Statistical Evaluation of the Effect of Ethanol in US Corn Production: A Kolmorogov-Smirnov Test for Independence on a Constrained Sum via the Empirical Copula.

Kobi Abayomi, Dexin Luo and Valerie M. Thomas

School of Industrial and Systems Engineering (ISYE),Georgia Institute of Technology, Atlanta, GA,U.S.

November 22, 2010

ABSTRACT

We investigate the dependence between ethanol (biofuel) production and competing uses for corn via a case study of U.S. corn production from 1980 to 2009. Agricultural models are often used to estimate this effect of biofuel production on crop production and, correspondingly, land use change: these models can be validated with statistical analysis of dependency — or competition — among crop uses. Standard statistical tests for independence are inappropriate, however, because the uses of an agricultural crop are constrained: total crop use is always the sum of the constituents. Commonly used methods for testing independence in compositional data rely upon restrictive distributional assumptions: either the multivariate log-normal distribution or a particular choice from the family of Liouville distributions.

We develop a general method for these *compositional distributions* via the Kolmorgorov-Smirnov probability distance and the empirical copula. We apply this method to determine competition between ethanol feedstock production and other alternate products of corn. We find evidence of competition among the constituents of corn yield, particularly with respect to ethanol production.

Keywords: Ethanol production, Biofuel, Independence tests, Empirical copula, Kolmogorov-Smirnov distance.

1 Introduction

1.1 Ethanol for Biofuel

The Energy Independence and Security Act (EISA) of 2007 mandates an increase in ethanol production to 36 billion gallons per year by 2022.(110th US Congress, 2007a) In 2008 the United States produced 9 billion gallons of ethanol fuel — an increase of more than 6000 percent since 1980.(Renewable Fuels Association, 2009) At the same time, the total U.S. corn

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

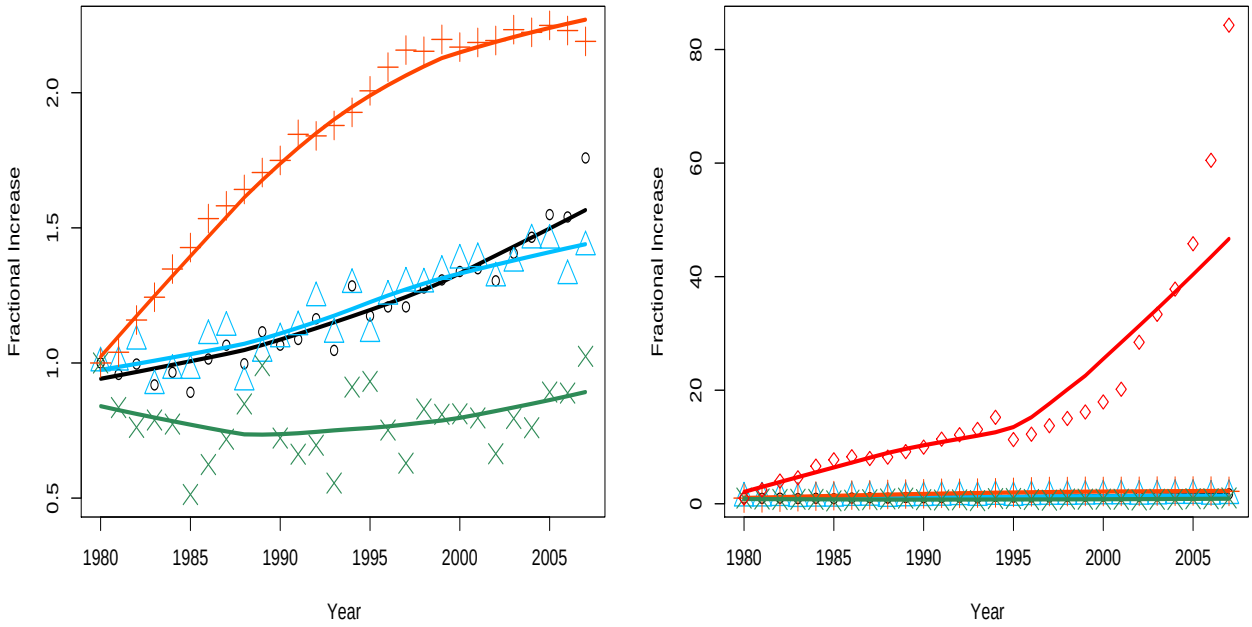


Figure 1: Fractional increase in corn production and constituents, 1980-2007, in color: Residual food, feed stocks, exports, ethanol, and total output. The smoothed curve suggests the trend. Yearly values are ratios to 1980 production of/by each constituent output. Total corn yield, food and feed stocks 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 (110th US Congress, 2007b), and addressed elsewhere (Tilman, Socolow, Foley, Hill, Larson, Lynd, Pacala, Reilly, Searchinger, Somerville and Williams, 2009). Currently, models of land use change from biofuel production either use basic assumptions about the extent to which biofuel production displaces other uses of the feedstock, or use agricultural economic models (Keeney and Hertel, 2009), (Agency, 2007), (Thompson W., 2009), (De La Torre Ugarte D. G., 2007). As biofuel production grows, there is potential to validate the assumptions of these models using data on crop production and biofuel production.

We develop a method for using constrained sum data to investigate how increased bioenergy production affects production of other agricultural crops. These phenomena can be examined from either a crop production perspective, as is done here, or from land use; these results are

parallel and mediated by yield. 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. The methodology can have wide applications and specific use in other settings: the influence of sugarcane-derived ethanol production on land use in Brazil, say; 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); the influence of biofuel production from non-food crops on food crop production.
Here, we use U.S. data on the use of corn for ethanol production to illustrate a method for statistical evaluation of how changes in crops and land use are affected by bioenergy development.

2 Dependence/Independence in Constrained Sum Data

2.1 Constituents of Corn Use: Compositional Data

The uses of corn considered in this paper include:

- Feed and residual use
- Exports
- Food and Industrial uses
- Ethanol

The category food and industrial uses include corn use for high fructose corn syrup (about 40 percent), glucose and dextrose (about 20 percent), starch (about 20 percent), alcohol for beverage and manufacturing (about 10 percent), and other (about 10 percent). The data are 30 years — from 1980 to 2009 — of the allocations of total U.S. corn production in units of mass (USDA, 2010). Preparatory illustrations of the fractional increases (1) and relative fractions (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 (Aitchison, 1997), (Barcelo-Vidal, Martn-Fernandez and Pawlowsky-Glahn, 2001), (Pawlowsky-Glahn and Mateu-Figueras, 2005), and especially (Aitchison, 1986)).

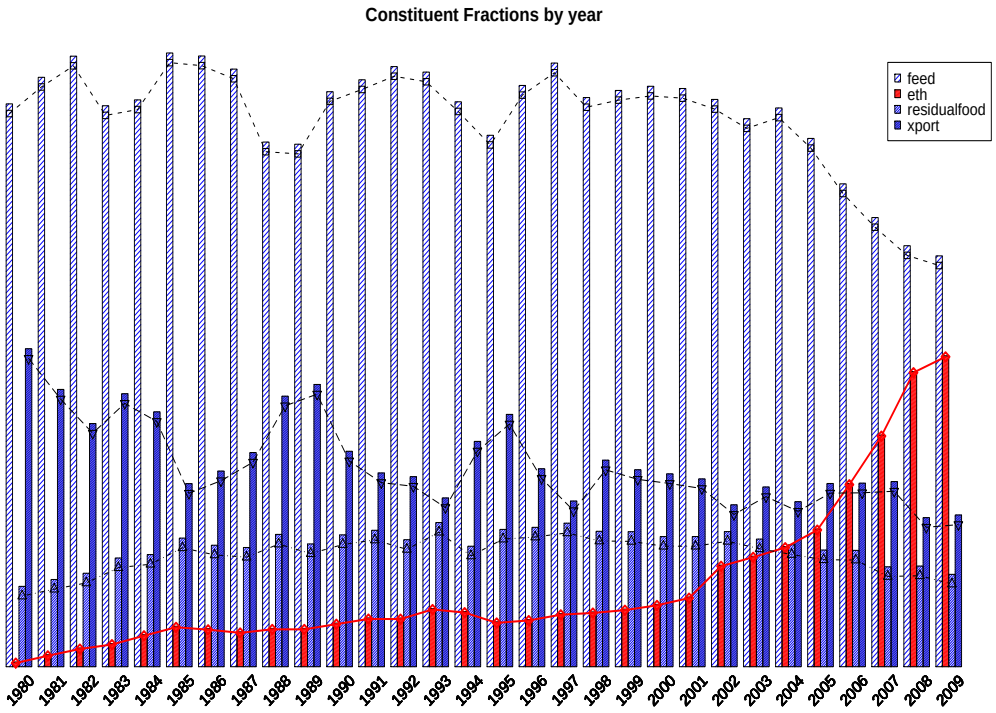


Figure 2: Joint distribution of fractional uses of corn, ethanol in red. The heights of the bars are the ratio of the constituent fraction to the total corn production (normalized to 1 for each year). The graph is a representation of a three-dimensional positive simplex, a four dimensional composition in Aitchison terminology: the heights are the joint distribution modulo the total production (and variation).(Aitchison, 1982) Dependence in this *compositional distribution* is a function of the statistic represented by this illustration.

Let

$$\mathbf{x} = (x_1, ..., x_k) \tag{2.1}$$

be a basis or open vector of positive quantities, $\mathbf{x} \in \mathbb{R}^{k+}$ — the k dimensional positive hyper-plane. In this example the positive quantities are the constituents of total corn production, in order (in units of mass): $\mathbf{x} = (x_{eth}, x_{food}, x_{feed}, x_{xport})$; corn to ethanol production, food and industrial, feed, and export.
Let

$$y_j = x_j / \sum_j^k x_j \tag{2.2}$$

with $\mathbf{y} = (y_1, ..., y_k)$ the vector of fractions; in the Aitchison (Aitchison, 1981b) terminology the composition of the basis $\mathbf{x}$. Here the y_j are the (relative to the total) fractions of ethanol, food and industrial, feed, and exports (mass) of corn illustrated in 2.

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 (Aitchison, 2000)). This characteristically constrains or wholly excludes standard methods and tests for multivariate independence, such as 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).

2.2 Distributional Models for Compositional Data

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 the resultant distribution. The log-ratio transformation is popular; tests of independence are on the constrained covariance matrix of the transformed data (see (Aitchison, 1981b)). 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 a broader class, i.e. the Liouville distribution (see (Connor and Mosimann, 1969) and (Rayens and Srinivasan, 1994), or (Barndorff-Nielsen and Jorgensen, 1991)).

2.2.1 Transform the simplex to the plane

A log-ratio transformation sets

$$v_{\underset{m}{j}} = log(\frac{y_j}{y_m}) = log\, y_j - log\, y_m \tag{2.3}$$

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.
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 , and merits component-wise inspection. This augmentation is, in fact, heuristically similar to

subcompositional independence, introduced in (Aitchison, 1982). *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.
Aitchison (Aitchison, 1981a) 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:

$$\Sigma_{\mathbf{v}} \propto diag(\omega_1, ..., \omega_k) + \omega_{k+1}, \tag{2.4}$$

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

2.2.2 Address the simplex directly

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 2.5 and 2.6) 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 (Rayens and Srinivasan, 1994) and (Aitchison, 1981b) for illustration).
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:

$$dF(\mathbf{y}) \propto (1 - \sum_j y_j)^{\alpha_{k+1}-1} \cdot \prod_j y_j^{\alpha_j - 1} \tag{2.5}$$

with $\alpha_0 \equiv 0$ (see (Gupta and Richards, 2001)), and $dF(\cdot)$ the density.
Connor and Mossiman characterize a vector of proportions $(y_1, ..., y_{k+1})$ as Dirichlet distributed (Connor and Mosimann, 1969), (James and Mosimann, 1980). The Dirichlet class, however, is only suitable for modeling random vectors that have substantial independence. This dependency characterization is insufficient for non-independent, or non-*completely neutral* proportions (see (Aitchison, 1981b) and as well (Rayens and Srinivasan, 1994)): compositional data that are positively associated cannot be modeled via the Dirichlet distribution which precludes an immediate (parametric) test of independence.

Rayens and Srinivasan propose the generalized Liouville distribution — a generalization of the Dirichlet distribution — as a richer model for compositional data under dependence (Rayens and Srinivasan, 1994).
A Liouville distribution is

$$dF \propto h(\sum_j y_j) \prod y_j^{\alpha_j - 1} \tag{2.6}$$

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 $\alpha_{k+1} = 1$.

3 Methodology: Independence via Probability Distance

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 a simple version of 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 $\mathbf{y}$.
We, in an alloy and extension of Aitchison's (impose a log-normal distribution after transformation) and Rayens' (fit a model directly to the simplex) methods, test for independence using distance on probability measures. This can be seen as blend of the 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 (KS) probability measure of distance. This method allows us quickly construct sub-compositions of the data, address independence testing beyond neutrality, without fitting a particular (non-neutral) Liouville distribution.
In the first step, we apply Aitchison's log-ratio transform m-fold times on the observed data, holding each component of the data as the residual singly. Using this transformation, we can find the empirical joint distributions of the compositions with one constituent as 'residual', i.e. the conditional joint dependence among the constituents, with respect to the 'residual' constituent. We calculate the empirical copula cumulative distribution function (CDF) $F_n(\mathbf{t})$ and the CDF for each margin, $F_j(t_j)$, and use it to calculate the KS or $\mathcal{L}^1$ metric as a distance from independence.
We then exploit the special role of the Dirichlet as the *neutral* distribution to generate marginally equivalent multivariate replicates. We can then generate T, say, replicates setting $\alpha = \bar{\mathbf{y}} = (\bar{y_1}, ...\bar{y_k})$ as the coefficients of Dirichlet distribution, where $\bar{y_i}$ is the mean of vector y_i. These replicates offer a simulated distribution for the distance from independence in a non-neutral setting: these k-dimensional random, sum-constrained, positive replicates are

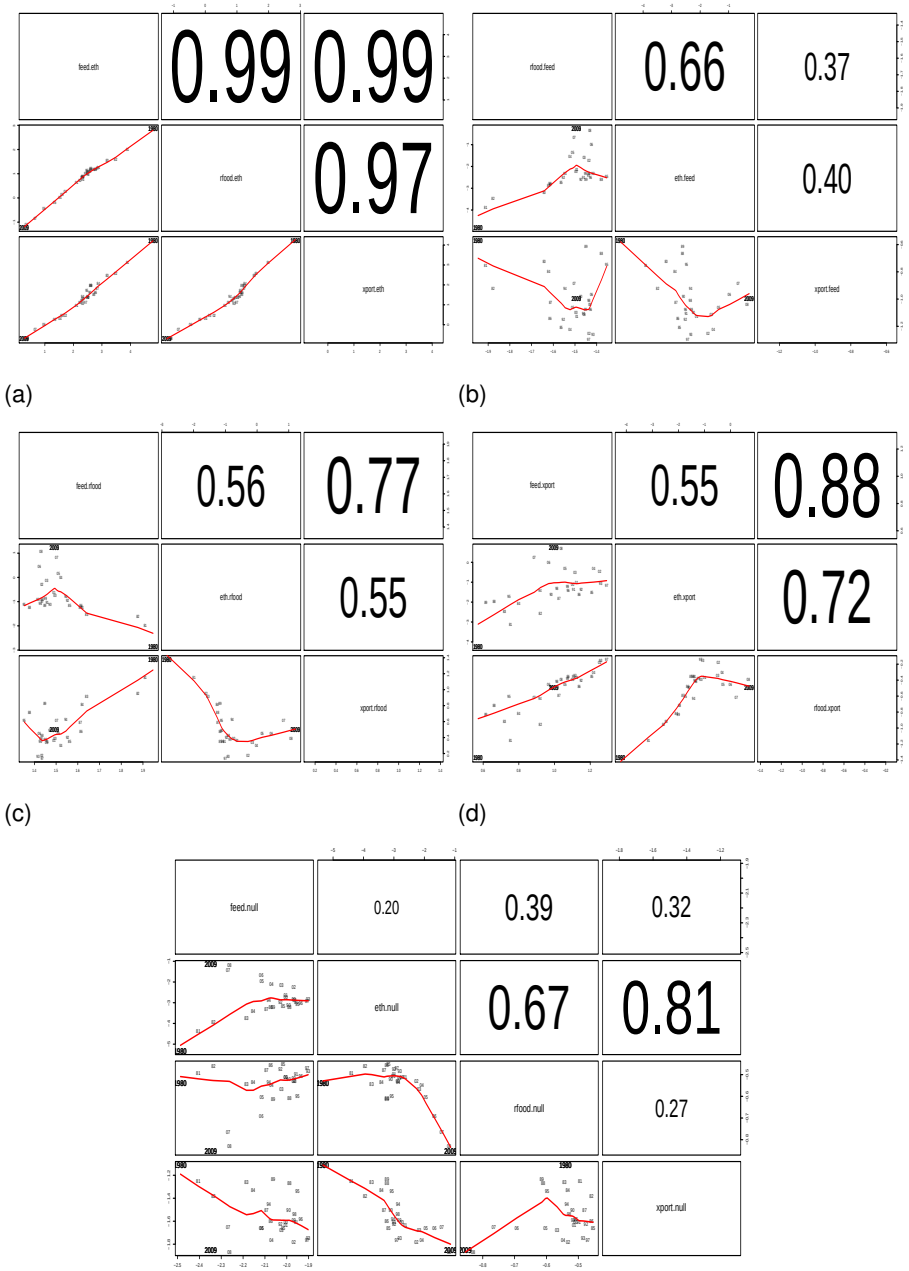


Figure 3: Scatterplots of log-ratios of constituents of total corn production (v_{j_m}), labeled by years. $m = (eth,\ feed,\ food,\ export,\ total)$ in plots (a)-(e), in order. Observed log-ratios are labeled by year. Loess smoothed curve in red. 1980 and 2009 are highlighted darker. Correlation coefficients (for linear pairwise dependence) are in each opposite panel - note pairwise correlations are inflated for compositional data. The dependence among the log-ratios appears to vary by choice of 'residual' m.

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 *distance measures* — computed on these replicates — to envelop dependency beyond mere neutrality The last step is to use the simulated results for KS distances of all the multivariate replicates and the KS distance calculated from the observed data to calculate the p-value, or probability

of dependency, sub-composition by sub-composition.
This technique 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 Rayens and Srinivasan (Rayens and Srinivasan, 1994).

3.1 Our Scheme

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 2.5, or from any positive distributions with a sum constraint. Gamma and Beta distributions are candidates: (Gelman, 2004) 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 ((Rayens and Srinivasan, 1994)). 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 $\underset{m}{\mathbf{v}_j}$ 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. (2.6)) is akin to choosing the residual of the simplex.

3.2 KS distance from Independence as Measure of Association

The *Kolmorogov-Smirnov* distance is:

$$D_n = \sup_t |F_n(t) - F(t)| \tag{3.1}$$

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 (Williams, 2001). A multivariate version of this statistic is

$$D_{n,k} = \sup_{\mathbf{t}} |F_n(\mathbf{t}) - F(\mathbf{t})| \tag{3.2}$$

For $\mathbf{t} = (t_1, t_2)$ the distance is a probability measure on Kendall's distributions; Chi-Square convergence does not hold (Nelsen, 2003). Similar — multivariate — versions of the Kolmorogov-Smirnov distance are investigated in ((Justel A, 1997)) and ((Fasano and Franceschini, 1987)). The first paper relies upon Rosenblatt's iterative transformation of the data ((Rosenblatt, 1952)) 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

$$C(\mathbf{u}) = F(F_1(v_1), ..., F_k(v_k)) \tag{3.3}$$

the mapping from $\mathbb{I}^k$ to $\mathbb{I}$; the shape of the joint distribution F fixed to the unit hypercube $\mathbb{I}^k$ (Nelsen, 1999). The Kolmogorov-Smirnov statistic (distance) for multivariate independence can be written:

$$D^{\Pi}_{n,k} = \sup_{\mathbf{t}} |F_n(\mathbf{t}) - \prod_j F_j(t_j)|. \tag{3.4}$$

Using (3.3), this is, now for $\mathbf{v}$

$$D^{\Pi}_{n,k} = \sup_{\mathbf{u}} |C_n(\mathbf{u}) - \prod_j u_j|. \tag{3.5}$$

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

$$C_n(\mathbf{u}) = \frac{\#\{\mathbf{t} \mid t_1 \leq F_1^{-1}(u_1), ..., t_k \leq F_k^{-1}(u_k)\}}{n} \tag{3.6}$$

where $\#\{\cdot\}$ is cardinality and F^{-1} is the inverse distribution function (see (Deheuvels, 1979) and (Wolff, 1980)). 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 $\mathbf{y}$.

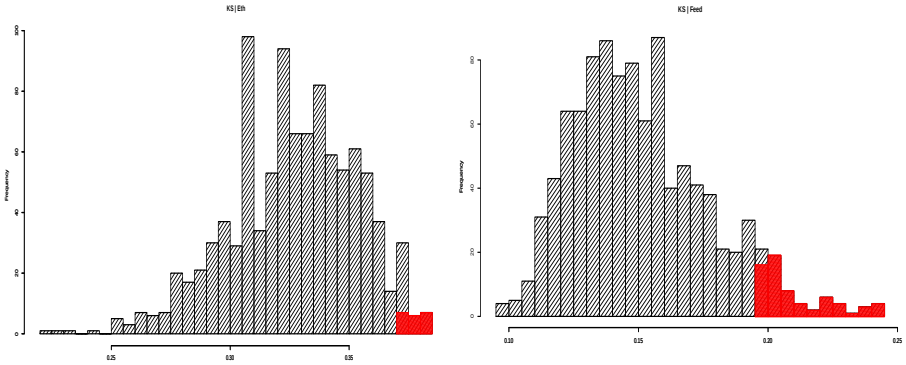
- Generate T Dirichlet replicates, parameter $\hat{\alpha}$, each of dimension $n \times k$: $(\mathbf{y}^{\hat{\alpha},1}, ..., \mathbf{y}^{\alpha,T})$.
- Compute $m = 1...k$ versions of Aitchison's log-ratios on the replicates: $\mathbf{v}_m^{\hat{\alpha},1}..., \mathbf{v}_m^{\hat{\alpha},T}$
- For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

$$D_{n,k}^{\Pi} = \sup_{\mathbf{t}} |C_n(\mathbf{u}^{\alpha}) - \prod_j u_j^{\hat{\alpha}_j}|. \tag{3.7}$$

where $u^{\hat{\alpha}_j} = F_{n,j}(v_j^{\hat{\alpha}_j})$ as a semi-parametric version of equation (3.4).
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, $D_{\substack{n,k \\ m}}^{\Pi,1}, ..., D_{\substack{n,k \\ m}}^{\Pi,T}$, 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 ((Genest and Rmillard, 2004) and (Genest, Quessy and Rmillard, 2006)) relies upon distributional specification of the copula (Kendall's type distributions, see (Nelsen, 2003) and (Genest and Rivest, 2001)) 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.

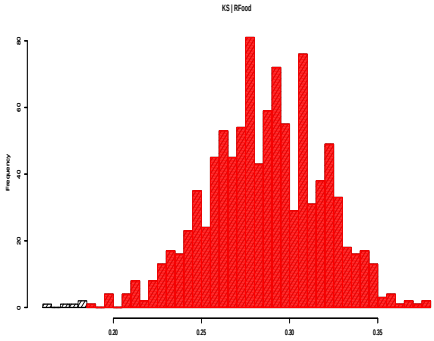
4 Results

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 (Cleveland, 1979)). 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.

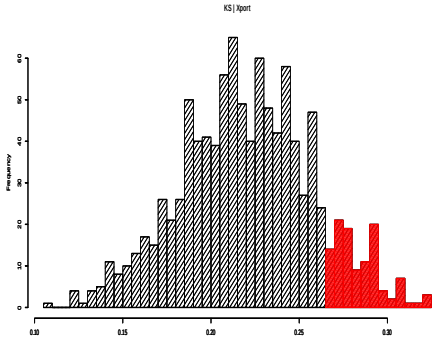


(a)

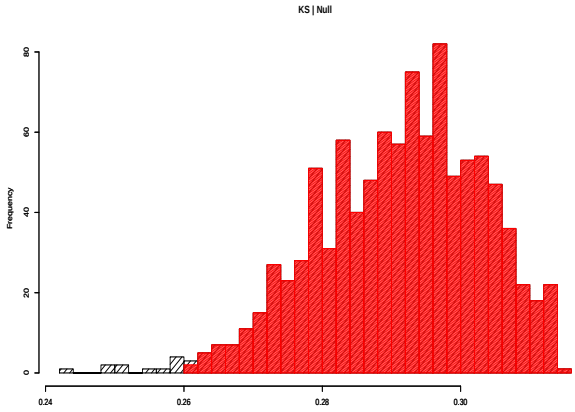
(b)



(c)

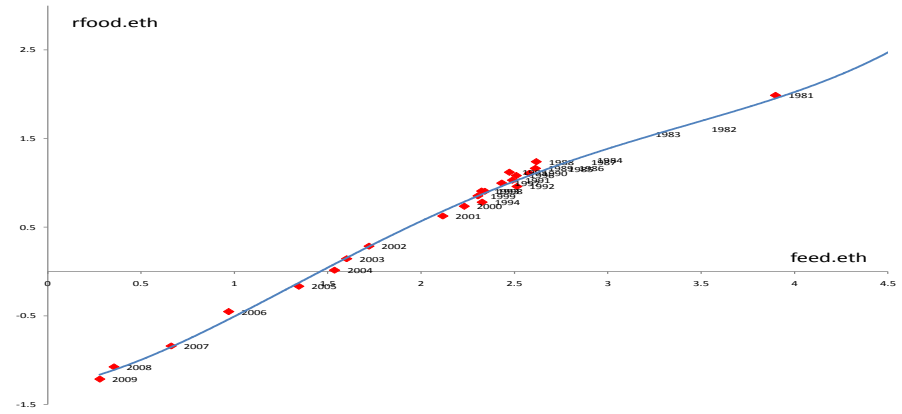


(d)

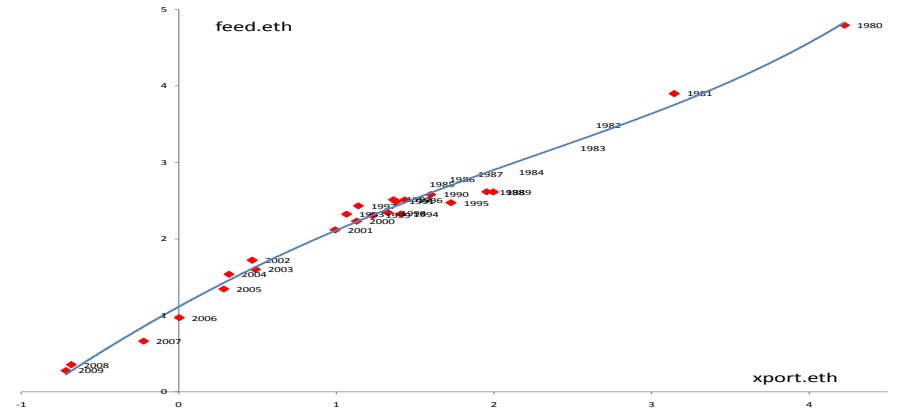


(e)

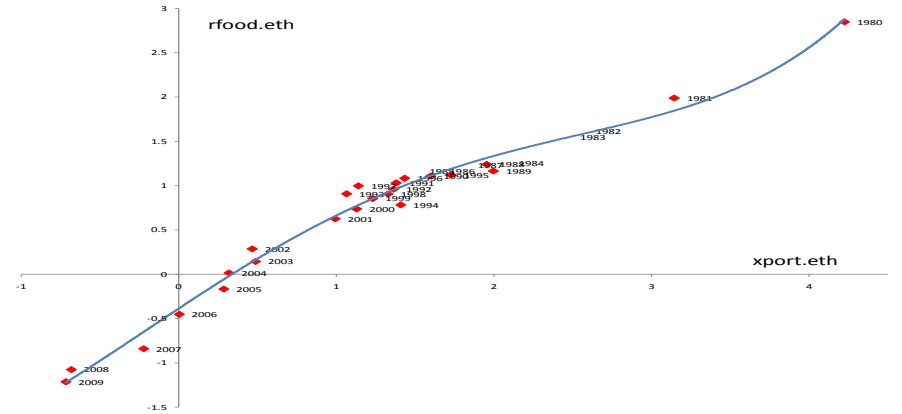
Figure 4: Histograms of the distributions of distances from independence-neutrality $(D^{\Pi,1}_{n,k_m}, ..., D^{\Pi,T}_{n,k_m})$, for $m = (eth,\ feed,\ foodandindustrial,\ export,\ total)$ in plots (a)-(e), in order. Here $n = 30$, $k = 3, 4$, $T = 1000$. The area in red are values above the observed 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, feed, and exports are far from independence: the observed p-values — 0.024, 0.067, 0.112, in order.



(a)



(b)



(c)

Figure 5: Plots of sub-compositions with respect to ethanol. Panel (a) is the log-ratios of food and industrial vs. feed; panel (b) is export vs. feed; panel (c) is exports vs. food and industrial. 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.

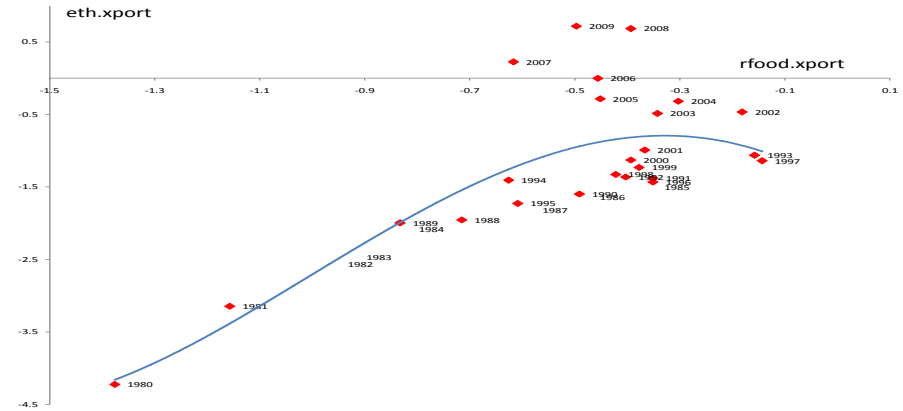
These data are the observed values of $\mathbf{v}_{\underset{m}{j}}$ — the log-ratios of the compositions.
We *do not fit* a logistic-normal distribution (Aitchison and Shen, 1980) to the data; in fact, we make no distributional assumption in the calculation of the distance from independence, beyond the utilization of random, marginal Dirichlets as replicates for the compositional data.

4.1 Statistical Dependency among Constituents of Corn Uses

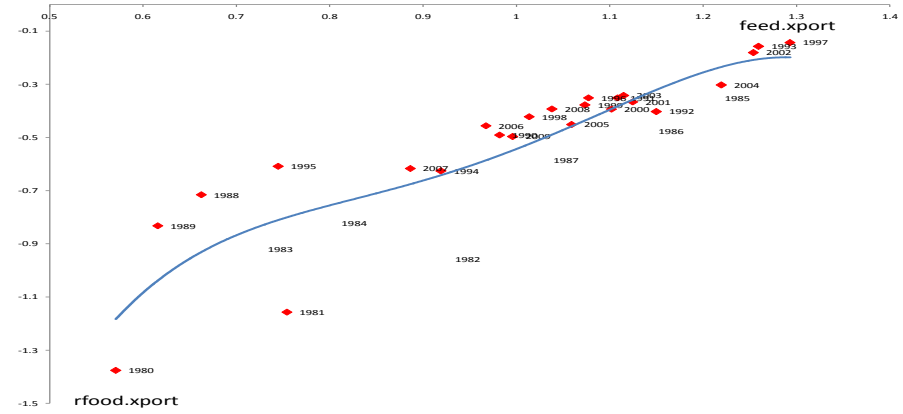
Figure 4 is the distributions of the statistic for the test of independence ($D^{\Pi}_{n,k}$); for the null composition — all of the constituents of corn uses together — and for each of the subcompositions. The observed statistic for each is highlighted by the leftmost border of the red shaded area: the shaded area shows the replicates that 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, feed and exported corn are highly significantly dependent, with p-values of $.024$, $.067$, and $.112$ respectively.
The remaining subcomposition — with respect to food and industrial — and the null composition have p-values for dependence of $.995$ and $.988$ which do not suggest dependence.
The distance statistics are scalar measures for the multivariate dependency within each (sub) composition. The joint distribution within each of the sub-compositions 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 food and industrial is insignificant. This is a possible indication that food and industrial allocations are not affected by competition among the other components.

4.2 Competition among Constituents of Corn Use

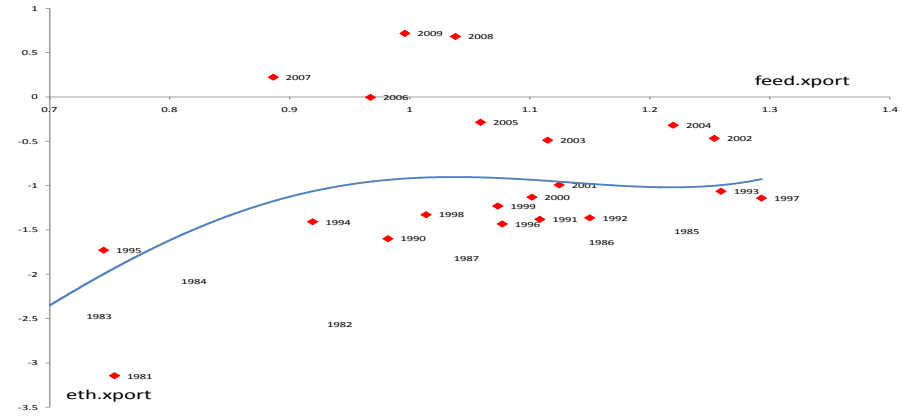
The panels in 5 are scatterplots of the log-ratios of the remaining constituents of corn-yield modulo the ethanol fraction, for every year. The illustrations suggest strong competition among the remaining constituents after the *fixed effect* of ethanol is removed: each fraction, labeled by year within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



(a)



(b)



(c)

Figure 6: Plots of sub-compositions with respect to exports. Panel (a) is the log-ratios of food and industrial vs. ethanol; panel (b) is 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.

The log-ratio pairs in each panel decrease in each panel. The allotments of corn to food and

industrial, feed stocks and exports are much greater than to ethanol in 1980; by 2004 the fraction to ethanol is greater than to food and industrial — by 2006 it is greater than to feed and nearly equal to exports.

These plots are evidence of a crowding effect, or competition among the remaining constituents of corn uses. These effects, while perhaps monotone at each pair (food and industrial vs. feed, exports vs. feed, exports vs. food and industrial) are non-constant. The rates of *decrease* — the slopes, say — of the pairwise log-ratios, from 1980-2007, are 0.89, 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 4 — suggests overall competition with respect to ethanol is highly significant.

Making use of this dependency finding, we can express how increased use of corn for ethanol affected the other uses of corn. Comparing the actual composition to what it would have been if the fractional distribution had remained constant, one might say that each ton of corn used to produce ethanol was associated with the loss of 0.49 tons of feed, loss of 0.54 tons of exports, and a gain of 0.03 tons of corn for food and industrial uses. These values should be interpreted with caution; they vary over time' choosing a different set of dates will result in different values. Moreover, since corn yield increased substantially over this time period, an alternative approach would be to determine how much ttotal feed, exports, and food and industrial use changed: between 1980 and 2009 the amount of corn used for feed increased by 18 percent, the amount of corn used for food and industrial use increased by 106 percent, and the amount exported decreased by 14 percent. From this perspective each ton of corn used to produce ethanol was associated with an increase of 0.31 tons of feed, an increase of 0.15 tons for food and industrial use, and a loss of 0.08 tons of exports. The first approach takes yield growth as independent of the ethanol program; the second approach associates all of the yield growth with the ethanol program.

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 (food and industrial vs. ethanol, food and industrial vs. feed, and ethanol vs. feed) 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 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.

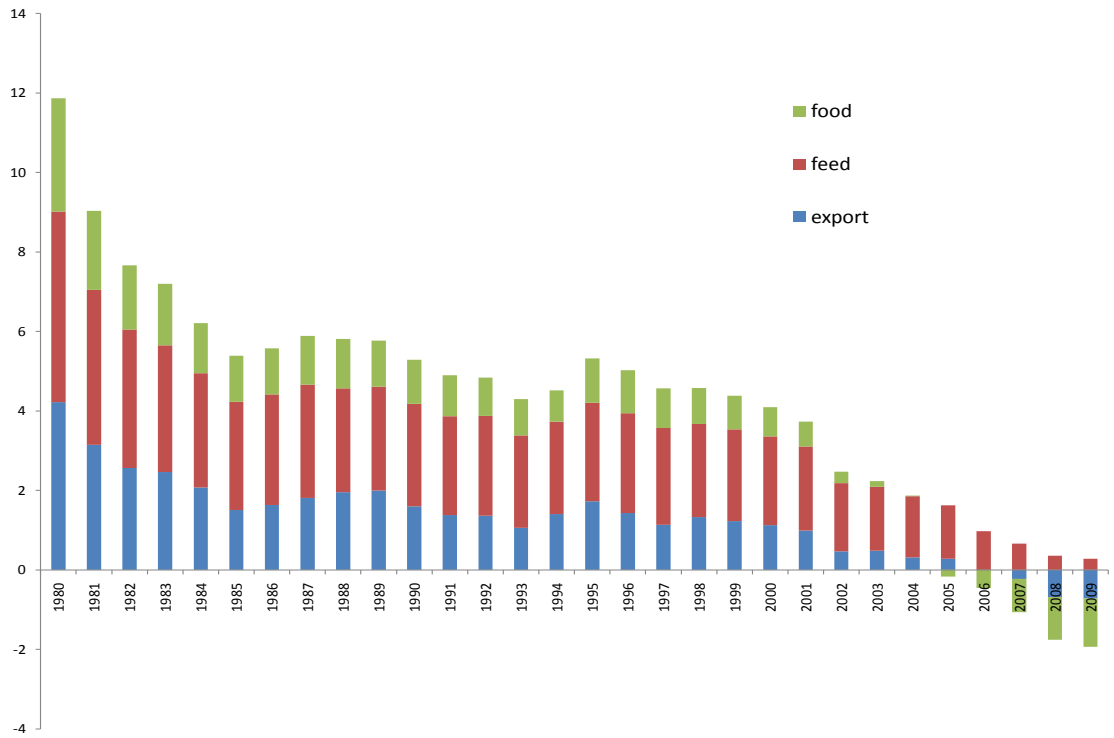


Figure 7: Log-ratios of subcomposition with respect to ethanol, feed and residue in red. The lengths within the bars are the yearly values of v_{j_m}; the by-year log ratio of the constituent fraction to with respect to ethanol. Positive values indicate amounts greater than ethanol, negative values indicate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

5 Discussion

The Dirichlet distribution is easy to simulate from directly or indirectly (see (Gelman, 2004)). 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).

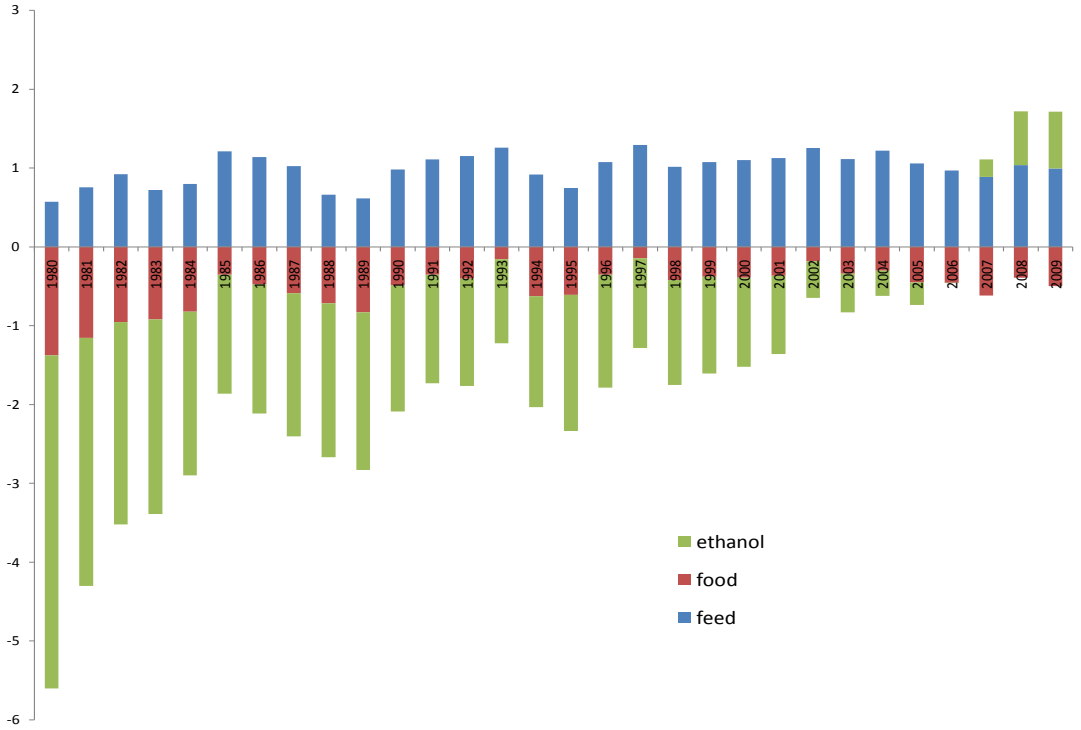


Figure 8: Log-ratios of subcomposition with respect to exports, food and industrial uses in red. The lengths within the bars are the yearly values of v_{j_m}; the by-year log ratio of the constituent fraction to with respect to exports. Positive values indicate amounts greater than exports, negative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

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 ($\grave{a}\ la$ Aitchison, see(Aitchison, 1981b)) for multivariate data in high dimensions: dependency in this setup is not restricted to an elliptical shape.

The method can be used to assess the affect of past biofuel production on other uses of the crop. It does not predict future effects of biofuel production; the relationships among crop uses can change over time. As figure 2 shows, during 1980 to 2000, use of corn for ethanol and for food and industrial uses grew while use of corn for feed remained largely constant and exports declined. But from 2000 to 2009, use of corn for ethanol increased, exports remained basically steady, food and industrial use declined somewhat and use of corn for feed declined. So comparisons to models of the effect of future growth in ethanol production on world corn markets would require additional interpretation and development.

This method provides a basis for determining dependence of changes in one crop use or land use on changes in another. Here, for example, we showed that the allocation of corn to food and industrial use may not have been significantly affected by the growth of bioethanol production. Future applications of this method could address the effects of corn-derived ethanol from a land use perspective, could include additional related crops such as soy, and could address the extent to which land use changes in Brazil have been influenced by the development of cane-derived ethanol production.

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, and 6 suggest temporally sensitive dependency patterns. This is a point of departure for future investigation. Characterizing competition among bioenergy and other agricultural activities and land uses will be important as bioenergy production approaches naturally constrained production ceilings.

References

110th US Congress 2007a. *Energy Independence and Security Act of 2007. Publication 110-140, 110th Congress*.

110th US Congress 2007b. Food and energy security act of 2007, *Report of the Committee on Agriculture, Nutrition, and Forestry. Publication 110-220. 110th Congress*.

Agency, U. E. P. 2007. Agricultural sector impacts, *Regulatory Impact Analysis: Renewable Fuel Standard Program*.

Aitchison, J. 1981a. Distributions on the simplex for the analysis of neutrality, *Statistical distributions in scientific work proceedings of the NATO Advanced Study Institute*, the Universita degli Studi di Trieste, Trieste.

Aitchison, J. 1981b. A new approach to null correlations of proportions, *Mathematical Geology* **13**(2): 175–189.

Aitchison, J. 1982. The statistical analysis of compositional data, *Journal of the Royal Statistical Society, Ser. B* **44**: 139–177.

Aitchison, J. 1986. *The Statistical Analysis of Compositional Data*, Monographs on Statistics and Applied Probability, Chapman and Hall Ltf. London.

Aitchison, J. 1997. The one-hour course in compositional data analysis or compositional data analysis is simple, *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.

Aitchison, J. 2000. Logratio analysis and compositional distance, *Mathematical Geology* **32**(3): 271–275.

Aitchison, J. and Shen, S. M. 1980. Logistic-normal distributions, *Biometrika* **67**(2): 261–272.

Barcelo-Vidal, C., Martn-Fernandez, J. A. and Pawlowsky-Glahn, V. 2001. Mathematical foundations of compositional data analysis, *Proceedings of IAMG'01 - the sixth annual conference of the International Association for Mathematical Geology*.

Barndorff-Nielsen, O. and Jorgensen, B. 1991. Some parametric models on the simplex, *Journal of Multivariate Analysis* **36**(1): 106–1016.

Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**: 829–836.

Connor, R. and Mosimann, J. 1969. Concepts of independence for proportions with a generalization of dirichlet distribution, *Journal of the American Statistical Association* **64**(325): 194–&.

De La Torre Ugarte D. G., English B. C., J. K. 2007. Sixty billion gallons by 2030: Economic and agricultural impacts of ethanol and biodiesel explansion, *Am. J. Agric. Econ.* **89**: 1290–1295.

Deheuvels, P. 1979. La fonction de dependance empirique et ses proprietes: Un test non parametrique d'independance, *Acad. Roy. Belg. Bull. Cl. Sci.* **65**: 274–292.

Fasano, G. and Franceschini, A. 1987. A multivariate version of the kolmorogov-smirnov test, *Notes of the Royal Astronomical Society* **225**: 155–170.

Gelman, A. 2004. *Bayesian Data Analysis*, 2 edn, Chapman and Hall.

Genest, C., Quessy, J. and Rmillard, B. 2006. Goodness-of-fit procedures of copula models based on the probability integral transform, *Scandinavian Journal of Statistics* **33**: 337–366.

Genest, C. and Rivest, L. P. 2001. On the multivariate probability integral transform, *Statistics and Probability Letters* **53**: 391–399.

Genest, C. and Rmillard, B. 2004. Test of independence and randomness based on the empirical copula process, *TEST* **13**: 335–369.

Gupta, R. and Richards, D. 2001. The history of the dirichlet and liouville distributions, *International Statistical Review* **69**(3): 433–446.

James, I. and Mosimann, J. 1980. A new characterization of the dirichlet distribution through neutrality, *The Annals of Statistics* **8**(1): 183–189.

Justel A, Pena D, Z. R. 1997. A multivariate kolmogorov-smirnov test of goodness of fit, *Statistics and Probability Letters* pp. 251–259.

Keeney, R. and Hertel, T. W. 2009. The indirect land use impacts of united states biofuel policies: The importance of acreage, yield, and bilateral trade responses, *Am. J. Agric. Econ.* **91**(4): 895–909.

Nelsen, R. 1999. *An Introduction to Copulas*, Springer.

Nelsen, R. 2003. Kendall distribution functions, *Statistics and Probability Letters.* **65**: 263–268.

Pawlowsky-Glahn, V. and Mateu-Figueras, G. 2005. The statistical analysis on coordinates in constrained spaces, *55th session of the International Statistical Institute*, Sydney Convention and Exhibition Centre, Sydney, Australia.

Rayens, W. and Srinivasan, C. 1994. Dependence properties of generalized liouville distributions on the simplex, *Journal of the American Statistical Association* **89**(428): 1465–1470.

Renewable Fuels Association 2009. *Historic U. S. Fuel Ethanol Production*.

Rosenblatt, M. 1952. Remarks on a multivariate transformation, *The Annals of Mathematical Statistics* pp. 470–472.

Thompson W., Meyer S., W. P. 2009. How does petroleum price and corn yield volatility affect ethanol markets with and without an ethanol use mandate?, *Energy Policy* **37(2)**: 745–749.

Tilman, D., Socolow, R., Foley, J. A., Hill, J., Larson, E., Lynd, L., Pacala, S., Reilly, J., Searchinger, T., Somerville, C. and Williams, R. 2009. Beneficial Biofuels-The Food, Energy, and Environment Trilemma, *SCIENCE* **325**(5938): 270–271.

USDA, E. R. S. 2010. Feed grains database, custom queries, `http://www.ers.usda.gov/Data/FeedGrains/`.

Williams, D. 2001. *Weighing the Odds: A Course in Probability and Statistics*, Cambridge University Press. Cambridge.

Wolff, E. F. 1980. N-dimensional measures of dependence, *Stochastica* **4**: 10–35.