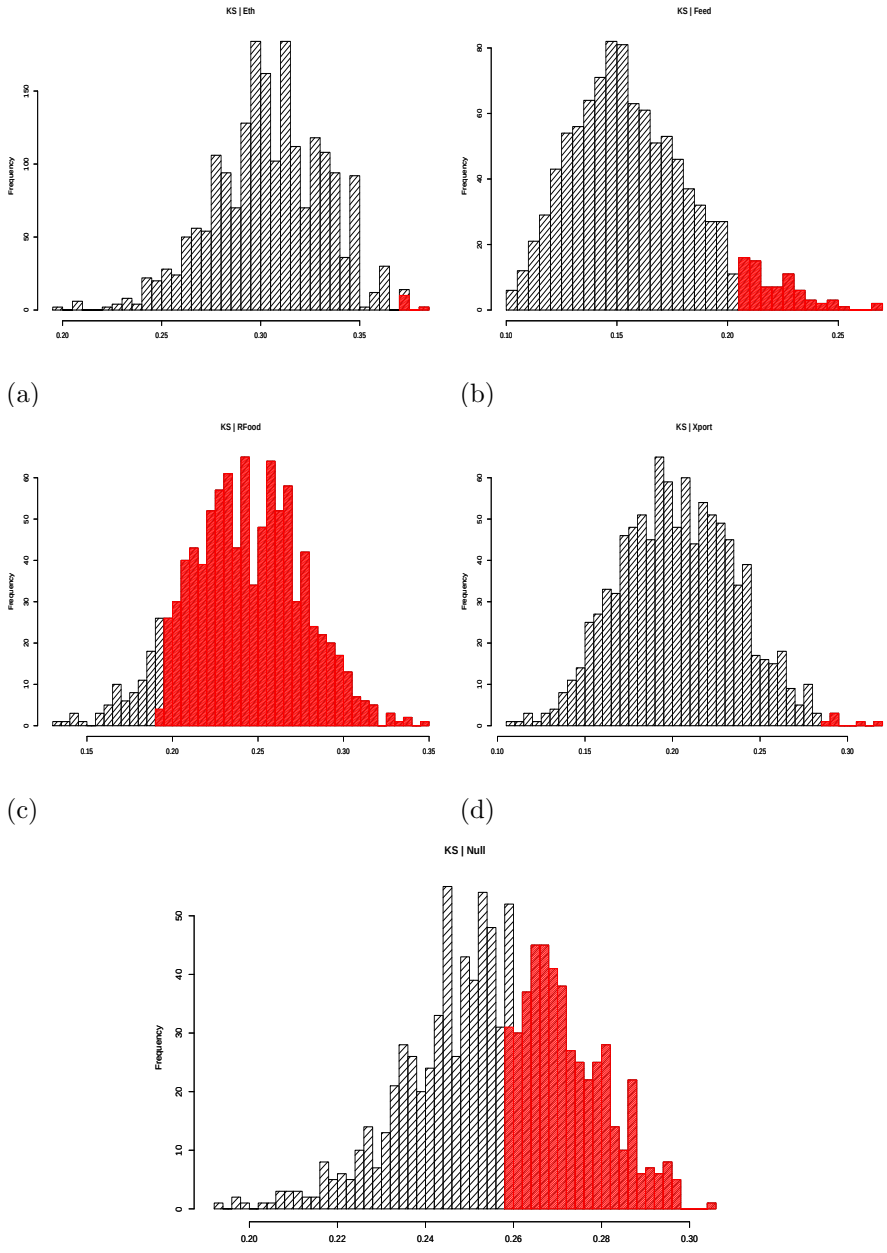
Statistical Dependency among Constituents of Corn Yield

Figure 4 is the distributions of the statistic for the test of independence ($D^{\Pi}_{n,k}$); for the null composition — all of the constituents of corn yield together — and for each of the subcompositions. The observed statistic for each is highlighted by the leftmost border of the red shaded area: the shaded area are the replicates which are greater than the observed distance. The statistic increases with distance from independence; thus the shaded areas are simulated *p-values* for the compositions.

The subcompositions with respect to ethanol and exported corn are highly significantly dependent — p-values of .002 and .007. The subcomposition with respect to feed stock has a p-value of .073.

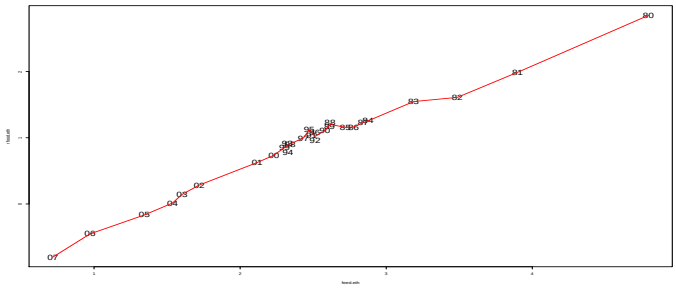
The remaining subcomposition — with respect to residual food — and the null composition have p-values for dependence of .909 and .473 which do not suggest evidence of dependency.

The distance statistics are scalar measures for the multivariate dependency within each (sub)composition. The joint distribution within each of the subcompositions

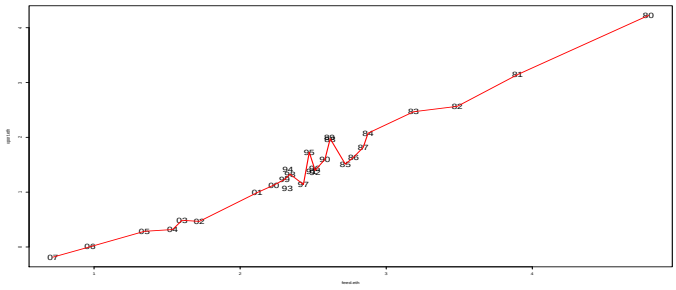


(a) (b) (c) (d) (e)

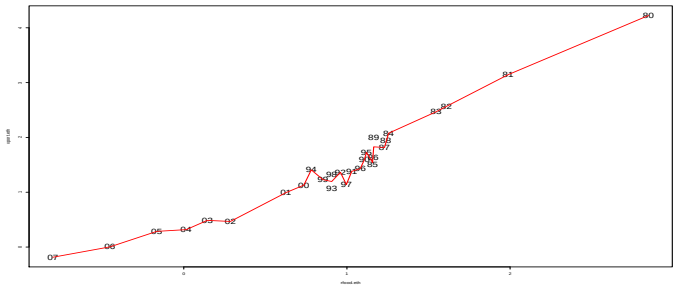
Figure 4: Histograms of the distributions of distances from independence-neutrality $(D^{\Pi,1}_{n,k_m}, ..., D^{\Pi,T}_{n,k_m})$, for $m = (eth,\ feed,\ residual\ food,\ export,\ total)$ in plots (a)-(e), in order. Here $n = 28$, $k = 3, 4$, $T = 1000$. The area in red are values above the observed D^{Π} for each choice of 'residual'. These areas are analogs of *p-values* for the test of distance from independence. The joint distributions of the compositions with respect to ethanol, animal feed stocks, and exports are far from independence: the observed p-values — 0.002, 0.006, 0.073, in order.



(a)

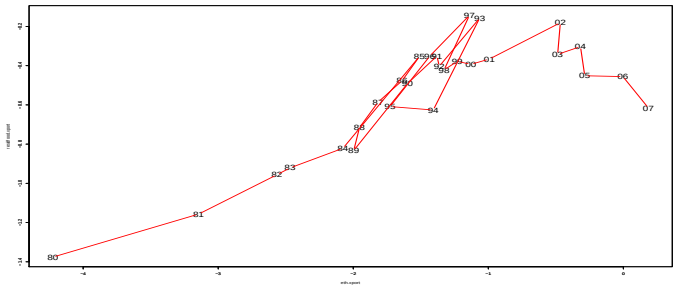


(b)

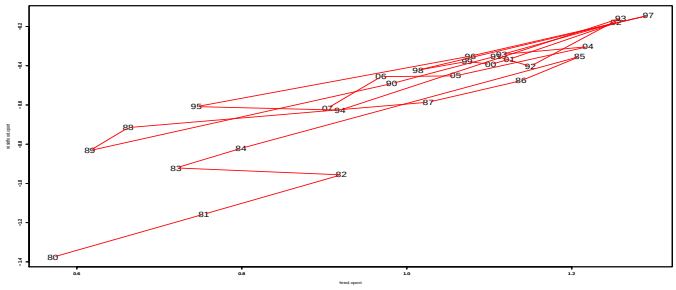


(c)

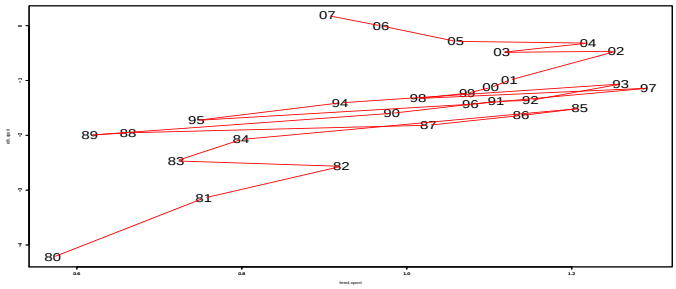
Figure 5: Plots of sub-compositions with respect to ethanol. Panel (a) is the log-ratios of residual food vs. feed; panel (b) is export vs. feed; panel (c) is exports vs. residual food. The graphs illustrate the competition or dependency among the remaining constituents after the residual or *fixed effect* of corn allocated to ethanol production has been account for (in each year). Labels are abbreviated years. The rates of decrease in the log ratios at each pair, from 1980-2007, are .89, 1.08 and 1.21, in order.



(a)



(b)



(c)

Figure 6: Plots of sub-compositions with respect to exports. Panel (a) is the log-ratios of residual food vs. ethanol; panel (b) is residual food vs. feed stocks; panel (c) is ethanol vs. feed stocks. The graphs illustrate the competition or dependency among the remaining constituents after the residual or *fixed effect* of corn allocated to exports production has been accounted for (in each year). Labels are abbreviated years.

furnishes the conditional dependency with respect to the residual component. The

(highly) significant values of distance from independence for the subcompositions with respect to ethanol, exports and feed stocks indicate that the values of these components are strongly associated. The existence of competition among these components, in relative fraction of corn yield, is an interpretation. In fact, the significant dependence in the subcomposition with respect to ethanol suggests strong competition among the remaining constituents once the ethanol fraction is accounted for. Conversely, the observed value of the dependency statistic for the subcomposition with respect to the residual food stocks is insignificant. This is a possible indication that the food stock allocations are not affected by competition among the other components.

Competition among Constituents of Corn Yield

The panels in Figure 5 are scatterplots of the log-ratios of the remaining constituents of corn-yield modulo the ethanol fraction, for every year. The illustrations suggest strong competition among the remaining constituents after the *fixed effect* of ethanol is removed: each fraction, labeled by year within each panel, decreases almost monotonically from 1980 to 2007.

The log-ratio pairs in each panel decrease in each panel. The allotments of corn yield to residual food, feed stocks and exports are much greater than to that of ethanol in 1980; by 2004 the fraction to ethanol is greater than to residual food — by 2006 it is greater than to feed stocks and nearly equal to exports.

These plots are evidence of a crowding effect, or competition among the remaining constituents of corn-yield. These effects, while perhaps monotone at each pair (residual food vs. feed, exports vs. feed stocks, exports vs. residual food) are non-constant. The rates of *decrease* — the slopes, say — of the pairwise log-ratios, from 1980-2007, are .89, 1.08 and 1.21, in order. These can be loosely interpreted as the magnitudes of the competition among the pairwise constituents. The observed distance statistic — illustrated in panel (a) of Figure 4 — suggests overall competition with respect to ethanol is highly significant.

Interpreting competition among the constituents when ethanol is included as a *variable* effect is less straightforward. Figure 6, for example, is an illustration of the log-ratio pairs (residual food vs. ethanol, residual food vs. feed stocks, and ethanol vs. feed stocks) for the subcomposition with respect to exports. The panels in Figure 6 suggest an overall trend of associated and decreasing log-ratios over time: the line connecting each data year, in order, is turbulently non-monotonic. The rates of *increase* of the pairwise log-ratios over 1980-2007 are 4.27, 1.71 and 0.17

respectively. The observed distance statistic for this subcomposition — with respect to exports — is illustrated in panel (d) of Figure 4. The significant distance from independence and increasing 'slope' of the pairwise log-ratios suggests that the constituents of corn yield, modulo exports, are occupying an increasing and greater share of overall corn production.

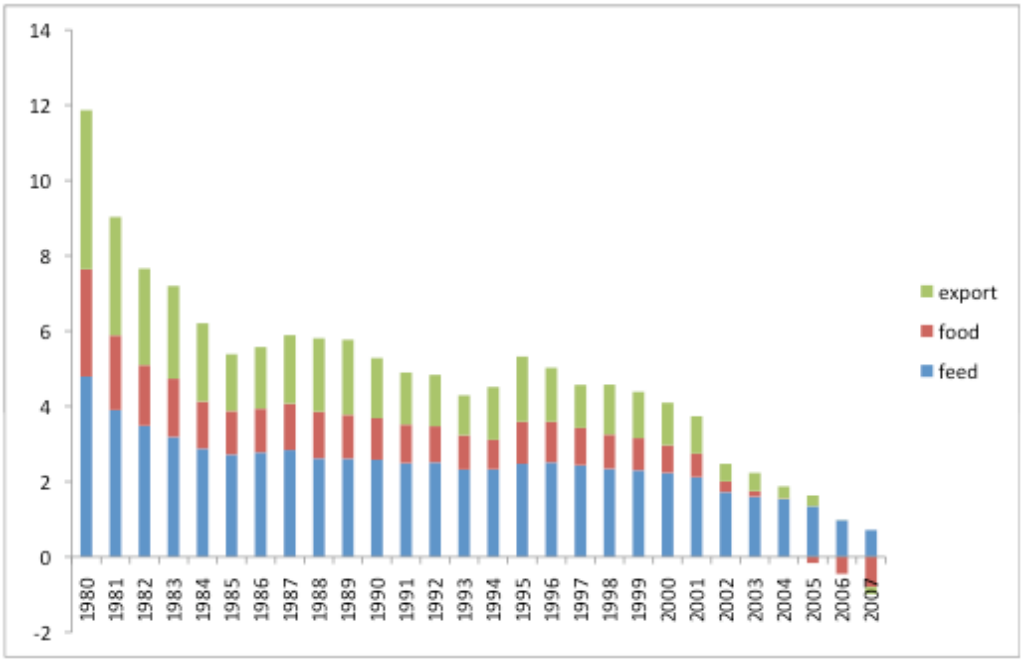


Figure 7: Log-ratios of subcomposition with respect to ethanol, food in red. The lengths within the bars are the yearly values of v_{j_m}; the by-year log ratio of the constituent fraction to with respect to ethanol. Positive values indicate amounts greater than ethanol, negative values indicate smaller values. The graph illustrates the effect of ethanol production on the remaining constituents of corn yield: the increase in ethanol production appears to be largely compensated for by the decrease in export and food production until year 2000. After 2000, ethanol production crowds out feed stocks as well; from 2005 ethanol production is greater than that for food corn.

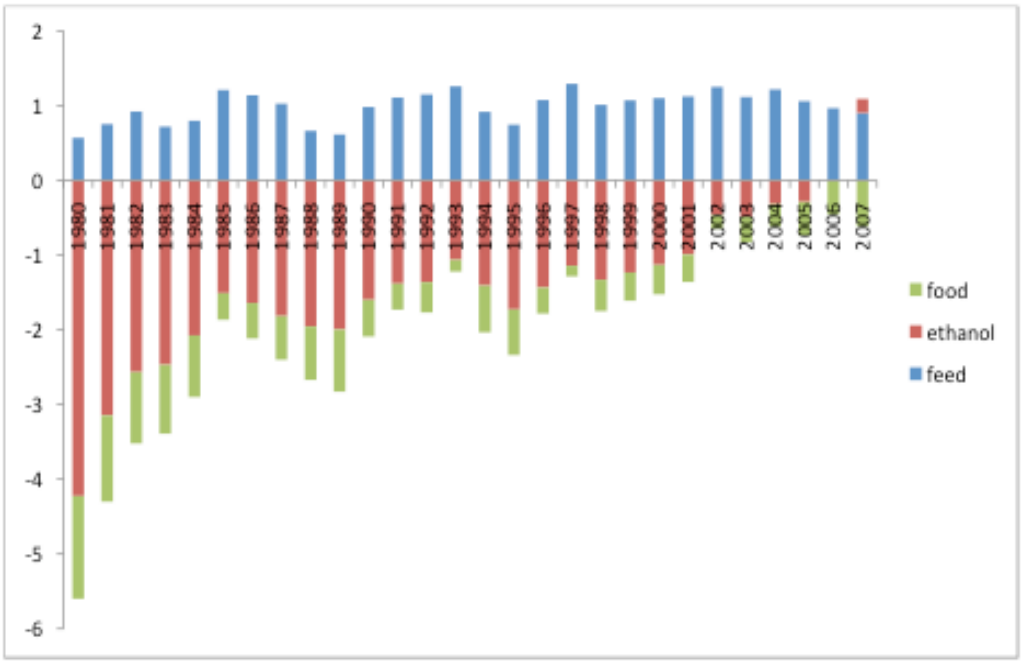


Figure 8: Log-ratios of subcomposition with respect to exports, ethanol in red. The lengths within the bars are the yearly values of v_{j_m}; the by-year log ratio of the constituent fraction to with respect to exports. Positive values indicate amounts greater than exports, negative values indicate smaller values. The graph illustrates the effect of corn exports on the remaining constituents of corn yield: the increase in ethanol production is apparent. In 2005 and 2006, ethanol production is greater than corn exports.

Discussion

The Dirichlet distribution is easy to simulate from directly or indirectly (see [15] and Appendix 2-3). We fit the Dirichlet to the margins of the composition and generate replicates from this fit; we prefer to simulate once and generate the m log-ratios from the replicates. This is a motion towards investigating complete subcompositional

dependence: these log-ratios can be generated for all subsets.

The alternative is to simulate from the Dirichlet margins 2^m times for all subcompositions, or fit the Liouville directly. Either method requires more complex computations: numeric or probabilistic integration to estimate parameters. Exploiting simulations here — i.e. randomly generating the distributions for the statistics of distance from independence for each of the subcompositions — accounts for sampling error in the data (estimation of the parameters for the *marginal* Dirichlets) and expands the class from which the replicates are drawn (beyond the *joint* Dirichlet).

We calculate the distance from independence via a norm on the probability integral transformed data, via the copula. The resultant distance is invariant to increasing transformations, like the log-ratio, equivalent on either the marginally Dirichlet replicates or their transformed copies, and semi-parametric. This method is preferable to tests of independence via null-correlation (*à la* Aitchison, see[12]) for multivariate data in high dimensions: dependency in this setup is not restricted to an elliptical shape.

The dependency statistics here treat time as unordered; as well, the illustrations of pairwise competition via the log-ratio scatterplots do not model time-dependent variability beyond inspection. The labels in Figures 3, 5, 6 suggest temporally sensitive dependency patterns. This is an important point of departure for future investigation. Characterizing competition among the outputs of corn yield will be extremely important as corn producers approach naturally constrained production ceilings.

Appendix

Appendix-1 Compositional Data and Subcompositions

Aitchison also defines $y_{k+1} = 1 - \sum_j^k y_j$; in this example $\sum_j^k y_j = 1$ since the total corn yield is just the sum of the constituents. Thusly, here, $(y_1, ..., y_{k-1}) \in \mathbb{S}_{k-1}$ — the $k-1$-dimensional simplex — versus the Aitchison method where $\mathbf{y} \in \mathbb{S}_k$.

In the original notation v_j is the log of the relative fraction of constituent j to the residual component of the basis, which is in the augmented simplex $\mathbb{S}_k^* = \{\mathbf{y}, y_{k+1}\} = \{y : y_j, (j = 1.., k+1), \sum_j y_j = 1\}$: $\mathbb{S}_k$ and $\mathbb{S}_k^*$ generate the same equivalence classes; the augmentation $\mathbb{S}_k^*$ is overdetermined. Here, however, the conditional distributions of the components are not assumed equivalent — in fact the conditional dependence appears to vary by choice of 'residual' m (see Figure 3), and merits component-wise

inspection. This augmentation is, in fact, heuristically similar to *subcompositional independence*, introduced in [14]. *Complete subcompositional independence* is independence among each of the $2^k - 1$ subsets of the composition.

The log-ratio transformation, $v(\cdot)$, maps the k-dimensional simplex ($\mathbb{S}^k$) to the k-dimensional real plane ($\mathbb{R}^k$); the logistic function ($y_j = \frac{e^{v_j}}{1+\sum_j v_j}$) is the inverse function. Thus the proportions, on the simplex, are mapped to the real hyperplane.

Appendix-2 Dependency under Log-Ratio Transform

The variance of a composition under a log-transform is $\sigma_{jj} = \omega_j + \omega_{k+1}$ and the covariances are $\sigma_{jl} = \omega_{k+1}$ as in [12].

A test of independence in this setting is with respect to a null hypothesis where $\Sigma_{\mathbf{v}}$ — the covariance matrix of the transformed composition — is constrained to the positive orthant and proportional to the units of the residual component. Aitchison proposes a Wald type likelihood ratio test, where the test statistic is iteratively estimated due to the constraints on the support of the parameter space ($\Sigma_{\mathbf{v}} \geq \mathbf{0}$ and proportional to the choice of y_{k+1}).

In contrast with Aitchison's likelihood ratio test for dependency in the composition, on the data transformed to $\mathbb{R}^k$, Rayens' fits a Liouville distribution (a generalization of the familiar Dirichlet distributions for proportions: see Equations 4 and 6) to Dirichlet marginals by choice of dependency function g to the data on $\mathbb{S}^k$. Both approaches are iterative: the parameters of both the Liouville and Aitchison's constrained log-normal must be estimated numerically (see [13] and [12] for illustration).

While Aitchison's use of the well-known log-normal distribution leads directly to independence testing via Wald's test, Rayens' approach for the Liouville distribution does not suggest a natural procedure for independence declaration. In fact, any reasonable procedure must restrict h within a class (linear, say) and test on introduced hyperparameters.

We choose to estimate dependency on log-ratio transformed data *without* assuming a log-normal distribution. We estimate dependency in the composition via replicates drawn from the Liouville family of distributions. In the Liouville family h is an additional parameter of interest for estimation — the choice of h governs the admissible dependence structure for $\mathbf{y}$.

Our approach is to approximate the broader Liouville class by introducing additional in marginally Dirichlet replicates.

- The estimates $\hat{\alpha} = (\hat{\alpha}_1, ..., \hat{\alpha}_k)$ of $\alpha = (\alpha_1, ..., \alpha_k)$ are the sample means of the composition data $\mathbf{y}$; the sampling error is order $n^{-1/2}$.
- Marginal Dirichlets can be generated directly from a version of Equation 4, or from any positive distributions with a sum constraint. Gamma and Beta distributions are candidates: [15] demonstrates a hierarchical approach with hyperparameters.
- To approximate Liouville replicates:
  (i) For/at each set of replicates $\mathbf{y}^{\hat{\alpha},t}$, $t = 1, ...T$, pick at random $1 \leq j, j' \leq k$
  (ii) Pick random $\epsilon \in [0, 1]$
  (iii) Reassign $y_j = y_j + \epsilon$ and $y_{j'} = y_{j'} - \epsilon$

The neutrality of the Dirichlet is a weaker independence than the full subcompositional independence available in the generalized Liouville family ([13]). Our procedure, modulo the randomization mechanism for ϵ, introduces broader dependency in marginal Dirichlets without resorting to direct fitting of one Liouville distribution or another (choice of h). As $T \to \infty$ the replicates will yield — infinitely often — non-neutral replicates.

While the logistic-normal class is closed to subcompositions, the dependency within $\mathbf{v}_{\substack{j \\ m}}$ is not invariant to choice of m. We exploit the log-ratio transformation to easily investigate subcompositional dependency (here for subcompositions of dimension 3), not to test independence via the logistic normal distribution. This operates in context with the generalized Liouville family, where the choice of h (see eq. (6)) is akin to choosing the residual of the simplex.

Appendix-3 KS distance from Independence as Measure of Association

The *Kolmorogov-Smirnov* distance is:

$$D_n = \sup_t |F_n(t) - F(t)| \tag{7}$$

Asymptotic convergence of this distance to a Chi-Squared distribution under the hypothesis $\mathbf{v}$ are generated with common distribution F is a well-known result [19]. A multivariate version of this statistic is

$$D_{n,k} = \sup_{\mathbf{t}} |F_n(\mathbf{t}) - F(\mathbf{t})| \tag{8}$$

For $\mathbf{t} = (t_1, t_2)$ the distance is a probability measure on Kendall's distributions; Chi-Square convergence does not hold [20]. Similar — multivariate — versions of the Kolmorogov-Smirnov distance are investigated in ([24]) and ([25]). The first paper relies upon Rosenblatt's iterative transformation of the data ([26]) by conditionally independent cumulative distributions and the second requires Gaussian data; neither paper offers distributional results for $k > 2$.

Let $\mathbf{u} = (u_1, ..., u_k)$, where each $u_j = F_j(v_j)$, F_j the distribution function for v_j. Let the joint distribution for $\mathbf{v}$ be $F(\mathbf{v})$. The *copula* for $\mathbf{u}$ is

$$C(\mathbf{u}) = F(F_1(v_1), ..., F_k(v_k)) \tag{9}$$

the mapping from $\mathbb{I}^k$ to $\mathbb{I}$; the shape of the joint distribution F fixed to the unit hypercube $\mathbb{I}^k$ [21]. The Kolmogorov-Smirnov statistic (distance) for multivariate independence can be written:

$$D^{\Pi}_{n,k} = \sup_{\mathbf{t}} |F_n(\mathbf{t}) - \prod_j F_j(t_j)|. \tag{10}$$

Using (9), this is, now for $\mathbf{v}$

$$D^{\Pi}_{n,k} = \sup_{\mathbf{u}} |C_n(\mathbf{u}) - \prod_j u_j|. \tag{11}$$

by definition of multivariate independence, with $C_n(\cdot)$ a multivariate version of the *empirical copula*:

$$C_n(\mathbf{u}) = \frac{\#\{\mathbf{t} \mid t_1 \leq F_1^{-1}(u_1), ..., t_k \leq F_k^{-1}(u_k)\}}{n} \tag{12}$$

where $\#\{\cdot\}$ is cardinality and F^{-1} is the inverse distribution function (see [22] and [23]). This statistic is the $\mathcal{L}_\infty$ distance between the empirical joint and independent distributions with equivalent margins. Our procedure:

- Fit a Dirichlet distribution (i.e. estimate $\hat{\alpha} = (\hat{\alpha}_1, ..., \hat{\alpha}_k)$ for $\alpha = (\alpha_1, ..., \alpha_k)$) to the composition data $\mathbf{y}$.

- Generate T Dirichlet replicates, parameter $\hat{\alpha}$, each of dimension $n \times k$: $(\mathbf{y}^{\hat{\alpha},1}, ..., \mathbf{y}^{\alpha,T})$.
- Compute $m = 1...k$ versions of Aitchison's log-ratios on the replicates: $\mathbf{v}_m^{\hat{\alpha},1} ..., \mathbf{v}_m^{\hat{\alpha},T}$
- For $m = 1..k$ compute $D_{\underset{m}{n,k}}^{\Pi,1}, ..., D_{\underset{m}{n,k}}^{\Pi,T}$ of

$$D_{n,k}^{\Pi} = \sup_{\mathbf{t}} |C_n(\mathbf{u}^{\alpha}) - \prod_j u_j^{\hat{\alpha}_j}|. \tag{13}$$

where $u^{\hat{\alpha}_j} = F_{n,j}(v_j^{\hat{\alpha}_j})$ as a semi-parametric version of equation (10).

This yields a distribution for the statistic under an independence hypothesis among the compositions — a direct result of the neutrality of the Dirichlet distribution. Moreover, the m versions of the statistic, $D_{\underset{m}{n,k}}^{\Pi,1}, ..., D_{\underset{m}{n,k}}^{\Pi,T}$, are proxies for tests of complete subcompositional independence. These statistics are calculated on the log-ratios ($\mathbf{v}$) of the replicates, and not the Dirichlet draws for this reason: picking each of m components to serve as 'residual' in via the basis ($\mathbf{x}$) or composition ($\mathbf{y}$) requires m estimates of α and m-fold random draws.

More recent work ([27] and [28]) relies upon distributional specification of the copula (Kendall's type distributions, see [20] and [29]) and the resultant transformed processes do not yield distributions for the multivariate Kolmorogov-Smirnov distance, do not specifically address dependency in the compositional data setting, and illustrate only $k = 2, 2$ and 5.

In short we offer a test of dependence via the $\mathcal{L}_\infty$ norm: this is the Kolmogorov-Smirnov distance. This test of dependence is semi-parametric in that the replicates are generated via $\hat{\alpha}$ but the distance from independence is calculated via the empirical probability integral transform or the multivariate order statistics using the empirical copula. The distance statistic $D_{n,k}^{\Pi}$ is Euclidean, but on the probability measure space — i.e. modulo the appropriate and flexible choice for the fixed marginals of $\mathbf{v}$.

We prefer this test for multivariate composition data, especially as k increases. For large k the support of elliptical distributions (such as the normal and log normal) — the setting for much analysis of compositional data — migrates into the extreme tails.

References

[1] Energy Independence and Security Act of 2007. Publication 110-140, 110th Congress (2007).

[2] Connor, R and Mosimann, J. Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution. *Journal of the American Statistical Association* **64**. 325. pp 194-206 (1969).

[3] Tilman, D., Socolow, R. et al. Beneficial Biofuels — The Food, Energy, and Environment Trilemma. *Science*, **325**. pp 270-271. (2009).

[4] Aitchison, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society, Series B.* **44**, 2. pp. 139-177 (1982).

[5] Renewable Fuels Association. "Historic U.S. Fuel Ethanol Production." (2009).

[6] Food and Energy Security Act of 2007: Report of the Committee on Agriculture, Nutrition, and Forestry. Publication 110-220. 110th Congress (2007).

[7] Aitchison, J. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltf. London (1986).

[8] Aitchison, J. The one-hour course in compositional data analysis or compositional data analysis is simple. In: Pawlowsky-Glahn, Proceedings of IAMG'97 — The third annual conference of the International Association for Mathematical Geology. Vol 1, 2 and addendum. International Center for Numerical Methods in Engineering (CIMNE), Barcelona. pp 3-35. (1997)

[9] Barcelo-Vidal et al. Mathematical Foundations of Compositional Data Analysis. In Ross, G. Proceedings of IAMG'01 - the sixth annual conference of the International Association for Mathematical Geology. p. 20. (2001)

[10] Pawlowsky-Glahn, V. and Mateu-Figueras, G. The Statistical Analysis on Coordinates in Consgtrained Spaces, in International Statistical Institute. 55th session of the International Statistical Institute. April, Sydney Convention & Exhibition Centre, Sydney, Australia. (2005)

[11] Aitchison, J et al. Logratio Analysis and Compositional Distance. *Mathematical Geology*, **32**, 3, pp 271-275. (2000)

[12] Aitchison, J. A New Approach to Null Correlations of Proportions. *Mathematical Geology* **13**, 2, pp 175-189. (1981)

[13] Rayens, W. & Srinivasan, C. Dependence Properties of Generalized Liouville Distributions on the Simplex. *Journal of the American Statistical Association*, **89**, 428, pp. 1465-1470. (1994).

[14] Aitchison, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society.*, Ser. B, **44**, pp. 139-177. (1982)

[15] Gelman, A., et al. Bayesian Data Analysis, Second Edition. Chapman and Hall. (2004)

[16] James, I & Mosimann, J. A New Characterization of the Dirichlet Distribution through Neutrality. *The Annals of Statistics*, **8**, 1, pp. 183-189. (1980)

[17] Aitchison, J. Distributions on the Simplex for the Analysis of Neutrality. In: Statistical distributions in scientific work proceedings of the NATO Advanced Study Institute held at the Universita degli Studi di Trieste, Trieste. (1981)

[18] Gupta, R. & Richards, D. The History of the Dirichlet and Liouville Distributions. *International Statistical Review*, **69**, 3, pp. 433-446. (2001)

[19] Williams, D. Weighing the Odds: A Course in Probability and Statistics. Cambridge University Press. Cambridge. (2001)

[20] Nelsen, R. et al. Kendall distribution functions. *Statistics & Probability Letters.*, **65**, pp. 263-268. (2003)

[21] Nelsen, R. An Introduction to Copulas. Springer. (1999)

[22] Deheuvels, P. La fonction de dpendance empirique et ses proprits: un test non paramtrique d'indpendance. *Acad. Roy. Belg. Bull. Cl. Sci.*, **65**, pp. 274-292. (1979)

[23] Wolff, E.F. N-Dimensional Measures of Dependence. *Stochastica*, **4**. (1980)

[24] Justel, A., Pea, D. & Zamar, R. A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics and Probability Letters*, **35**, pp. 251-259. (1997)

[25] Fasano, G & Franceschini, A. A multivariate version of the Kolmorogov-Smirnov test. *Notes of the Royal Astronomical Society*, **225**, pp. 155-170. (1987)

[26] Rosenblatt, M. Remarks on a Multivariate Transformation. *The Annals of Mathematical Statistics*, **23**, pp. 470-472. (1952)

[27] Genest, C. & Rmilliard, B. Test of Independence and Randomness Based on the Empirical Copula Process. *Sociedad de Estadstica e Investigacin Operativa*, **13**, pp. 335-269. (2004)

[28] Genest, C., Quessy, J. & Rmillard, B. Goodness-of-fit Procedures of Copula Models Based on the Probability Integral Transform. *Scandinavian Journal of Statistics*, **33**, pp. 337-366. (2006)

[29] Genest, C. & Rivest, L.P. On the multivariate probability integral transform *Statistics and Probability Letters*, **53**, pp. 391-399. (2001)

[30] Cleveland, W.S. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* **74**829836. (1979)

[31] Aitchison, J. & Shen, S.M. Logistic-Normal distributions. *Biometrika*, **67**, 2. pp. 261-272. (1980)

[32] USDA, Economic Research Service, *Feed Grains Database, custom queries*,