Quantifying rock fabrics – a test of independence of the spatial distribution of crystals

J.R. Mackenzie[1,*],
J.J. Egozcue[2] , **R. Heilbronner**[1] , **R. Hielscher**[3,4] , **A. Müller**[3] , **and H. Schaeben**[3]
[1]Basel University, Switzerland
[2]Catalonia University of Technology, Barcelona, Spain
[3]Freiberg University of Mining and Technology, Germany; *helmut.schaeben@geo.tu-freiberg.de*
[4]now with GSF–National Research Center for Environment and Health, Neuherberg, Germany
[*]missing in the Alps since Jul 2, 2006

Abstract

A novel test of spatial independence of the distribution of crystals or phases in rocks based on compositional statistics is introduced. It improves and generalizes the common joins–count statistics known from map analysis in geographic information systems.

Assigning phases independently to objects in $\mathbb{R}^D$ is modelled by a single-trial multinomial random function $\mathbf{Z}(\mathbf{x})$, where the probabilities of phases add to one and are explicitly modelled as compositions in the K–part simplex $\mathcal{S}^K$. Thus, apparent inconsistencies of the tests based on the conventional joins–count statistics and their possibly contradictory interpretations are avoided. In practical applications we assume that the probabilities of phases do not depend on the location but are identical everywhere in the domain of definition. Thus, the model involves the sum of r independent identical multinomial distributed 1-trial random variables which is an r-trial multinomial distributed random variable. The probabilities of the distribution of the r counts can be considered as a composition in the Q–part simplex $\mathcal{S}^Q$. They span the so called Hardy–Weinberg manifold $\mathcal{H}$ that is proved to be a $K-1$-affine subspace of $\mathcal{S}^Q$. This is a generalisation of the well-known Hardy–Weinberg law of genetics. If the assignment of phases accounts for some kind of spatial dependence, then the r-trial probabilities do not remain on $\mathcal{H}$. This suggests the use of the Aitchison distance between observed probabilities to $\mathcal{H}$ to test dependence. Moreover, when there is a spatial fluctuation of the multinomial probabilities, the observed r-trial probabilities move on $\mathcal{H}$. This shift can be used as to check for these fluctuations. A practical procedure and an algorithm to perform the test have been developed. Some cases applied to simulated and real data are presented.

Key words: Spatial distribution of crystals in rocks, spatial distribution of phases, joins–count statistics, multinomial distribution, Hardy–Weinberg law, Hardy–Weinberg manifold, Aitchison geometry.

1 Introduction

There are a number of deformation mechanisms that involve mixing of particles of different phases. For example, in cataclastic flow, particles are fragmented and displaced past each other, in diffusion creep, grains of one phase nucleate and grow between grains of other phases. The resulting mixtures may form spatially independent or spatially dependent patterns. If there are more than two different phases, deviation from independence can be realised in many ways. To infer the nature of the underlying process and to identify the active deformation mechanisms, it is necessary to find reliable statistics by which dependent and independent spatial distribution of phases can be distinguished from one another and where the deviation can be quantified.

Figure 1 displays an image of a thin section exposing calcite and dolomite grains.

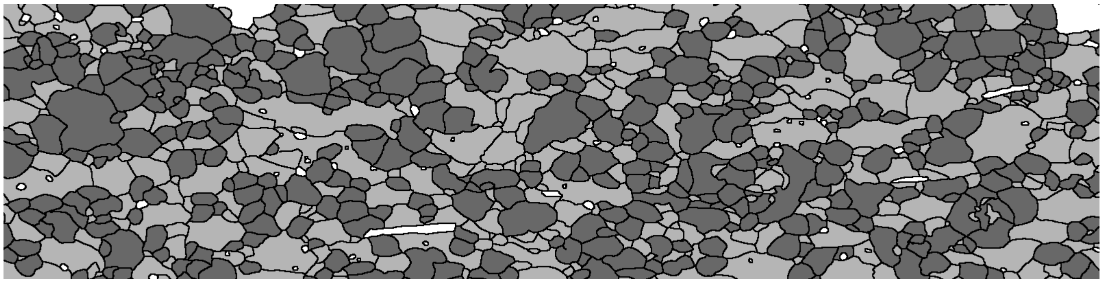

Figure 1: Binary image in enhanced pixel mode of calcite and dolomite grains.

The problem was explicitly stated and approached by Kretz (1969) and later e.g. by Jerram et al. (1996) and Mackenzie et al. (2005; 2006). Corresponding mathematical–statistical models date back to Moran (1948), Cliff and Ord (1973, 1981), and can be found in textbooks like Chilès and Delfiner (1999), Haining (2003), and many others. Map analysis in geographic information systems (GIS) applies common statistics to joins–count, cf. Bonham-Carter (1994), O'Sullivan and Unwin (2003). The latter report of apparent inconsistencies of the tests based on the joins–count statistics and their possibly contradictory interpretations. Another test has been developed by Hahn (1995) which avoids the reported inconsistencies by merging several statistics into a unique new one. She also defines a maximum–likelihood estimator for the usually unknown probabilities of a phases in terms of the boundaries of grains and not in terms of their total number or volume.

Usually, joins–count statistics are exemplified by analysing patterns composed of two phases in the plane. Non of the approaches accounts for the compositional character of the counts, and generalisations to more than two phases distributed in space seem to be largely involved (cf. Epperson, 2003).

In Section 2 the joins–count approach for two phases in the plane is reviewed. Then we generalise the approach to more than two phases in spaces of arbitrary dimension emphasising that counts are actually compositions. Thus we prepare the proper basics for compositional statistics of joins–counts. In Section 3, we introduce a measure for the deviation from independence by applying Aitchison geometry and the corresponding transformations, respectively, to determine the distance of an arbitrary composition from the Hardy–Weinberg manifold in the simplex representing independence. Nevertheless, a full account of the theory will be given elsewhere. Then, in Section 4, we set out to give a step-by-step procedure to test the null–hypotheses of independence. In Section 5, we present some preliminary results both for simulated and real–world data.

2 Common statistics of joins–counts

We assume that the unit square $[0, 1]^2 \subset \mathbb{R}^2$ is partitioned into polygonal cells given by their boundaries (vector mode) or regularly into square pixels (pixel mode). For cellular partitions we refer the reader to Mallet (2002). In the most simple case, one of two possible phases (states, colours) is assigned to each cell or pixel, respectively. The assignment of phases may be independent or account for some spatial dependence. Technically and more illustratively we refer to the assignment of phases as to colour–coding.

The subject of joins–count statistics is to test for independence of phases, states or colours. For reasons of simplicity we may refer to the two states as "black" and "white". Generally, the probabilities of their occurrence require a distinction whether their definition corresponds to frequencies given in terms of proportion of either numbers, volume, surface, area, or perimeter. Just in case of pixels, all definitions coincide.

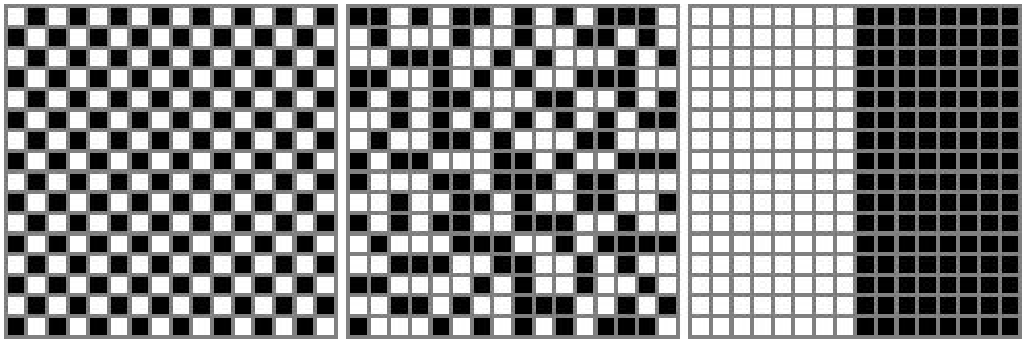

Figure 2: Binary image in pixel mode: (A) anticlustered (left); (B) independent (centre); (C) clustered (right).

Figure 2 shows three binary images displaying an anticlustered or negatively autocorrelated pattern (Fig. 2A), an apparently independent sometimes called random pattern (Fig. 2B), and a clustered or positively autocorrelated pattern (Fig. 2C). If the probabilities of the colours are constant over the square, i.e. if they are independent of the position of the pixel, then there are only two ways of deviation from independence in the case of a binary image. If the total number of colours increases, so do the possibilities for deviations from independence.

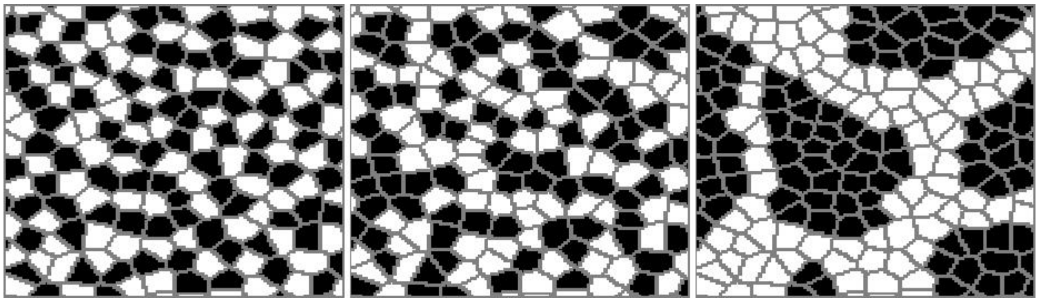

Figure 3: Binary image in enhanced pixel mode: (A) anticlustered (left; (B) independent (centre); (C) clustered (right).

In practical problems, we usually analyse digital images which come in pixel mode displaying polygonal cells ("grains") which are homogeneous with respect to their colours.

To apply joins–counts the boundaries between grains must be explicitly given as a third "neutral" colour just indicating the boundary. Otherwise joins between grains of the same colour could not be detected nor counted. For the time being we refer to this mode as “enhanced pixel mode” (Fig. 3).

The probability p of a colour is generally unknown and has to be estimated by $\widehat{p}$ form the given digital image. With respect to polygonal cells, there are several distinct approaches to estimate this probability. They basically differ in the way we think of probability, probability by total number, probability by volume or area, probability by surface or perimeter, etc. Usually the definition of probabilities with reference to proportions of perimeter applies.

Assuming that the probability for a cell or pixel to be black is known to be $p, 0 < p < 1$, and to be white is $1 - p$, the probabilities of joins of kinds can easily be given if the colour–coding is assumed to be independent. Then these probabilities are

$$\mathrm{P}(\text{black} \mid \text{black}) = p^2, \;\; \mathrm{P}(\text{black} \mid \text{white}) = 2p(1-p), \;\; \mathrm{P}(\text{white} \mid \text{white}) = (1-p)^2 \;, \qquad (1)$$

which can be identified to the Hardy–Weinberg law of genetics.

Let J denote the total number of all joins and J_{bb}, J_{bw}, J_{ww} the total number of black-black, black-white, and white-white joins, respectively, let i and j label the two sampling units (cells, pixels) being compared, and let a_i denote the attribute of sampling unit i with

$$a_i = \begin{cases} 1 & \text{if unit } i \text{ is black} \\ 0 & \text{otherwise} \end{cases} \;,$$

and let d_{ij} denote the entry in the contiguity matrix indicating or weighting the adjacency of sampling units i and j.

Then the joins–counts are defined as

$$J_{bb} = \frac{1}{2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} d_{ij}a_ia_j, \;\; J_{bw} = \frac{1}{2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} d_{ij}(a_i-a_j)^2, \;\; J_{ww} = \frac{1}{2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} d_{ij} - (j_{bb} + j_{bw}) \;,$$

with

$$J = J_{bb} + J_{bw} + J_{ww} \;, \qquad (2)$$

Obviously, the three statistics are not independent, a fact only noticed by few authors, e.g. Hahn (1995), and Fortin and Dale (2005, p. 119). In fact, they are obviously are compositions (2).

Then, the actual counts are compared with the expected counts jp^2, $j2p(1-p)$, $j(1-p)^2$ if p is known or with the estimated expected counts $j\widehat{p}^2$, $j2\widehat{p}(1-\widehat{p})$, $j(1-\widehat{p})^2$ if p is estimated by $\widehat{p}$, respectively, according to the distribution given by (1). The null–hypothesis is always independence. Hahn (1995) merges the three statistics into the unique test statistic

$$T = \frac{(J_{bb} - J\,\widehat{p}^2)^2}{J\widehat{p}^2} + \frac{(J_{bw} - J\,2\widehat{p}(1-\widehat{p}))^2}{J\,2\widehat{p}(1-\widehat{p})} + \frac{(J_{ww} - J\,(1-\widehat{p})^2)^2}{J\,(1-\widehat{p})^2} \;, \qquad (3)$$

with the maximum–likelihood estimator

$$\widehat{p} = \frac{2J_{bb} + J_{bw}}{2J} \;, \qquad (4)$$

in terms of joins–counts. With (4), the estimator (3) simplifies to

$$T = \frac{J(4J_{bb}J_{ww} - J_{bw}^2)^2}{(J_{bw} - 2J_{bb})^2(J_{bw} + 2J_{ww})^2} \;,$$

which is distributed as χ_1^2 given the the null–hypothesis of independence (Hahn, 1995). Others, e.g. O’Sullivan and Unwin (2003, pp. 190), check three test statistics independently and report that the three tests may appear to contradict one another. We suspect that apparent contradictions may origin in neglecting the dependence of the three statistics, i.e. their characteristics as being compositions.

3 Compositional statistics of joins–counts

A new development is presented here to test independence of colour–coding. Three main aspects are accounted for trying to generalise and improve the techniques described in Section 2. They are:

- The concept of joins is generalised to more general patterns, applicable both to pixel mode or vector mode.
- The space of the multinomial probabilities involved, i.e. probabilities of the colour–coding and the probabilities of generalised joins, is assumed to be the simplex with its associate Aitchison geometry.
- Equations of the equilibrium (1) (Hardy–Weinberg law) are generalised to more than two colours and joins larger than binary.
- Dependence is checked in two different aspects: dependence of the random colour–coding and dependence of the parameters of these multinomials along the given pattern.

Only a sketch of the whole theory is here presented.

3.1 Random colours on a set

Let $\mathbb{G}$ be a subset of $\mathbb{R}^D$. Frequently, but not necessarily, $\mathbb{G}$ will be assumed connected and bounded. Each point of $\mathbb{G}$ is denoted by $\mathbf{x} = (x_1, x_2, \ldots, x_D)$, where x_i is i-th coordinate of $\mathbf{x} \in \mathbb{R}^D$. A colour, k, from a set of different colours $\{1, \ldots, K\}$ is assigned to each point in $\mathbb{G}$. For instance, $\mathbb{G}$ can be tessellated into cells with a colour assigned. The colour assigned to the point $\mathbf{x}$ is that of the cell that contains the point. There are three natural questions about the way of assigning colours on $\mathbb{G}$,

a. Which are the proportions, $\mathbf{p} = (p_1, p_2, \ldots, p_K)$, of assigned colours for the whole $\mathbb{G}$?

b. Is the proportion $\mathbf{p}$ stable for any subset $\mathbb{G}_1 \subset \mathbb{G}$?, i.e. is there some trend or fluctuation of $\mathbf{p}$ when shifting from one point to another point across $\mathbb{G}$?

c. Irrespective of a possible trend in $\mathbf{p}$, is there some evidence against an independent assignment of colours?

If the method of colour–coding was that of the cells, it is obvious that the colouring of very close points should be largely dependent. However, when the points compared are more distant than the diameter of the cells, dependence could be drastically reduced and may vanish.

The framework where this problem can be stated is to consider a single-trial multinomial random function $\mathbf{Z}(\mathbf{x})$ on $\mathbb{G}$, where the multinomial parameters $\mathbf{p}(\mathbf{x})$ may depend on the position $\mathbf{x}$ within the $\mathbb{G}$. This means that

$$\mathrm{P}[\mathbf{Z}(\mathbf{x}) = \mathbf{z}_k | \mathbf{p}(\mathbf{x})] = p_k(\mathbf{x}) \ , \ \ k = 1, 2, \ldots, K \ ,$$

where $\mathbf{z}_k = (0, 0, \ldots, 1, \ldots, 0)$ with the only 1 placed in the k-component. The dependence in $\mathbf{x}$ of these expressions will be assumed and the argument $(\mathbf{x})$ will be suppressed if not necessary. On the other hand, vectors of probabilities can be considered to be elements of the simplex of K parts, $\mathcal{S}^K$, equipped with the so-called Aitchison geometry (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001; Aitchison et al., 2002; Egozcue et al., 2003) thus becoming an $(K-1)$-dimensional Euclidean space.

The key to confront the problem is to find out an univariate quantity measuring randomness of the colours, which do not depend on the eventually dependence of $\mathbf{p}$ on $\mathbf{x}$, i.e. invariant under shifting within $\mathbb{G}$. The main ideas come from the Hardy–Weinberg law of genetics (Hardy, 1908; Weinberg, 1908).

3.2 Grouping points in a lag–pattern

Let $\mathbf{x}_0$ be a point in $\mathbb{G}$ and a set of $r \geq 2$ lags $\mathbf{r}_j$, $j = 1, \dots, r$, such that $\mathbf{x}_j = \mathbf{x}_0 + \mathbf{r}_j$ is in $\mathbb{G}$. By default, the lag $r_1 = (0, 0, \dots, 0)$ is assumed to be in the set of lags. The set R of the lags $\mathbf{r}_j$ is called *lag–pattern*. While a lag–pattern may be admissible for some values of $\mathbf{x}_0$, it may not be admissible for others ones, due to the effect of the boundary of $\mathbb{G}$. If $\mathbb{G} = \mathbb{R}^D$, any lag–pattern is always admisible for every $\mathbf{x}_0$. However, in practice, grids are finite and boundary effects must be taken into account. The set of $\mathbb{G}$-points $\mathbf{x}_j = \mathbf{x}_0 + \mathbf{r}_j$, for a given $\mathbf{x}_0$ and R is denoted $R(\mathbf{x}_0)$.

The number of times a colour is assigned to a point in $R(\mathbf{x}_0)$ can be expressed as the sum of random vectors

$$\mathbf{N}(R(\mathbf{x}_0)) = \sum_{\mathbf{x} \in R(\mathbf{x}_0)} \mathbf{Z}(\mathbf{x}) \ .$$

This concept replaces here the idea of join–count. Now joins are more than binary and they can be placed in an arbitrary relative position.

Given a point $\mathbf{x}_0$ and a lag–pattern, multinomial random vectors $\mathbf{Z}(\mathbf{x})$ can be observed on the set $R(\mathbf{x}_0)$

When the multinomial random vectors $\mathbf{Z}(\mathbf{x})$ are assumed to be mutually independent, the distribution of $\mathbf{N}(R(\mathbf{x}_0))$ can be obtained. The sum of these observations, i.e. the number of times that each colour is observed on $R(\mathbf{x}_0)$, has a known distribution. However, this distribution is multinomial whenever the probabilities of each added 1-trial multinomial are equal. The following section describes properties of sums of single-trial multinomial vectors.

3.3 Sums of multinomials

Sums of independent one-trial multinomials are studied. This includes the case in which different probabilities are assumed for each added multinomial. However, attention will eventually be focused on the well-known case where all probabilities are equal.

Proposition 1. *Let $\mathbf{Z}_j$, $j = 1, 2, \dots, r$ be independent, single-trial multinomial random vectors of K-components, with probability parameters $\mathbf{p}(j) = (p_1(j), p_2(j), \dots, p_K(j))$ respectively. Then, the probability function of the random vector $\mathbf{N} = \sum_{j=1}^r \mathbf{Z}_j$ is given by the coefficients of the $(K-1)$–variable, r–degree polynomial*

$$\phi(\xi_1, \xi_2, \dots, \xi_K) = \prod_{j=1}^{r} \left(\sum_{i=1}^{K} \xi_i p_i(j) \right) = \sum_{n} (\xi_1^{n_1} \xi_2^{n_2} \cdots \xi_K^{n_k}) \cdot q_n \ , \tag{5}$$

where n stands for a numbering of the vector of indexes $\mathbf{n}$ ranging all terms such that $\sum_{k=1}^K n_k = r$. Specifically, $q_n = \mathrm{P}[\mathbf{N} = \mathbf{n} | \mathbf{p}(j), r]$.

Proof. Define the moment–generating function as

$$\phi(\xi_1, \xi_2, \dots, \xi_K) = \mathrm{E}\left[\xi_1^{N_1}, \xi_2^{N_2}, \dots, \xi_K^{N_K} \right] \ .$$

The moment–generating function of a one-trial multinomial with parameters in $\mathbf{p}(j)$ is

$$\phi(\xi_1, \xi_2, \dots, \xi_K; j) = \sum_{i=1}^{K} \xi_i p_i(j) \ .$$

The sum of independent one-trial multinomial random vectors has moment–generating function $\prod_j \phi(\xi_1, \xi_2, \dots, \xi_K; j)$. □

A conclusion is that $\mathbf{N}$ is not a multinomial random vector. A corollary of the previous result is the well-known additive property of the multinomial distribution.

Proposition 2. *In the conditions of Proposition 1, if* $\mathbf{p}(1) = \mathbf{p}(2) = \cdots = \mathbf{p}(r) = \mathbf{p}$, *then* $\mathbf{N}$ *is multinomial distributed with probability parameters* $\mathbf{p}$ *and* r *trials, i.e*

$$q_n = \mathrm{P}[\mathbf{N} = (n_1, n_2, \ldots, n_K)|\mathbf{p}, r] = \frac{r! \prod_{j=1}^{K} p_j^{n_k}}{\prod_{i=1}^{K} n_i!} \ , \ \sum_{i=1}^{K} n_i = r \ . \tag{6}$$

Proof. For equal $\mathbf{p}(j)$, (5) simplifies to

$$\phi(\xi_1, \xi_2, \ldots, \xi_K) = \left(\sum_{i=1}^{K} \xi_i p_i \right)^r \ . \tag{7}$$

After a little bit of combinatorics q_n can be identify with (6). □

A discrete probability function, as given in (5) or (6) can be considered as a composition in the simplex $\mathcal{S}^Q$, where Q stands for the number of probability points (non–null probabilities). In general $Q > K$, as determined in the following proposition.

Proposition 3. *Let* $\mathbf{N}$ *be a sum of* r *one-trial multinomial random vectors. The number of points in the sample space with non-null probability is*

$$Q = \frac{(r + K - 1)!}{r!(K - 1)!} \ . \tag{8}$$

Proof. Q is the number of additive partitions of K or a combination of two elements with repetitions r and $K - 1$. □

An important question is which are the properties that characterise the probability functions of a sum of independent one-trial multinomials. A key property is that r-trial multinomials are in a $(K - 1)$–dimensional linear manifold (affine subspace) of the simplex of Q parts as presented in the following new proposition.

Proposition 4. *Let* $M \subset \mathcal{S}^Q$ *be the set of compositions* $\mathbf{q} = (q_1, q_2, \ldots, q_Q)$ *such that the components are the* r*-trial multinomial probabilities, Eq. (6), for some parameter* $\mathbf{p} \in \mathcal{S}^K$. *Then,* M *is a* $(K-1)$*–dimensional linear manifold of the simplex of* $\mathcal{S}^Q$, *called Hardy–Weinberg linear manifold or* $\mathcal{H}$*–manifold (cf. Rocklin and Oster(1976), Akin and Szucs (1994)).*

Proof. Define the function $\mathbf{m} : \mathcal{S}^K \to \mathcal{S}^Q$, such that $\mathbf{q} = \mathbf{m}(\mathbf{p})$, i.e. it assigns the multinomial probabilities to a parameter vector $\mathbf{p} \in \mathcal{S}^K$. Let $\mathbf{p}_e = K^{-1}(1, 1, \ldots, 1)$ the neutral element in $\mathcal{S}^K$. The image $\mathbf{m}(\mathbf{p}_e)$ is not the neutral element in $\mathcal{S}^Q$ but the composition with components

$$q_n = \frac{r!}{K^r \prod_{i=1}^{K} n_i!} \ .$$

Denote $\mathbf{m}_0(\mathbf{p}) = \mathbf{m}(\mathbf{p}) \ominus \mathbf{q}_0$, so that the image of the neutral element in $\mathcal{S}^K$ is the neutral element in $\mathcal{S}^Q$. To prove the statement it suffices that $M_0 = M \ominus \mathbf{m}(\mathbf{p}_0)$ is a $(K-1)$–dimensional subspace of $\mathcal{S}^Q$.

To proof that M_0 is a subspace, it should be closed under powering and perturbation in $\mathcal{S}^Q$. The components of elements in M_0 have the form $q_{0n}(\mathbf{p}) = c \prod_{j=1}^{K} p_j^{n_j}$, where c is a constant equal for the Q components. For a composition in M_0 and $\alpha \in \mathbb{R}$, the α–power has components

$$c^\alpha \left(\prod_{j=1}^{K} p_j^{n_j} \right)^\alpha = c_1 \prod_{j=1}^{K} (p_j^\alpha)^{n_j} \ ,$$

where c_1 is equal for all components. This shows that M_0 is closed under powering and moreover, it corresponds to α–powering of the vector of parameters $\mathbf{p} \in \mathcal{S}^K$. This is $\alpha \odot \mathbf{m}_0(\mathbf{p}) = \mathbf{m}_0(\alpha \odot \mathbf{p})$.

Consider $\mathbf{m}_0(\mathbf{t}) \in M_0$ corresponding to the parameters $\mathbf{t} \in \mathcal{S}^K$. The perturbation in M_0, $\mathbf{m}_0(\mathbf{t}) \oplus \mathbf{m}_0(\mathbf{p})$ has components proportional to

$$\prod_{j=1}^{K} t_j^{n_j} \cdot \prod_{j=1}^{K} p_j^{n_j} = \prod_{j=1}^{K} (t_j \cdot p_j)^{n_j} \ ,$$

which implies

$$\mathbf{m}_0(\mathbf{t}) \oplus \mathbf{m}_0(\mathbf{p}) = \mathbf{m}_0(\mathbf{t} \oplus \mathbf{p}) \ .$$

Then, M_0 is closed under perturbation in $M_0 \subset \mathcal{S}^Q$ and perturbations in the subspace M_0 correspond to perturbations of the parameters in $\mathcal{S}^K$. Therefore, $\mathbf{q} = \mathbf{m}_0(\mathbf{p})$ an isomorphism between $\mathcal{S}^K$ and the subspace $M - \mathbf{q}_0 = M_0 \subset \mathcal{S}^Q$. As the dimension of $\mathcal{S}^K$ is $K-1$, this is also the dimension of M and M_0. □

More explicit characteristics of M can be found.

Proposition 5. *The composition*

$$\mathbf{q}_0 = \mathcal{C}\left[\ldots, \frac{r!}{n_1! \cdots n_K!}, \ldots\right] \ , \tag{9}$$

is a point of $M \subset \mathcal{S}^Q$. The subspace $M_0 = M \ominus \mathbf{q}_0 \subset \mathcal{S}^Q$ contains the compositions whose components are

$$\mathbf{q} \ominus \mathbf{q}_0 = \mathcal{C}\left[\ldots, \prod_{k=1}^{K} p_k^{n_k}, \ldots\right] \ , \ \sum_{k=1}^{K} n_k = r \tag{10}$$

for some preselected order of the combinations of $\mathbf{n}$. *Moreover,*

$$\mathrm{clr}(\mathbf{q} \ominus \mathbf{q}_0) = \left[\ldots, \sum_{k=1}^{K} \left(n_k - \frac{r}{K}\right) \ln p_k, \ldots\right] \ , \ \sum_{n=1}^{Q} \sum_{k=1}^{K} \left(n_k - \frac{r}{K}\right) \ln p_k = 0 \ ,$$

and, for the n_k's corresponding to the value of n,

$$\sum_{n=1}^{Q} \left(n_k - \frac{r}{K}\right) \ln p_k = 0 \ , \ k = 1, 2, \ldots, K \ . \tag{11}$$

Proof. Consider the neutral element of $\mathcal{S}^K$, $\mathbf{p}_e = K^{-1}[1, 1, \ldots, 1]$. Then $\mathbf{m}(\mathbf{p}_e) = \mathbf{q}_0$ and $\mathbf{m}_0(\mathbf{p}_e) = \mathbf{q}_e$ where $\mathbf{q}_e$ is the neutral element in $\mathcal{S}^Q$. In fact, when all probabilities p_i are equal, each component of $\mathbf{m}(\mathbf{p}_e)$ has the common factor $p_1^{n_1} p_2^{n_2} \cdots p_K^{n_K} = p^r$ and, after closure, the remaining coefficients are those of Equation (9). Carrying out the negative perturbation $\mathbf{q} \ominus \mathbf{q}_0$ Equation (10) follows.

In order to compute the $\mathrm{clr}(\mathbf{q} \ominus \mathbf{q}_0)$, first compute the geometric mean of all the components. The sum of powers of each component of $\mathbf{q} \ominus \mathbf{q}_0$ is r. The sum of all these powers over the Q components is Qr. In the product of all components, the power of each p_i are equal by symmetry. Then, in the product of the geometric mean, the power of each p_i is Qr/K. The geometric mean implies a power $1/Q$ and the geometric mean is

$$g = (p_1 p_2 \cdots p_K)^{r/K} \ .$$

Dividing each component of $\mathbf{q} \ominus \mathbf{q}_0$ by g and taking logarithms the expression of $\mathrm{clr}(\mathbf{q} \ominus \mathbf{q}_0)$ follows. Taking into account the symmetry in the p_i's, Equation (11) also follows.

□

4 Procedure

The aim is to check for evidences against the hypothesis that the colours of the points in $\mathbb{G}$ are assigned independently. This hypothesis is difficult to test in one single step and therefore a sequence of hypotheses of independence on different lag–patterns is checked instead.

For each lag–pattern three compositions in $\mathcal{S}^Q$ are obtained: (i) the multinomial parameter $\mathbf{m}(\widehat{\mathbf{p}}) \in M \subset \mathcal{S}^Q$ corresponding to the estimated probabilities of the colours, $\widehat{\mathbf{p}} \in \mathcal{S}^K$, obtained using all available points; (ii) for a given lag–pattern, the estimated multinomial probabilities $\widehat{\mathbf{q}} \in \mathcal{S}^Q$ using the counts of colours over the lag–pattern; and (iii) the orthogonal projection $\widehat{\mathbf{q}}_{\mathcal{H}}$ of $\widehat{\mathbf{q}}$ on the Hardy–Weinberg linear manifold M.

The squared distance $d_a^2(\widehat{\mathbf{q}}, \mathbf{m}(\widehat{\mathbf{p}}))$ measures the total deviation from independence for the lag–pattern over $\mathbb{G}$. The squared distance $d_a^2(\widehat{\mathbf{q}}_{\mathcal{H}}, \mathbf{m}(\widehat{\mathbf{p}}))$ measures the deviation along M due to either local changes of multinomial probabilities within the whole set $\mathbb{G}$ or, in particular, due to errors when estimating $\mathbf{p}$ by $\widehat{\mathbf{p}}$. The squared distance $d_a^2(\widehat{\mathbf{q}}, \widehat{\mathbf{q}}_{\mathcal{H}})$ measures the lack on independence of the colour–coding.

The orthogonality of the projection onto the $\mathcal{H}$–manifold implies (Pythagoras theorem)

$$\underbrace{d_a^2(\widehat{\mathbf{q}}, \mathbf{m}(\widehat{\mathbf{p}}))}_{\text{global dependence}} = \underbrace{d_a^2(\widehat{\mathbf{q}}_{\mathcal{H}}, \mathbf{m}(\widehat{\mathbf{p}}))}_{\text{fluctuation of probability}} + \underbrace{d_a^2(\widehat{\mathbf{q}}, \widehat{\mathbf{q}}_{\mathcal{H}})}_{\text{dependence}} . \tag{12}$$

In order to obtain these square distances it is convenient to shift the $\mathcal{H}$–manifold to the origen by subtracting $\mathbf{q}_0$, thus considering the subspace $M_0 = M \ominus \mathbf{q}_0$.

1. *Estimate the probabilities of each colour on* $\mathbb{G}$.

   Pixels, pixel mode: Count the total number of times each colour appears, c_i, and estimate $\widehat{\mathbf{p}}$ as
   $$\widehat{p}_i = \frac{c_i + \alpha}{n_G + \alpha K} \ , \ i = 1, 2, \ldots, K,$$
   being the more usual values of α, 0.5 and 1. The number of points considered in $\mathbb{G}$ is denoted n_G and K is the number of colours. This preliminary counting of colours following individual points in $\mathbb{G}$ can be identified as counting following the degenerate lag–pattern constituted by a single lag vector $\mathbf{r}_1 = \mathbf{0}$ (see items 2, 3, 4 and 5).

   Grains, enhanced pixel mode: Count the total number c_b of pixels indicating a grain boundary, and the total number of grain boundary pixels c_{ij} seperating grains of colours i, j, $i, j = 1, \ldots, K$, and estimate $\widehat{\mathbf{p}}$ as
   $$\widehat{p}_i = \frac{\sum_{j=1}^{K} c_{ij} + \alpha}{c_b + \alpha K} \ , \ i = 1, 2, \ldots, K.$$

2. *Define a sequence of lag–patterns to be tested.* Define a sequence of lag–patterns R_1, R_2, ..., R_ℓ. Conveniently, the complexity of the sequence is considered as being increasing. For instance $R_1(\mathbf{x}_0)$ may contain the neighbours (length of vectors $\mathbf{r}_i$ in the pattern of length 1) of $\mathbf{x}_0$ in the direction of the reference axes (or alternatively, following the direction of only one axis). The second $R_2(\mathbf{x}_0)$ may be the six points, regularly distributed at a distance of 2. Another possibility is to group a set of vectors on a straight line, etc. A lag–pattern for which independence is not rejected should be considered as an inadequate training pattern in a multi–point spatial analysis.

3. *Move lag–pattern over* $\mathbb{G}$. Search for all possible positions $\mathbf{x}_0$ for which the points $\mathbf{x}_0 + \mathbf{r}_i$, $i = 1, 2, \ldots, r$ are in $\mathbb{G}$.

4. *Compute the number of times colours appear on the lag–pattern for each admissible* $\mathbf{x}_0$. This is, compute the realization of $\mathbf{N}(R(\mathbf{x}_0))$ for each $\mathbf{x}_0$. Denote the number of admissible positions of the lag–pattern m_R.

5. *Estimate the probabilities of the Q probability points.* Using the data obtained in step 4, and assuming they are independent, estimate the probabilities for each probability point. There are Q probability points of the sum of r one-trial multinomials given by (8). Zeroes should be avoided. For instance (notation of section 3.3),

$$\widehat{\mathbf{q}}_n = \frac{m_n + \alpha}{m_R + \alpha Q} \ , \ \sum_{j=1}^{Q} m_j = m_R \ ,$$

where m_n is the number of times the combination of colours corresponding to the numbering n have been found when running $\mathbf{x}_0$ over $\mathbb{G}$.

6. *Compute the composition on the $\mathcal{H}$-manifold corresponding to $\widehat{\mathbf{p}}$.* The parts of $\mathbf{m}(\widehat{\mathbf{p}}) \in \mathcal{S}^Q$ are

$$\mathbf{m}(\widehat{\mathbf{p}})_n = \frac{r! \ \prod_{j=1}^{K} \widehat{p}_j^{n_k}}{\prod_{i=1}^{K} n_i!} \ , \ \sum_{i=1}^{K} n_i = r \ ,$$

where the subindex n denotes the component of the compositional vector following the numbering criterion, and after removing $\mathbf{q}_0$ is

$$\mathbf{m}_0(\widehat{\mathbf{p}}) = (\widehat{\mathbf{q}} \ominus \mathbf{q}_0) = \mathcal{C}\left[\ldots, \prod_{k=1}^{K} \widehat{p}_k^{n_k}, \ldots\right] \ , \ \sum_{k=1}^{K} n_k = r \ .$$

7. *Compute the Aitchison distance $d_a^2(\widehat{\mathbf{q}}, \mathbf{m}(\widehat{\mathbf{p}}))$.* This can be computed in a number of ways. Here the clr–way is suggested. First, compute the clr–coefficients of the two vectors $\widehat{\mathbf{q}}$, $\mathbf{m}(\widehat{\mathbf{p}})$ or better $\widehat{\mathbf{q}} \ominus \mathbf{q}_0$, $\mathbf{m}_0(\widehat{\mathbf{p}})$. Denote by $\mathbf{q}$ a generic vector in $\mathcal{S}^Q$; the clr–coefficients constitute a real vector of Q components

$$\mathrm{clr}(\mathbf{q})_n = \mathcal{C}\left[\ldots, \ln \frac{q_n}{g(\mathbf{q})}, \ldots\right] \ , \ g(\mathbf{q}) = \exp\left(\frac{1}{Q}\sum_{j=1}^{Q} \ln q_j\right) \ .$$

The Aichison distance between the two compositions is then the ordinary Euclidean distance between $\mathrm{clr}(\widehat{\mathbf{q}})$, $\mathrm{clr}(\mathbf{m}(\widehat{\mathbf{p}}))$ or $\mathrm{clr}(\widehat{\mathbf{q}} \ominus \mathbf{q}_0)$, $\mathrm{clr}(\mathbf{m}_0(\widehat{\mathbf{p}}))$

8. *Find an orthonormal basis of the $\mathcal{H}$-manifold or the shifted subspace $M_0 = M \ominus \mathbf{q}_0$.* To get $K-1$ independent vectors in M_0, consider, e.g., an orthogonal basis in $\mathcal{S}^K$, $\mathbf{b}_i$, $i = 1, 2, \ldots, K-1$. They can be obtained using some sequential binary partition (Egozcue and Pawlowsky-Glahn, 2005). Compute the $K-1$ compositions $\mathbf{m}_0(\mathbf{b}_i) \in \mathcal{S}^Q$, and their respective $\mathrm{clr}(\mathbf{m}_0(\mathbf{b}_i))$. Group these clr–coefficients in a matrix B by rows. They are independent vectors of $\mathbb{R}^Q$ and are in the clr–transformed subspace $\mathrm{clr}(M_0)$. B is a $(K-1, Q)$ matrix, with rank $K-1$ provided $K < Q$. Recall that the rows of B add to zero due to clr-transformation. Proceed to the SVD of B^t given by

$$\underbrace{B^t}_{(Q,K-1)} = \underbrace{U}_{(Q,K-1)} \cdot \underbrace{D}_{(K-1,K-1)} \cdot \underbrace{V}_{(K-1,K-1)} \ ,$$

where the diagonal matrix D has positive entries (singular values), U has unitary and orthogonal real column–vectors, and V has unitary and orthogonal real row–vectors. Therefore, the columns of U, $\mathbf{u}_i$, $i = 1, 2, \ldots, K-1$ are the clr–coefficients of $K-1$ orthonormal compositional vectors in the subspace $\mathcal{H}_0$.

9. *Find the projection of $\widehat{\mathbf{q}}$ on M.* These is done in terms of clr–coefficients. Use $\mathrm{clr}(\widehat{\mathbf{q}})$ to get the orthogonal projections

$$\gamma_i = \langle \widehat{\mathbf{q}}, \mathrm{clr}^{-1}(\mathbf{u}_i)\rangle_a = \langle \mathrm{clr}(\widehat{\mathbf{q}} \ominus \mathbf{q}_0), \mathbf{u}_i \rangle \ , \ i = 1, 2, \ldots, K-1$$

so that

$$\mathbf{q}_{\mathcal{H}} \ominus \mathbf{q}_0 = \bigoplus_{i=1}^{K-1} \gamma_i \odot \mathrm{clr}^{-1}(\mathbf{u}_i) \ ,$$

and, therefore, the clr–expression is

$$\mathrm{clr}(\mathbf{q}_{\mathcal{H}} \ominus \mathbf{q}_0) = \sum_{i=1}^{K-1} \gamma_i \mathbf{u}_i \ .$$

10. *Compute the distances $d_a^2(\widehat{\mathbf{q}}_{\mathcal{H}}, \mathbf{m}(\widehat{\mathbf{p}}))$ and $d_a^2(\widehat{\mathbf{q}}, \widehat{\mathbf{q}}_{\mathcal{H}})$.* The first one is the Euclidean distance between $\mathrm{clr}(\mathbf{q}_{\mathcal{H}} \ominus \mathbf{q}_0)$ and $\mathrm{clr}(\mathbf{m}_0(\widehat{\mathbf{p}}))$ computed previously in step 7. The second distance is also computed in terms of clr–coefficients of $\widehat{\mathbf{q}}_{\mathcal{H}} \ominus \mathbf{q}_0$ (step 9) carrying out the substraction

$$\mathrm{clr}(\mathbf{q}_{\mathcal{H}} \ominus \mathbf{q}_0) - \mathrm{clr}(\widehat{\mathbf{q}} \ominus \mathbf{q}_0) = \mathrm{clr}(\mathbf{q}_{\mathcal{H}}) - \mathrm{clr}(\widehat{\mathbf{q}}) \ .$$

The required distance is found as the Euclidean distance of the respective clr–vectors.

11. *Find the p–value in a test of hypothesis.* Consider the following tests of hypothesis:

    – H_0^{gl} : the colour–coding is *globally* independent in $\mathbb{G}$ following the lag–pattern R. The natural test statistics for this hypothesis is $d_a^2(\widehat{\mathbf{q}}, \mathbf{m}(\widehat{\mathbf{p}}))$, with rejection for large values.

    – H_0^{spc} : the probabilities of each colour are spatially invariant on $\mathbb{G}$ with respect the lag–pattern R. The appropriate test statistics is $d_a^2(\widehat{\mathbf{q}}_{\mathcal{H}}, \mathbf{m}(\widehat{\mathbf{p}}))$, with rejection for large values.

    – H_0^{smpl} : the multinomial sampling of the colour–coding is independent in $\mathbb{G}$ with respect to the lag–pattern R. The test statistic is $d_a^2(\widehat{\mathbf{q}}, \widehat{\mathbf{q}}_{\mathcal{H}})$, with rejection for large values.

    The distribution of the three mentioned test statistics is not exactly known although one can assume that they are asymptotically χ^2. Moreover, the asymptotic conditions will be seldom attained whenever the number K of colours and the number r of points in the lag–pattern are not very small. The reason is that, for large Q, there will be small probabilities that should be estimated with a very large number of multinomial trials. Therefore, the distribution of these test statistics should be studied by simulation under the conditions of the null hypothesis.

The mentioned three test statistics are linked by the Pythagoras theorem (12) and cannot be independent. The distribution of the test statistics, and the estimation of the multinomial parameters are critical issues that should be addressed in the future.

5 Practical applications

Following the procedure outlined in the previous Section 4, we shall analyse some preliminary examples and simulated binary images, i.e. two–dimensional binary patterns, as their results can easily be visualised in the three part simplex $\mathcal{S}^3$ of compositions $q = [q_1, q_2, q_3]$, $q_1 + q_2 + q_3 = 1$, containing the one–dimensional Hardy–Weinberg manifold $M \subset \mathcal{S}^3$ defined by compositions of the form $q = [p^2, (1-p)^2, 2p(1-p1)]$ characterising the independent case.

In case of grains, they have been generated by Voronoi (–Thiessen–Dirichlet) cells, even though we never employ their feature as being convex.

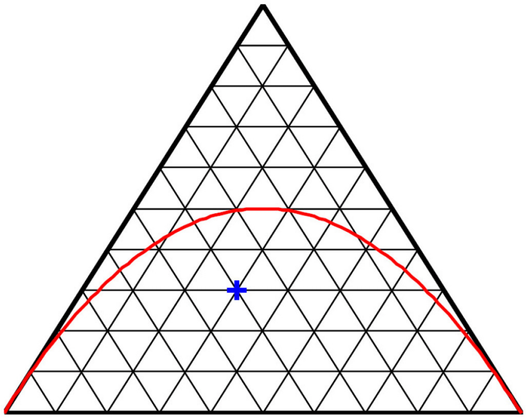

Figure 4: The three–part simplex $\mathcal{S}^3$, the one–dimensional Hardy–Weinberg manifold $\mathcal{H}$ (red line), and an arbitrary composition q of three parts (blue cross).

5.1 General considerations

Figure 4 shows the two–dimensional simplex $\mathcal{S}^3$, the one–dimensional Hardy–Weinberg manifold $\mathcal{H}$ corresponding to independent colour coding, and an arbitrary composition of three parts indicating deviation from independence.

Just by visual inspection we may conclude that the colour–coding represented by the composition $q = [q_1, q_2, q_3]$ tends to some extent towards clustering. A point above the line representing the Hardy–Weinberg manifold would indicate a tendency towards anticlustering. In the next step we apply compositional geometry to quantify the deviation in terms of the distance of the composition q from the Hardy–Weinberg manifold $\mathcal{H}$. Eventually, we would like to test whether the deviation is significant or not by comparison of the actual distance with quantiles of an appropriate distribution.

To construct an orthonormal basis of the H–manifold, or more specifically the shifted subspace $M_0 = M \ominus \mathbf{q}_0$, we proceed as follows. We choose $\mathbf{b}_0 = (1, 1, \ldots, 1)^t$ and the $K-1$ basis vectors $\mathbf{b}_i,\ \ i = 1, \ldots, K-1$ as

$$\mathbf{b}_1 = (2, 1, \ldots, 1)^t,\ \ \mathbf{b}_2 = (2, 2, 1, \ldots, 1)^t,\ \ \mathbf{b}_3 = (2, 2, 2, \ldots, 1)^t,\ \ \ldots, \mathbf{b}_{K-1} = (2, 2, \ldots, 2, 1)^t,$$

and apply the mapping $\mathbf{m} : \mathcal{S}^K \rightarrow \mathcal{S}^Q$ of Proposition (4) to get

$$\mathbf{v}_i = \mathbf{m}(\mathbf{b}_i) - \mathbf{m}(\mathbf{b}_0)\ ,\ \ i = 1, 2, \ldots, K,$$

which are elements of $\mathcal{S}^Q$. After applying the clr–transform, we have $K-1$ linearly independent vectors in $\mathbb{R}^Q$. Next we use the Gram–Schmidt orthogonalisation to accomplish a set of orthonormal vectors $\mathbf{u}_i$. Applying exponentiation we transform them back to M. Then we use this orthonormal basis to compute the required distances related to an actual count on a lag–pattern.

Eventually, a sign has been assigned to the Aitchison distance such that it is negaitive if the composition q is above the line representing the Hardy–Weinberg manifold, and positive otherwise.

5.2 Examples

5.2.1 Simulation of independence

To get some first ideas we take a look at the most simple case of binary images. To this end we simulate a total of 1.200 binary images with independent colour coding for different probabilites $\mathbf{p}$. We actually choose (i) $\mathbf{p} = (0.5, 0.5)^t$, (ii) $\mathbf{p} = (0.45, 0.55)^t$, (iii) $\mathbf{p} = (0.3, 0.7)^t$, and (iv) (i) $\mathbf{p} = (0.1, 0.9)^t$, and generate 300 images for each case.

Choosing a lag–pattern of five pixels in a row or a column, respectively, moving the centre pixel along the boundary pixels, and registering the colours in the first and fifth pixel only, i.e. $r = 2$, we degenerate our approach to mimic joins–counts. More specifically, we count the total number c_b of pixels indicating a grain boundary, and the total number of grain boundary pixels c_{ij} seperating grains of colours $i, j, \;\; i, j = 1, \ldots, K$, and estimate $\widehat{\mathbf{p}}$ as

$$\widehat{p}_i = \frac{\sum_{j=1}^{K} c_{ij} + \alpha}{c_b + \alpha K}\ ,\ i = 1, 2, \ldots, K.$$

Figure 5 displays an enlarged part of a binary image and the lag–pattern used to be evaluated, here to actually count the colours, thus simplifies to ordinary counting of joins.

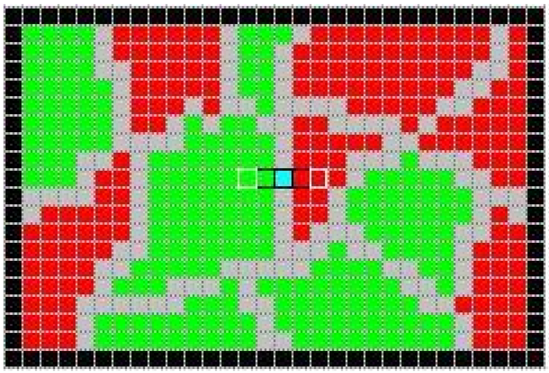

Figure 5: The 1×5–lag pattern with $r = 2$ to count joins parallel to the x–axis in a binary image given in enhanced pixel mode.

The results are displayed in Figure 6.

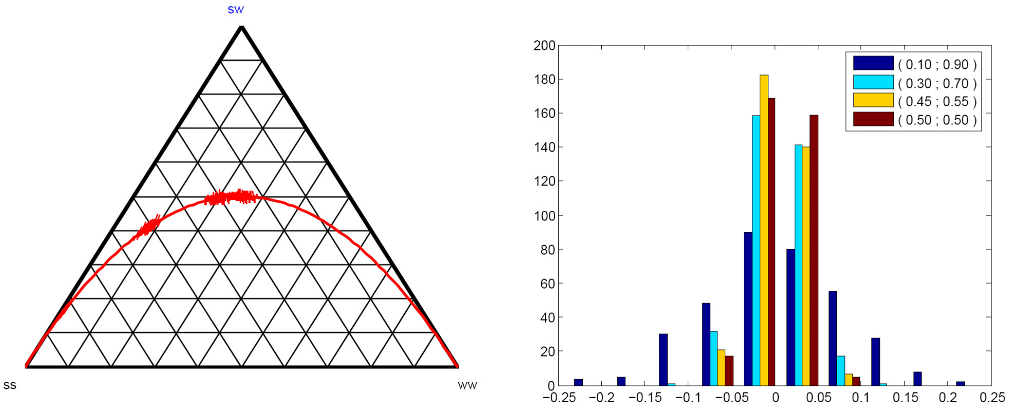

Figure 6: Simulated independet colour coding for several values of p: (A) Joins–counts in the simplex completed by the Hardy–Weinberg manifold (left); (B) Histogram of signed Aitchison distances for several values of p (right).

From the histogram we may conclude that the variance of signed distances increases as p decreases. This seems plausible as changing the colour–coding for just one grain may largely change the distance if p is small while the distance hardly changes if p is close to 0.5 (Fig. 7).

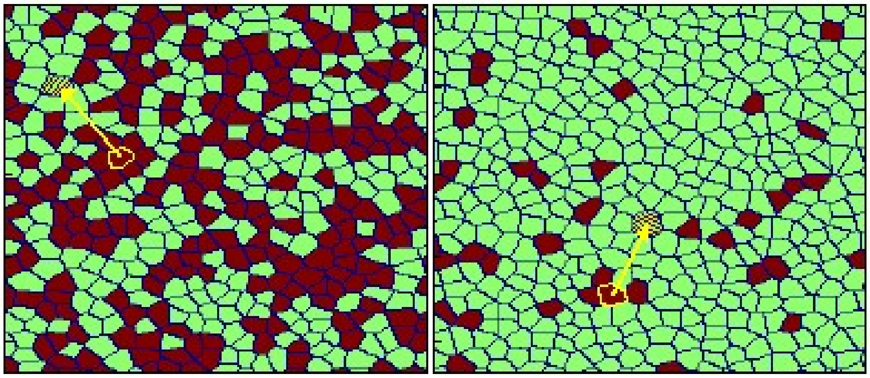

Figure 7: Two simulations: (A) with $p = 0.5$ (left); (B) with $p = 0.1$ (right).

5.2.2 Independence vs. dependence in enhanced pixel mode

Next we analyse three binary images, all three of them simulated with $p = 0.5$ (Fig. 8).

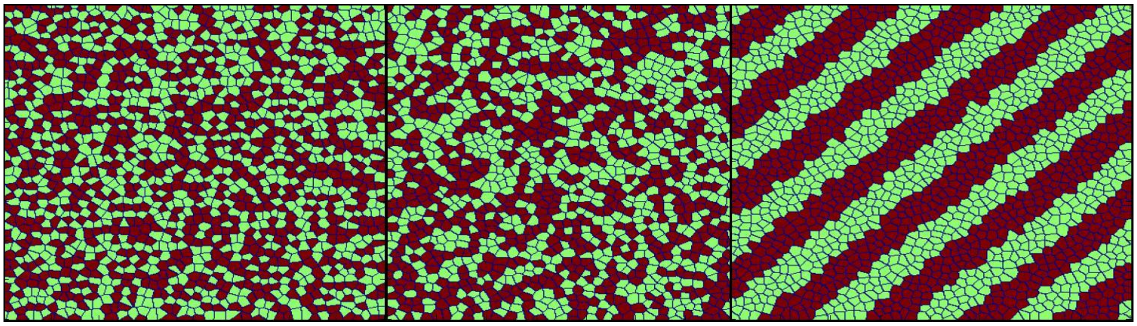

Figure 8: Three simulated binary images for $p = 0.5$: (A) anticlustered (left); (B) independent (centre); (C) clusteres (right).

The numerical results are as follows:
Case (A)
We estimated $\widehat{\mathbf{p}}$ with $\widehat{p}_1 = 0.5088$, $\widehat{p}_2 = 0.4912$, computed

$$\begin{array}{rrrr} \widehat{\mathbf{q}}: & \widehat{q}_{11} = 0.2114 & \widehat{q}_{12} = 0.5949 & \widehat{q}_{22} = 0.1937 \\ \mathbf{m}(\widehat{\mathbf{p}}): & \mathbf{m}(\widehat{\mathbf{p}})_{11} = 0.2589 & \mathbf{m}(\widehat{\mathbf{p}})_{12} = 0.4998 & \mathbf{m}(\widehat{\mathbf{p}})_{22} = 0.2413 \\ \widehat{\mathbf{q}}_{\mathcal{H}}: & \widehat{q}_{\mathcal{H}11} = 0.2610 & \widehat{q}_{\mathcal{H}12} = 0.4998 & \widehat{q}_{\mathcal{H}22} = 0.2392 \end{array}$$

and got the distances

$$d_a(\widehat{\mathbf{q}}, \mathbf{m}(\widehat{\mathbf{q}})) = 0.3149, \;\; d_a(\widehat{\mathbf{q}}, \widehat{\mathbf{q}}_{\mathcal{H}}) = 0.3147$$

Case (B)
We estimated $\widehat{\mathbf{p}}$ with $\widehat{p}_1 = 0.5117$, $\widehat{p}_2 = 0.4883$, computed

$$\begin{array}{rrrr} \widehat{\mathbf{q}}: & \widehat{q}_{11} = 0.2667 & \widehat{q}_{12} = 0.4901 & \widehat{q}_{22} = 0.2432 \\ \mathbf{m}(\widehat{\mathbf{p}}): & \mathbf{m}(\widehat{\mathbf{p}})_{11} = 0.2618 & \mathbf{m}(\widehat{\mathbf{p}})_{12} = 0.4997 & \mathbf{m}(\widehat{\mathbf{p}})_{22} = 0.2385 \\ \widehat{\mathbf{q}}_{\mathcal{H}}: & \widehat{q}_{\mathcal{H}11} = 0.2616 & \widehat{q}_{\mathcal{H}12} = 0.4997 & \widehat{q}_{\mathcal{H}22} = 0.2386 \end{array}$$

and got the distances

$$d_a(\widehat{\mathbf{q}}, \mathbf{m}(\widehat{\mathbf{p}})) = 0.03152, \;\; d_a(\widehat{\mathbf{q}}, \widehat{\mathbf{q}}_{\mathcal{H}}) = 0.031498$$

Case (C)
We estimated $\widehat{\mathbf{p}}$: with $\widehat{p}_1 = 0.5002$, $\widehat{p}_2 = 0.4998$, computed

$$\begin{array}{rrrr} \widehat{\mathbf{q}}: & \widehat{q}_{11} = 0.4159 & \widehat{q}_{12} = 0.1687 & \widehat{q}_{22} = 0.4154 \\ \mathbf{m}(\widehat{\mathbf{p}}): & \mathbf{m}(\widehat{\mathbf{p}})_{11} = 0.2502 & \mathbf{m}(\widehat{\mathbf{p}})_{12} = 0.5 & \mathbf{m}(\widehat{\mathbf{p}})_{22} = 0.2498 \\ \widehat{\mathbf{q}}_{\mathcal{H}}: & \widehat{q}_{\mathcal{H}11} = 0.2501 & \widehat{q}_{\mathcal{H}12} = 0.5 & \widehat{q}_{\mathcal{H}22} = 0.2499 \end{array}$$

and got the distances

$$d_a(\widehat{\mathbf{q}}, \mathbf{m}(\widehat{\mathbf{p}})) = 1.302, \;\; d_a(\widehat{\mathbf{q}}, \widehat{\mathbf{q}}_{\mathcal{H}}) = 1.302$$

While there is no evidence against the null–hypothesis in case A, we are led to reject it for cases A and C with respect to the lag–pattern exposed in Figure 5.

For the second example we analyse two binary images which exemplifiy the many possibilities of deviation from independence yet including instances of independence (Fig. 9).

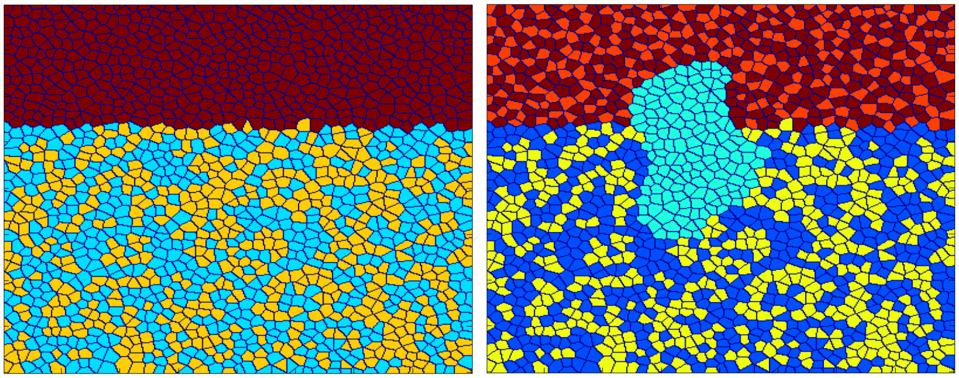

Figure 9: Two simulated images: (A) three colours, colours 1 and 2 are independent with respect to each other, colour 3 does not mix with colours 1 and 2 (left); (B) five colours, colours 1 and 3 are independent with respect to each other, colours 4 and 5 anticlustered, colours 1 and 3 do not mix with colours 4 and 5, colour 2 does not mix with any other colour (right).

For the case of tree colours we compute the distances

$$d_a(\widehat{\mathbf{q}}, \mathbf{m}(\widehat{\mathbf{p}})) = 4.4321, \;\; d_a(\widehat{\mathbf{q}}, \widehat{\mathbf{p}}_{\mathcal{H}}) = 4.3537$$

and for the case of five colours

$$d_a(\widehat{\mathbf{q}}, \mathbf{m}(\widehat{\mathbf{p}})) = 7.6569, \;\; d_a(\widehat{\mathbf{q}}, \widehat{\mathbf{p}}_{\mathcal{H}}) = 7.5632$$

These large distances lead us to infer that the null-hypothesis of indepedence should probably be rejected with respect to the lag–pattern exposed in Figure 5; they do not provide much insight into the actual pattern of deviation from independence. However, the various compostions $\mathbf{q}$ arranged as matrices may provide a more detailed analysis.

For the example with three colours we get

$$\widehat{\mathbf{q}} = \left(\begin{array}{cc|c} \mathbf{0.18636} & & \\ \mathbf{0.31101} & \mathbf{0.16524} & \\ \hline 0.01023 & 0.00748 & \mathbf{0.31966} \end{array} \right) \quad \text{and} \quad \widehat{\mathbf{q}}_{\mathcal{H}} = \left(\begin{array}{ccc} 0.16204 & & \\ 0.29014 & 0.12988 & \\ 0.19086 & 0.17087 & 0.05620 \end{array} \right)$$

indicating almost complete separation of the third color from the others, and no evidence of dependence of colours one and two with respect to the lag–pattern exposed in Figure 5.

For the example with five colours we get

$$\widehat{\mathbf{q}} = \left(\begin{array}{ccc|cc} \mathbf{0.17184} & & & & \\ 0.00678 & \mathbf{0.085621} & & & \\ \mathbf{0.27966} & 0.005469 & \mathbf{0.14365} & & \\ \hline 0.00371 & 0.002020 & 0.00161 & \mathbf{0.02772} & \\ 0.00584 & 0.005496 & 0.00223 & \mathbf{0.19765} & \mathbf{0.06067} \end{array} \right)$$

and

$$\widehat{\mathbf{q}}_{\mathcal{H}} = \left(\begin{array}{ccccc} 0.00994 & & & & \\ 0.08443 & 0.01791 & & & \\ 0.14184 & 0.06020 & 0.05056 & & \\ 0.07868 & 0.03339 & 0.05609 & 0.01556 & \\ 0.12688 & 0.05384 & 0.09045 & 0.05018 & 0.04045 \end{array} \right)$$

confirming and quantifying the visual inspection with respect to the lag–pattern exposed in Figure 5.

5.2.3 Independence vs. dependence in pure pixel mode

Next we analyse two binary images in "pure" pixel mode (Fig. 10) with a knight–shaped lag–pattern with $r = 4$.

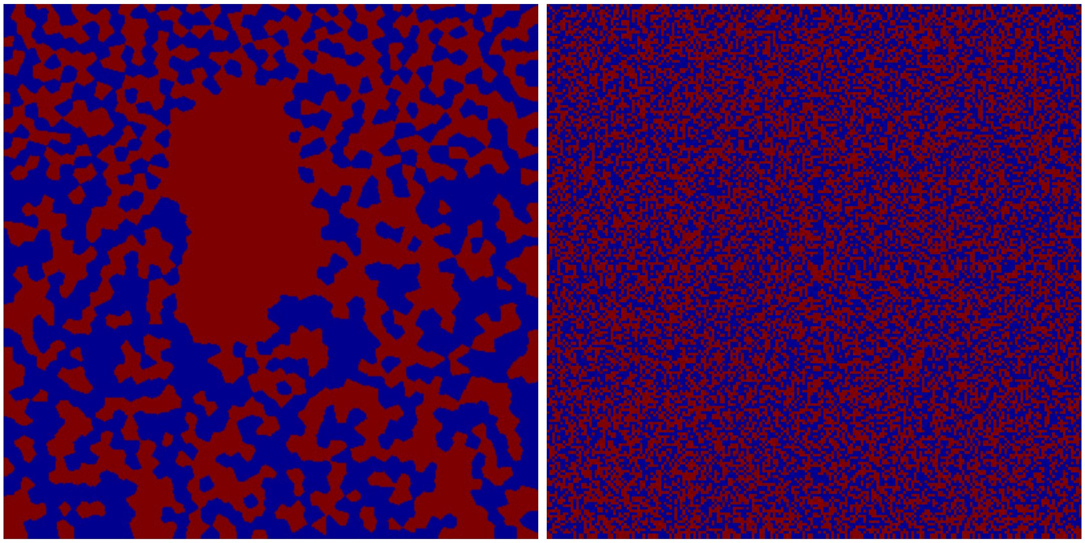

Figure 10: Simualted binary images in "true" pixel mode: (A) clustered,, $p = 0.5$ (left); (B) independent, $p = 0.5$ (right).

Let $\widehat{\mathbf{q}}_i$ denote the relative frequency of $i - 1$ red and $5 - i$ blue pixels within the knight–shaped lag–pattern comprising a total of 4 pixels, and let $\mathbf{m}(\widehat{\mathbf{p}})$ be the corresponding vector of probabilities according to the estimated $\widehat{\mathbf{p}}$ assuming independence.

For case A it is $\widehat{p}_1 = \widehat{p}_{\text{red}} = 0.5339$ and $\widehat{p}_2 = \widehat{p}_{\text{blue}} = 0.4661$. Then

$$\begin{aligned} \widehat{\mathbf{q}} &= (0.3907, 0.0512, 0.0468, 0.0515, 0.4598)^t \\ \mathbf{m}(\widehat{\mathbf{p}}) &= (0.0471, 0.2162, 0.3715, 0.2837, 0.0812)^t \end{aligned}$$

resulting in the Aitchison distance of $d_a(\widehat{\mathbf{q}}, \mathbf{m}(\widehat{\mathbf{p}})) = 4.0466$ suggesting rejection of the null–hypothesis of independence.

For case B it is $\widehat{p}_1 = \widehat{p}_{\text{red}} = 0.5009$ and $\widehat{p}_2 = \widehat{p}_{\text{blue}} = 0.4991$, and

$$\begin{aligned} \widehat{\mathbf{q}} &= (0.0622, 0.2499, 0.3792, 0.2522, 0.0627)^t \\ \mathbf{m}(\widehat{\mathbf{p}}) &= (0.0625, 0.2500, 0.3750, 0.2500, 0.0625)^t \end{aligned}$$

resulting in the Aitchison distance of $d_a(\widehat{\mathbf{q}}, \mathbf{m}(\widehat{\mathbf{p}})) = 0.0117$ indicating missing evidence to reject the null–hypothesis of independence.

6 Conclusions

We introduced a novel test of spatial independence of the distribution of phases (states, colours) based on compositional statistics. In the most simple case it degenerates to joins–count statistics known from map analysis in geographic information systems. Applying a compositional statistical model the case of independent spatial distribution of phases is characterized as Hardy–Weinberg manifold $\mathcal{H}$ as it generalizes the well-known Hardy–Weinberg law of genetics.

Then we suggest to measure deviation from independence by the actual Aitchison distance to the Hardy–Weinberg manifold. Large distances suggest that there is evidence against the hypothesis of independence. Unfortunately, the distribution of this distance in case of independence is not known. Therefore we did not yet succeed to design a test–of–significance.

In case of digital images in pixel mode the method readily applies. Its proper application would consist in a systematic sequence of trials with different lag–patterns until evidence against the hypothesis of independence reveals itself. The most efficient application would be to cast any suspicion of spatial dependence into a lag–pattern specifically designed for the purpose to confirm and quantify the user's suspicion.

In case of digital images displaying grains and their boundaries explicitly ("enhanced" pixel mode) special provision must be developed to account for the grain boundaries. For the time being, applications are confined to the most simple lag–pattern with $r = 2$ when the method degenerates to a consistent version of joins–counts. Generally, properly designed lag–patterns and their evaluation adjusted to the presence of explicit boundaries should detect any kind of dependence analogously to the "pure" pixel mode.

More involved applications including an extension from 2d images to 3d models will be presented in a forthcoming communication.

7 Acknowledgement – A tribute to James Mackenzie

This contribution originates in a PhD project commenced by our colleague James Mackenzie who is missing in the Alps since July 2, 2006. When he was with the Geoscience Department of Basel University, Switzerland, he focused on numerical simulations of micro fabrics to better understand natural ones. Unfortunately, his fate seems to prevent him from completing the project. Our way to pay tribute to James' achievements is to communicate his approach and preliminary results to the scientific community. Any shortcomings are our responsibility.

References

Aitchison, J., Barceló-Vidal, C., Egozcue, J.J., Pawlowsky-Glahn, V. (2002). A concise guide for the algebraic–geometric structure of the simplex, the sample space for compositional data analysis. In U. Bayer, H. Burger and W. Skala (Eds.), *Proceedings of the eigth annual conference of* *the International Association for Mathematical Geology, Volume I and II*, pp. 387–392. Berlin: Selbstverlag der Alfred-Wegener-Stiftung.

Akin, E., Szucs, J.M. (1994). Approaches to the Hardy–Weinberg manifold. *Journal of Mathematical Biology 32*, pp. 633–643.

Billheimer, D., Guttorp, P., Fagan, W. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association 96*(0), pp. 1205–1214.

Bonham-Carter, G.F. (1994). *Geographic information systems for geoscientist.* Pergamon.

Chilès, J.-P., Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty.* J. Wiley & Sons.

Cliff, A.D., Ord, J.K. (1973). *Spatial autocorrelation.* London: Pion.

Cliff, A.D., Ord, J.K. (1981). *Spatial processes.* London: Pion.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*(3), pp. 279–300.

Epperson, B.K., (2003). Covariances among join–count spatial autocorrelation measures. *Theoretical Population Biology 64*, pp. 81-–87.

Fortin, M.-J., Dale, M.R.T. (2005). *Spatial analysis: A guide for ecologists.* Cambrideg: Cambridge University Press.

Hahn, U. (1995). Two phase planar tesselalations – An“infection” model and a test of randomness. In K. Nielsen (Ed.), *Abstracts of the Ninth International Congress for Stereology*, Proceedings of the 9th International Congress on Stereology, Panum Institute, Copenhagen, Denmark, August 20-25. Copenhagen: Stougaard Jensen/Scantryk.

Haining, R.P. (2003). *Spatial data analysis: Theory and practice.* Cambridge: Cambridge University Press.

Hardy, G.H. (1908). Mendelian proportions in a mixed population. *Science 28*(0), pp. 49–50.

Jerram, D.A., Cheadle, M.J., Hunter, R.H., Elliott, M.T. (1996). The spatial distribution of grains and crystals in rocks. *Contributions to Mineralogy and Petrology 125*(0), pp. 60–74.

Kretz, R. (1969). On the spatial distribution of crystals in rocks. *Lithos 2*(0), pp. 39–66.

Mackenzie, J.R., Heilbronner, R., Stünitz, H. (2005). Quantifying the spatial distribution of phases in rocks – a new look at an old approach. *Geophysical Research Abstracts 7*(0).

Mackenzie, J.R., Heilbronner, R., Stünitz, H. (2006). Simulating crystalline microstructures – an aid in quantifying the spatial distribution of phases in two-phase crystalline rocks. *Geophysical Research Abstracts 8*(0).

Mallet, J.-L. (2002). *Geomodeling.* Oxford: Oxford University Press.

Moran, P.A.P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B 37*(0), pp. 243–251.

O'Sullivan, D., Unwin, D.J. (2003). *Geographic information analysis.* J. Wiley & Sons.

Pawlowsky-Glahn, V., Egozcue, J.J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA) 15*(5), pp. 384–398.

Rocklin, S., Oster, G. (1976). Competition between Phenotypes. *Journal of Mathematical Biology 3*, pp. 225–261.

Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. *Jahreshefte d. Vereins vaterl. Naturkunde in Württemberg 64*(0), pp. 369–382.