Principles of an image-based algorithm for the quantification of dependencies between particle selections in sampling studies

by D.S. Dihalu* and B. Geelhoed*

Synopsis

A generalization of Gy's model for the fundamental sampling error introduced the new 'parameter for the dependent selection of particles', denoted as C_{ij}. This allows for modeling deviations from the ideal situation where the selection of a pair of particles is composed of two independent selections. The generalized model potentially leads to more accurate variance estimates in the case of clustering of particles, differences in densities or sizes of the particles or repulsive inter-particle forces. A straightforward and practically applicable method is needed for the determination of this parameter for miscellaneous mixtures in industrial settings.

In this contribution, the feasibility of using digital image analysis to determine this parameter C_{ij}, will be demonstrated. Line transect sampling of a digital image was used to construct a transition probability matrix. A new algorithm to derive quantitative estimates for C_{ij} will be presented and discussed.

The applicability was verified by a photograph of zirconium silicate particles of sizes typical for industries dealing with pharmaceutical, food/feed, and environmental applications. Conditions affecting the practical applicability are identified and potential pitfalls will be discussed, including e.g. how a potential unrepresentative surface can affect the quality of the estimate of C_{ij}.

Introduction

During the sampling of various kinds of particulate materials, there is a risk that the estimated properties of the batch deviate from the true properties of the batch from which the sample was drawn. Therefore, sampling errors introduce a risk of potentially unreliable decisions. A good measure for the potential magnitude of the sampling error is the sampling variance[1].

The most common and widely applied method to estimate the sampling variance is based on the theory reported by Pierre Gy[2]. In this theory, the variance V of the sample concentration c_{sample} can be calculated using the following formula, which is used in the absence of grouping and segregation:

$$V(c_{sample}) = \frac{1-q}{qM_{batch}^2}\sum_{i=1}^{N_{batch}} m_i^2 (c_i - c_{batch})^2 \quad [1]$$

with q = (first-order) inclusion probability
M_{batch} = mass of the batch
m_i = mass of particle i in the population
c_i = concentration of particle i
c_{batch} = concentration of the batch

To account for grouping and segregation, Gy introduced a correction factor $(1+\varepsilon\gamma)$:

$$V(c_{sample}) = (1+\varepsilon\gamma)\frac{1-q}{qM_{batch}^2} \times \sum_{i=1}^{N_{batch}} m_i^2 (c_i - c_{batch})^2 \quad [2]$$

with ε = segregation constant
γ = grouping factor (approximate number of particles in a group)

In Gy's theory, no equations are given to predict the factors ε and γ, based on the properties of the population that is sampled. This suggests that Gy intended these factors (ε and γ) to be empirically determined, based on judgement and previous experience. However, a recent generalization of the underlying core model of Gy's theory was proposed[3], which can deal with grouping, clustering and segregation of particles without having to resort to empirical correction factors (like ε and γ).

In this generalization, it was concluded that one of the parameters that is needed to find the variance is a parameter that indicates to what extent the selection of particles need to be regarded as dependent probabilistic events. Three situations can occur within a batch containing particles of type i and j regarding the parameter for the dependent selection of particles, denoted by C_{ij}, indicating:

(i) $C_{ij} < 0$ grouping of particles of type i and j
(ii) $C_{ij} = 0$ no grouping and no segregation of particles of type i and j
(iii) $0 < C_{ij} \leq 1$ segregation between particles of type i and j.

* *Faculty of Applied Sciences, Delft University of Technology, The Netherlands.*

Principles of an image-based algorithm

It was shown that the sampling variance is given by:

$$V(c_{sample}) = \frac{1}{M_{sample}^2}\sum_{i=1}^{T} N_i m_i^2 (c_i - c_{sample})^2 - \frac{1}{M_{sample}^2}\sum_{i=1}^{T}\sum_{j=1}^{T} N_i N_j C_{ij} m_i m_j (c_i - c_{sample}) \times (c_j - c_{sample}) \quad [3]$$

with T = the number of particle classes
N_i = the (expected) number of particles in the sample belonging to the i-th particle class
m_i = the particle mass of a particle belonging to the i-th particle class
c_i = the mass concentration of the property of interest in a particle belonging to the i-th particle class
M_{sample} = the (expected) mass of a sample
C_{ij} = the parameter for the dependent selection of particles.

It has been discussed that setting $C_{ij} = 0$ in the above equation, leads to a form that is equivalent to Gy's equation (Equation [1]) when the effect of a finite population correction is neglected[3].

In order to let this equation be of use in industry, it is important that its input parameters can be practically determined. Because the generalization introduces a new parameter (C_{ij}), the question arises how to determine this new parameter in practice. Two basic ideas have been put forward to find quantitative values for C_{ij}[4]: (i) a modelling approach based on the known physical particle properties, and (ii) a direct approach based on image analysis in which the spatial distribution of particles is directly observable.

Especially the second method is of interest, because, in practice, digital photographs of mixtures are already taken routinely in many industries, or can be taken without taking too much extra effort. Furthermore, digital photographs have as an advantage that the analysis can be performed at any moment, and these analyzes can be repeated as often as desired. These properties have provoked many studies, e.g. the work done by Korpelainen *et al.*[5], who used the technique of image analysis to evaluate the parameters of an adapted version of Equation [1].

In this work, attention will be paid to the feasibility of using image analysis to determine the parameter for dependent selection of particles. Since this method is based on an image, the analysis is restricted to the particles that are shown within the regions of the picture. Therefore, it is not very likely that this image will be applied to describe 'long-range quality fluctuations'.[2]

This article will thus focus on the evaluation of C_{ij} on basis of an image. Because in some images it can readily be observed by the eye whether there is grouping and segregation, it is expected that precise numerical estimates for C_{ij} can be obtained based on the information contained in such images, if an appropriate and to be constructed mathematical algorithm is used.

In this article, we report the construction of the first mathematical algorithm that can be used to get numerical estimates of C_{ij} based on the information contained in a digital image.

Theory

In this section, the theory underlying the new mathematical algorithm is given and then the algorithm is constructed. The theory is split into four parts, i.e.:

- Transition probabilities
- Markov chain Monte Carlo simulations
- Line transect sampling
- Image analysis

Transition probabilities

When particles of type i and j are grouped ($C_{ij} < 0$), they are more likely to be found close to each other. Conversely, when particles are segregated ($0 < C_{ij} \leq 1$), they are more likely to be further away from each other. A new idea is to describe this effect as well using a transition probability between particle types. In Figure 1, two different strings of particles of two types (LARGE = 1 and SMALL = 2) are depicted. As can be seen directly, the probability of finding a transition from type one to type two or from type two to type one is higher in the picture on the left. Transitions from type one to type one or from type two to type two are more likely to be found in the string on the right. As a consequence, the corresponding transition probabilities will be different in both cases.

Although the parameter C_{ij} is the central new core parameter of the generalization of Gy's model, the theory in this section will make use of the transition probability as an auxiliary parameter.

We report a new mathematical notation here. In this new notation, the transition probability is denoted as $P(A \rightarrow B)$, which indicates the probability of going from state A (start-state) to state B (end state). The state A represents a chain of particles that has been selected so far, while B represents the updated chain after a next particle has been selected.

We also define here the order of a transition probability as the number of particles that are present in state A. The advantage of the innovative notation introduced in this article is that it can be easily used to describe different orders of transition probabilities. In order to clarify the principles of higher orders of transition probabilities, three cases are depicted.

In Figure 2, the first three orders of transition probabilities are illustrated.

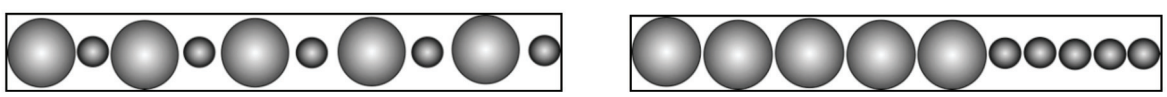

Figure 1—Two strings of particles of different types: LARGE(1) and SMALL (2). In the left string heterogeneous transitions are more likely to be found than in the right string. On the other hand, homogeneous transitions will be found in the right situation more than in the left one

Principles of an image-based algorithm

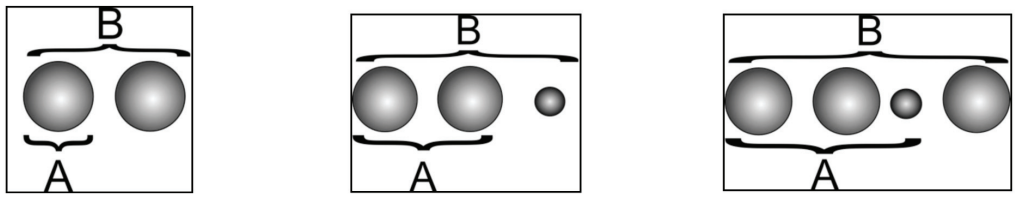

Figure 2—Visualization of the first three orders of transition probabilities: (1) first-order transition probability, (2) second-order transition probability, and (3) third-order transition probability

If sampling is visualized as a string of particle selections, the probability of selecting a next particle is given by a transition probability of order equal to the number of particles so far selected. This implies that very high orders will have to be taken into account, because a sample can in some cases contain 10^6 particles or more. Here, a simplification is proposed to deal with this apparent complexity. The idea is to assume that the probability of going to state B depends only on the last (N+1) particles in state B. In this approximation, the N-th order transition probability is given by the here-proposed symbol $P_N(A \rightarrow B)$. In a mathematical manner, this means that P_N must have following property for the approximation to hold:

$$P_N(A \rightarrow B) = P_N(C \rightarrow D) \approx P(A \rightarrow B)$$

where the chain of the last N particles in states A and C are equal and the chain of the last (N+1) particles in states B and D are equal as well. The symbol '≈' indicates that P_N is an approximation of the true transition probability, P.

It is expected that the approximation becomes more accurate if a higher N is selected. Therefore, it is expected that $P_N(A \rightarrow B)$ becomes a good approximation for the actual transition probability if N is selected high enough.

From now on, we will assume that if two start states end with the same sequence of N particles, they are the same (i.e. A=C) and if two end states end with the same sequence of (N+1) particles, they are also the same (i.e. B=D). It is noted that the function P_N can then be represented by a matrix, containing a number of rows equal to the number of possible states for A, and a number of columns equal to the possible number of states for B and that A and B can be represented by vectors.

As an example, consider a first-order transition probability, P_1, and denote the number of particle classes by T. If the probability of going from state A = (i) to state B = (j,k) is denoted as p_{ijk}, then the matrix P_1 for T=3 can be represented by:

$$P_1 = \begin{bmatrix} p_{111} & p_{112} & p_{113} & p_{121} & p_{122} & p_{123} & p_{131} & p_{132} & p_{133} \\ p_{211} & p_{212} & p_{213} & p_{221} & p_{222} & p_{223} & p_{231} & p_{232} & p_{233} \\ p_{311} & p_{312} & p_{313} & p_{321} & p_{322} & p_{323} & p_{331} & p_{332} & p_{333} \end{bmatrix}$$

Note that p_{ijk}=0 when $i \neq j$. Hence, the above matrix reduces to:

$$P_1 = \begin{bmatrix} p_{111} & p_{112} & p_{113} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{221} & p_{222} & p_{223} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_{331} & p_{332} & p_{333} \end{bmatrix}$$

If state A is a 'chain' of one particle, the vector [1,0,0] represents A=(1); the vector [0,1,0] represents A=(2); and the vector [0,0,1] represents A=(3). If B is a chain of two particles, the vector B will be [f_{11},f_{12},f_{13},f_{21},f_{22},f_{23},f_{31},f_{32}, f_{33}], where f_{ij} = 1 if state B=(i,j) and zero otherwise.

For example, let A=(2), then the vector representation of A is zero everywhere expect the second entry, i.e. A=[0,1,0]. Let B=(2,3), then the vector representation of B=[0,0,0,0,0, 1,0,0,0]. It follows that:

$$P_1(A \rightarrow B) = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} p_{111} & p_{112} & p_{113} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p_{221} & p_{222} & p_{223} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_{331} & p_{332} & p_{333} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = p_{223}$$

The matrix multiplication results in the correct transition probability (p_{223} in this example).

Markov chain Monte Carlo simulations

If the numerical value of the transition matrix $P_N(A \rightarrow B)$ is known, computer simulations of the sampling process are applied to obtain the parameter C_{ij}. The underlying idea of the simulations is to use the relation between C_{ij} and the expected values of N_i and covariances between N_i and N_j: (see e.g. Geelhoed[6]):

$$E(N_iN_j) - E(N_i)E(N_j) = \Delta_{ij}E(N_i) - C_{ij}\,E(N_i)E(N_j)$$

with
- N_i and N_j = the number of particles in the sample belonging to the i-th and j-th particle class respectively
- $E(N_i)$ and $E(N_j)$ = the expected value of the number of particles in the sample belonging to the i-th and j-th particle class respectively
- $E(N_iN_j) - E(N_i)E(Nj)$ is the covariance between N_i and N_j
- Δ_{ij} = the Kronecker delta which is one when i=j and zero otherwise.

The expected value and covariance are estimated by replicated simulation of random and independent samples. This allows assessing a numerical estimate for C_{ij} using the above equation. The steps of the simulation process are:

- Step 1: simulate a (random) sample based on a given transition probability matrix, P_N. Particles are selected until a certain sample mass is reached. The first N particles in the chain are selected at steady state.
- Step 2: record the numbers of particles (N_i) belonging to each particle type in the sample.
- Step 3: repeat steps 1 and 2 many times (at least 10^4 times)
- Step 4: based on the recorded data, estimate $E(N_iN_j)$, $E(N_i)$ and $E(N_j)$.
- Step 5: based on the relation between $E(N_iN_j)$, $E(N_i)$ and $E(N_j)$ and C_{ij} (see above) calculate C_{ij}.

Principles of an image-based algorithm

It can be seen that the algorithm requires as input parameters: the particle masses, the sample mass, and the transition matrix $P_N(A \rightarrow B)$.

The next section will focus on getting $P_N(A \rightarrow B)$ based on an image.

Line transect sampling

In order to extract $P_N(A \rightarrow B)$ from a digital image of a particulate mixture, it is proposed here to make use of the so-called line transect sampling method. This method is also mentioned in literature as 'line intercept sampling' and 'line intersect sampling' (see e.g. Kaiser[7]). Line transect sampling is mainly applied to estimate means (see e.g. Pontius[8]). Although line transect sampling has proven to be a reliable, versatile, and easy to implement method to analyze an area containing various objects of interest[9], it has never been applied to estimating variances during particulate material sampling before.

Note that, in this work, it is proposed to take a line transect sample for estimating c_{ij}. The concentration in a sample, c_{sample}, (see also Equation [3]), will still be obtained from a sample taken in a traditional way.

Line transect sampling of particles on an image

We here propose that, in the simplest case, a straight line is drawn throughout a specified area, and by travelling along this line an ordered collection of particles of interest is observed. We propose to make use of the fact that the order of the particles on the defined line transect indicates how particles of a certain type are located. The underlying idea is that an increased occurrence of finding particles of certain types next to each other will be observed whenever these particle types are clustered. Conversely, a decreased occurrence of finding two particles of certain types next to each other will be observed in the case of segregation of these types.

Every line transect sample starts with initializing a line that goes through the collection of particles by following a predefined path. However, by letting a line going through the whole group of particles, bigger particles will have a higher probability of being transected by the line, leading to a potentially size-biased sample. Ideally, however, each particle must have an equal probability of ending up in the sample that is taken along a specified line. In order to overcome this potential size-biased sampling, a constant size-independent detection radius needs to be defined that surrounds the centre of each particle[7]. Every particle centre is therefore surrounded by a disc with radius equal to the detection radius.

Every particle for which this disc is intersected by the line transect, will end up in the sample, whether or not the particle itself is intersected. This will result in the virtual enlargement or shrinkage of particles and, as a consequence, the effect of size bias is overcome.

As a result, three possible situations can be distinguished when a line is approaching a particle with a surrounding boundary: (i) transection and selection, (ii) no transection and no selection, or (iii) selection, but no transection. The three possible situations are illustrated in Figure 3.

The underlying idea of using line transect sampling to estimate $P_N(A \rightarrow B)$ is to count for each possible transition $A \rightarrow B$ the number of occurrences. Dividing this number of occurrences by the total number of occurrence of state A on the line (i.e. all states that end with the same N particles as state A) will yield an estimate of $P_N(A \rightarrow B)$. The transition probability can thus be estimated with:

$$P_N(A \rightarrow B) = \frac{N_{A \rightarrow B}}{N_A} \qquad [4]$$

where $N_{A \rightarrow B}$ indicates the number of occurrences of going from state A to B and N_A is the number of occurrences of state A that are found on the path taken. For example, if N=1 and state A consist of particle of class i, $N_A = N_i$.

It is clear from this idea that one needs to observe 'enough' occurrences of going from state A to B in order to get a statistically reliable estimate of $P_N(A \rightarrow B)$. We propose here to use, as a first, but crude, approximation, Poisson statistics to estimate the standard deviation of the estimated value of $P_N(A \rightarrow B)$. In other words, the relative standard deviation in the estimate of $P_N(A \rightarrow B)$ is $\sqrt{\frac{1}{N_{A \rightarrow B}}}$. To get a relative standard deviation of at least 10%, one needs to have at least 100 occurrences of the corresponding transition.

We constructed a computer program for automated line transect sampling, implementing the above procedure for obtaining an ordered sample, and for calculation of transition probabilities. However, some additional caution is required to assure that the line transect sampling procedure renders reliable results. Ultimately, there are four practical guidelines formulated here:

- For each transition $A \rightarrow B$, with non-zero $P_N(A \rightarrow B)$, at least 100 instances should be found on the line.
- The line should cover the whole area of the picture as evenly as possible.
- The line needs to be defined such, that there are no points of intersection with itself.

  A path is regarded to be appropriately chosen if the occurrence of particles that are transected more than once is small or even negligible. Otherwise, the true

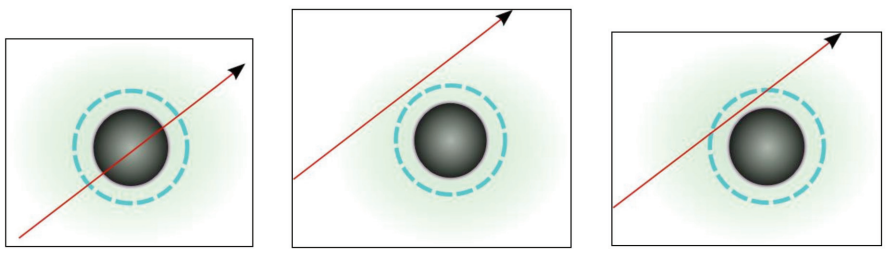

Figure 3—Abstraction of three possible cases that can occur when the line approaches a particle with its detection radius boundary: (i) transection and selection (left), (ii) no transection nor selection (middle), and (iii) selection, but no transection (right)

Principles of an image-based algorithm

ordering of the particles is not reflected. The line should therefore not intersect itself.

- The detection radius should be selected appropriately (not too large or too small), so that the ordering of particles along the line reflects the actual ordering of particles in the image.
- Images should be isotropic (possibly the technique of line transect sampling can also be applied to also study isotropicity ; this is subject to future research). If this is the case, the direction of the line transect will not critically influence the outcome.

At the end of this section, we will go into more detail about the above guidelines, but for the moment, keeping the above practical guidelines in mind, the entire procedure of taking a line transect sample can be described first (see also Figure 4 for three schematic examples):

- First, a line is defined, which may consist of a collection of straight lines.
- Second, a chain of particles is identified, which is the chain of particles that is the (unordered) line transect sample.
- Third, the unordered line transect sample is ordered, based the projection of the particle centres on the line transect. The distance traversed along the line transect can be thought of as the 'time' of selection; particles that are selected 'before' other particles are also put before these other particles in the ordering of the line transect sample.
- In the special case when two particles are selected at the same 'time', the distance between the particle and the line will determine the ordering (smallest distance will be first).
- The different transitions in which one is interested are counted. Based on these counts, $P_N(A \rightarrow B)$ is estimated using Equation[4] .

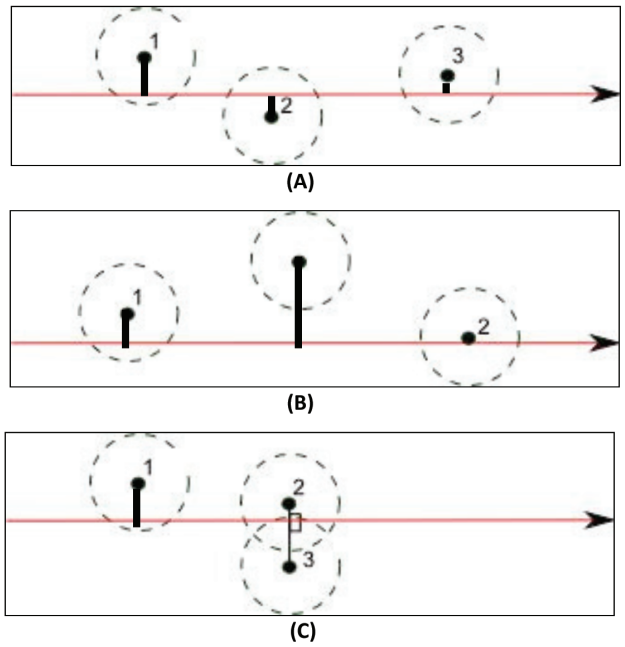

Figure 4—An example of finding the ordering of the line transect sample. Three particles are depicted using their centres and detection discs, and projections on the line transect. A horizontal line is depicted, which represents (part of) the line transect. The example also represents directions of the line transect other than horizontal (i.e. the image can be rotated). The numbers indicate the ordering of the particles, where 1 is the first and 3 is the last particle to be selected. The particle that has its projection first on the line will be selected first (A and B). In the case of particles with the same projection on the line transect, the particle with the shortest distance from its centre towards the line will be selected first (C)

Worked-out example of line transect sampling

An example is given now. In the simple case that is depicted in Figure 5, two types of particles, SMALL (1) and LARGE (2), are arranged randomly (but in a regular packing) and one specific path is followed (direction is shown by the arrow). Furthermore, the boundary of the detection radius that depends on the size of the particle it is surrounding is displayed with dashed lines for every particle.

The line transects the collection of particles and their boundaries and, as a consequence, the following order of particles is included in the sample:

(1, 1, 1, 1, 2, 1, 2, 1, 1, 2, 2, 1, 2, 1, 2, 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1)

Now the elements of the $P_1(A \rightarrow B)$ matrix of this binary mixture can be determined straightforwardly. The first step now is to distinguish and classify the different transitions that are obtained.

In this particular case, four types of transitions are possible, i.e. $[(1) \rightarrow (1,1)]$, denoted by $N_{1 \rightarrow 12}$; $[(1) \rightarrow (1,2)]$, denoted by $N_{1 \rightarrow 12}$; $[(2) \rightarrow (2,1)]$, denoted by $N_{2 \rightarrow 21}$; and $[(2) \rightarrow (2,2)]$, denoted by $N_{2 \rightarrow 22}$. Subsequently, the number of particles of one particle type transected by the line is to be counted, rendering the corresponding values for N_1 and N_2. The obtained values for the parameters are listed in Table I.

At this moment, all the necessary parameters are

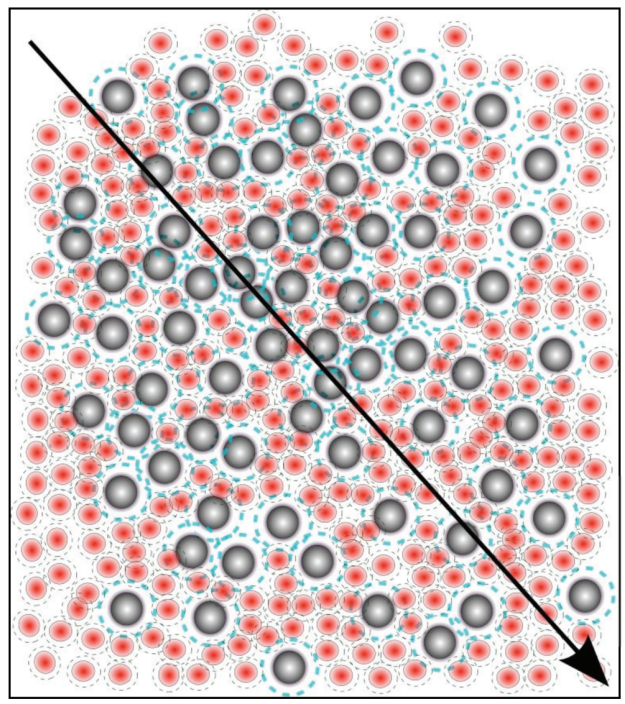

Figure 5—Collection of (1) small particles and (2) large particles together with their (dashed) corresponding detection radius boundaries

Table I

Values for $N_{i\rightarrow ij}$ and N_j for the situation that is examined (as was shown in Figure 5)

N_1	N_2	$N_{2\rightarrow 21}$	$N_{1\rightarrow 11}$	$N_{1\rightarrow 12}$	$N_{2\rightarrow 22}$
22	8	6	15	6	2

Table II

Values of the transition probabilities estimated by application of Equation [4] with the parameters found in Table I

$P_1(1\rightarrow 11)$	0.68
$P_1(2\rightarrow 21)$	0.75
$P_1(1\rightarrow 12)$	0.27
$P_1(2\rightarrow 22)$	0.25

obtained and the transition probability of each transition can be estimated by application of Equation [4]. The transition probabilities that were found are denoted in Table II.

It is noted that the expected size bias is observed, because the ratio of N_1/N_2=22/8 is lower than the corresponding ratio of large particles: small particles in the total image, which has a ratio of N_1/N_2=304/72. Selecting a constant size-independent detection radius will, however, resolve this problem.

The example, although useful to illustrate the concept of estimating transition probabilities using line transect sampling, is also not so realistic because of the low number of observed transitions used to estimate the transition probabilities. This probably resulted in highly inaccurate estimates of transition probability.

Order of transition probabilities

A point of discussion is how large N must be in $P_N(A\rightarrow B)$ so that the approximation $P(A\rightarrow B) \approx P_N(A\rightarrow B)$ is accurate enough. Intuitively, one would expect that the higher N is, the more accurate $P_N(A\rightarrow B)$ approximates $P(A\rightarrow B)$. However, the number of particles in the line transect sample will to some extent be a limiting factor for N. As an example, suppose T=3 and N=1. There are then 3 possible states of A, 9 possible states for B, and 9 possible transitions for which $P_N(A\rightarrow B)$ is non-zero. Increasing N from one to two will result in 9 possible states for A, 27 possible states for B, and 27 possible transitions with non-zero $P_N(A\rightarrow B)$. It can be seen that the number of possible transitions with non-zero $P_N(A\rightarrow B)$ increases with increasing N. If N is increased beyond a certain value, the number of possible transitions will be so high, that given the limited number of particles in the sample, it will not be possible anymore to have 100 occurrences of each possible transition (first practical guideline). The value of N is therefore limited by the finite sample size of the line transect sample.

Image analysis

Before the line transect sample step can be performed, first the coordinates of each particle need to be known. Many algorithms have been developed over the years to adjust images in such a manner that it is possible to extract information from them. With modern image analysis software packages it is possible to account for e.g. gamma correction, shadow, contrast, and contours on a photograph[10]. Furthermore, modern PC performance makes it possible to quickly perform automatic evaluation, in particular the automatic recognition of particles of interest. The particles of interest can be recognized automatically on the basis of their brightness, colour and other descriptive parameters (e.g. particle shape and particle size). The main goal of the use of image analysis software in this work is to automatically find the coordinates of particles of interest. The automatic measurement with the image analysis program consists broadly of three steps that are briefly listed here:

- *Image enhancement*—the optical image enhancement is performed with functions that are available for brightness adjustment, contrast enhancement and gamma correction. In order to improve the automatic recognition performance, all images are smoothed (with a sigma filter[10], illumination errors are corrected by shading correction, and object edges are sharpened by the delineation function of the image analysis software used.
- *Segmentation*—In this step the particles of interest are separated from the background and from each other by clicking on or outlining reference objects. Furthermore, particles are interactively separated and artifacts are deleted.
- *Particle characterization/classification*—the conditions for classifying particles are entered in the program. This is a kind of 'object filter'[10], which can be used, for example, to specify that only objects within a certain size category should be counted. The program is then given the instruction to record the centre coordinates of each particle and the class.

Overall procedure

Schematically, the three stages of the automated line transect sampling methodology are:

(1) Find coordinates of particles
(2) Take line transect sample and calculate the transition probability $P_N(A\rightarrow B)$
(3) Calculate the parameter for the dependent selection of particles c_{ij}

For the first part (1), image analysis software was used that can read in digital photographs, and has good analysis options. An algorithm was available that can automatically select and classify particles of various size and colours. One of the outputs of this program are the coordinates of each individual particle. Subsequently, the coordinates of each particle were used for step 2 of the procedure.

For the line transect sample taking step (2), a program was constructed that makes use of the coordinates of the separate particles. A line transect that can be specified by the

Principles of an image-based algorithm

user beforehand will be drawn throughout the field of coordination points and the order and type of particles is collected. Finally (3), the dependent selection of particles parameter was determined with the mathematical algorithm that was constructed here and explained earlier.

Experimental part

An experimental illustration of the overall procedure previously described is presented in this section. An image is selected of particles with sizes that are typical for mixtures used in industries dealing with pharmaceutical, food/feed, and environmental applications. In the experimental results section, both the transition probabilities and the parameter for the dependent selection of particles are evaluated and discussed. Finally, the variance is calculated with Equation [3].

Experimental procedure

The picture that was analyzed can be found below (Figure 6). Although here the same material is used as in other work[11], it is noted that this picture was not taken during the experimental work performed there.

From this photograph there appears to be clustering of small particles probably due to the different particle sizes, and one can even see a network of lanes of small particles. Hence, based on this visual appearance, one would expect to find a significantly higher value for P(1→11) than for P(2→21).

The mass of a small particle was taken as 2 mg (m_1= 2 mg) and the mass of a large particle was taken as 20 mg (m_2=20 mg). In total a sample was drawn of 200 g (M= 200 000 mg). The concentration of each small particle was considered to be zero (c_1=0), whereas the concentration of every large particle was considered to be one (c_2=1). The average number of particles of type one is 50 000 (N_1=50 000), while the average number of particles of type two was taken as 5 000 (N_2 = 5 000). In this binary mixture, T has the value of two (T = 2) and c_{sample} is approximately 0.50. The detection radius here was taken as the mean radius of the large particles.

The zirconium silicate mixture of Figure 6 was analyzed as follows. First, a specified line transect path was taken throughout the total area (see Figure 7) and then the line transect sample was collected. Subsequently, the transition probabilities were estimated and the corresponding dependent selection for particles was evaluated. In this case, the small particles were considered to be of type 1, whereas the big particles were classified as type 2.

Experimental results

The picture of the zirconium silicate mixture (Figure 6) that was analyzed contained 12 030 particles in total. A line transect sample containing 2 225 small and 560 large particles was obtained. The small particles are regarded to be of 'type 1', while the large particles are of 'type 2'. The first-order transition probabilities were determined (see Table III).

With computer simulations of the sampling process, the corresponding parameters for the dependent selection of particles were found, which are listed in Table IV.

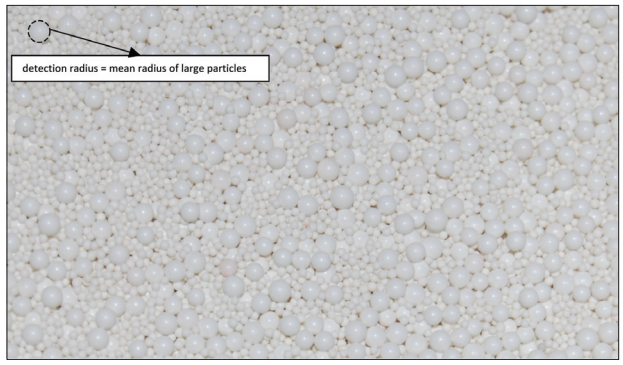

Figure 6—Zirconium silicate particles of several sizes (enlarged excerpt of the original photograph). As is indicated in the box, the detection radius that was taken during this analysis was chosen to be as large as the mean radius of the large particles

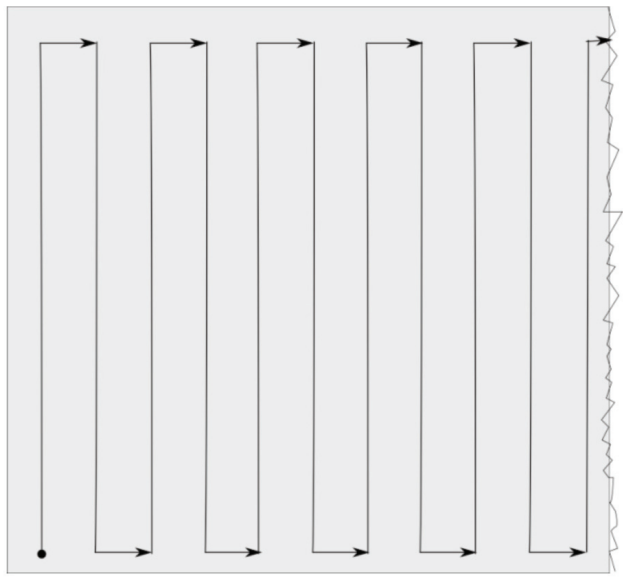

Figure 7—Overview of the line transect that was taken in the picture. The total length of the path depends on the width of the picture

Table III

The first-order transition probabilities found for the line transect sample generated from the zirconium silicate picture

$P(1\rightarrow11)$	0.799
$P(2\rightarrow21)$	0.798
$P(1\rightarrow12)$	0.201
$P(2\rightarrow22)$	0.202

Table IV

The C_{ij} values found for the line transect sample generated from the zirconium silicate picture

C_{11}	-5.50×10^{-5}
$C_{12}=C_{21}$	3.58×10^{-5}
C_{22}	1.25×10^{-4}

Principles of an image-based algorithm

Contrary to the appearance of clustering of small particles in the photograph depicted in Figure 6, $P(1 \rightarrow 11)$ is not significantly larger than $P(2 \rightarrow 21)$.

The variance, V_{LTS}, can be calculated using Equation [3]:

$V_{LTS} = 1.39 \times 10^{-5}$

Discussion

In addition to the practical guidelines that were given for the line transect sample, more improvements can be implemented. First of all, by letting the line be as large as possible, this line will resemble more the 'true' distribution of the particles on the picture. If it can be realized in practice, more than 100 of all possible combinations should be on the line.

Besides straight lines, it could be chosen to sample the area by taking other paths, e.g. spider web, growing circles, etc.

One possible solution that will always hold for the second and third practical guideline implicitly is to take a path that is constructed along the so-called travelling salesman problem[12]. The main advantage of this path is the fact that each particle on the picture will end up in the sample, together with information on its neighbouring particle, and thus enabling the necessary information about the transition probabilities based on all the particles on the picture.

The method presented in this work for evaluating C_{ij} in view of estimating the sampling variance using Equation [3] relies on an image of a representative surface of the population to be sampled. The arrangement of particles will have an influence on the resulting variance prediction. It is possible, however, that certain sources of variation manifest themselves after the image was taken. These sources of variation will possibly influence the sampling variance, but are not taken into account by the method proposed here. These issues could introduce additional sources of variation that cannot be seen on a static photograph. In those cases, application of the line transect sampling technique should be complemented with the assessment of the additional variances introduced after the image was taken.

A final remark that has to be made here is that the line transect sample is for estimating C_{ij} only. The sample which is used to obtain c_{sample} will still need to be taken in a traditional manner.

Results and conclusion

The main result of this work is a new algorithm that can use the information in an image to numerically evaluate the parameter for the dependent selection of particles. In order to construct the algorithm, attention was paid to higher order transition probabilities and a new notation was proposed.

The algorithm consists of the following steps:

(1) Find coordinates of particles
(2) Take line transect sample
(3) Calculate the transition probability $P_N(A \rightarrow B)$ and the parameter for the dependent selection of particles C_{ij}

Implementation of the value for the parameter of the dependent selection of particles can then be used to calculate the sampling variance with the generalization of the model of Gy[3].

Two images were analyzed using the procedure developed here for illustrative purposes. It has there been shown that the novel mathematical algorithm can be used in a direct way to analyze images of mixtures of particulate materials for the grouping and segregation of the different particle types present.

Acknowledgement

This work was performed as part of a project supported by the Netherlands Technology Foundation STW, under STW grant 7457. NFI, Deltares, Nutreco, Hosokawa Micron B.V. and Organon are members of the users' committee of this project.

References

1. Geelhoed, B. The construction of variance estimators for particulate material sampling, arXiv:1005.2968v1 [stat.AP], http://arxiv.org/pdf/1005.2968v1.pdf, 2008.

2. Gy, P. *Sampling of particulate materials, theory and practice*, Elsevier, Amsterdam, 1979, 1982.

3. Geelhoed, B. A generalisation of Gy's model for the fundamental sampling error. *Second World Conference on Sampling and Blending*. The Australasian Institute of Mining and Metallurgy, ISBN 1-920806-28-8, 2005. pp. 19–25.

4. Geelhoed, B. Variable second-order inclusion probabilities as a tool to predict the sampling variance. *Third World Conference on Sampling and Blending*. Porto Alegre, 2007. pp. 82–9.

5. Korpelainen, M., Reinikainen, S-P., Laukkanen, J,. and Minkkinen, P. Estimation of Uncertainty of Concentration Estimates Obtained by Image Analysis, *Journal of Chemometrics*, vol. 16, 2002. pp. 548–554.

6. Geelhoed, B. Variable second-order inclusion probabilities during the sampling of industrial mixtures of particles, *Applied Stochastic Models in Business and Industry*, vol. 22, 2006. pp. 495–501.

7. Kaiser, L. Unbiased Estimation in Line-Intercept Sampling, *Biometrics*, vol. 39, 1983. pp. 965–976.

8. Pontius, J. Estimation of the mean in line intercept sampling. *Environmental and Ecological Statistics* 5, 1998. pp. 371–379.

9. Buckland, S.T. *Introduction to distance sampling: estimating abundance of biological populations*, New York, Oxford University Press; 2001.

10. Carl Zeiss Axiovision User's Guide Release 4.7 2008.

11. Geelhoed, B., Koster-Ammerlaan, M.J.J., Kraaijveld, G.J.C., Bode, P., Dihalu, D.S., and Cheng, H. An experimental comparison of Gy's sampling model with a more general model for particulate material sampling. *Fourth World Conference on Sampling and Blending*, Cape Town, South Africa. WCSB4 Conference Proceedings, 2009. pp. 27–38.

12. Applegate, D.L. *The Traveling Salesman Problem*, Princeton University Press 2006. ◆