

GMOs analysis in large kernel lots: modelling sampling of non-randomly distributed contaminants

C. Paoletti¹, M. Donatelli², E. Grazioli¹ & G. Van den Eede¹

¹European Commission, DG JRC, IHCP, B&GMOs Unit, Via E. Fermi 1 – 21020 Ispra (VA), Italy. E-mail: claudia.paoletti@jrc.it

²Research Institute for Industrial Crops (ISCI), Via di Corticella, 133 – 40128 Bologna, Italy

Introduction

The definition of internationally harmonized strategies for the evaluation of GMO safety is a priority and there is a strong interest in sampling schemes to ensure accuracy and precision of GM surveys. Our work focuses on one critical aspect of GMO control: the definition of sampling protocols for GMO detection and/or quantification. Several guidelines defining sampling strategies for purity analyses are adopted for the detection of GM materials (Kay and Paoletti, 2001), while waiting for the *ad hoc* protocols.

Kernel lot sampling is a complex multi-stage procedure that should reduce a lot to an analytical sample, of suitable working size, representing the lot. Most kernel sampling plans are based upon the assumption of random distribution of GMOs so that the mean, the standard deviation and both the producer and consumer risks can be estimated according to the Binomial or the Poisson distributions (Remund et al., 2001). Given the high likelihood of non-detectable *strata* of GM material in kernel lots (Lischer, 2001), assuming randomness is very risky because even modest deviations from randomness have a strong effect on the accuracy (GMO %) and precision (variance of GMO %) of GMO estimates (Paoletti *et al.*, 2003).

Here we present a model to estimate the sampling error associated to different sampling protocols, in terms of both number and size of samples taken from the lot (primary samples), applicable to any consignment of particulate material with respect to any kind of contamination, including GMOs. The novelty of our approach is the freedom from any distribution constraints.

Results

Preliminary results from our simulations done with KeSTE (Kernel Sampling Technique Evaluation <http://www.sipeaa.it/ASP/ASP2/KeSTE.asp>; Paoletti *et al.*, 2003) indicated that the pattern of convergence to the “true” contamination value is similar for different contamination levels and lot heterogeneity scenarios. Specifically, the spread of the

simulation results (*SD* - standard deviation of the estimates) for each sampling numerosity (number of primary samples taken from a lot) provides an indication of the possible sampling error (*SE*). When a given population is sampled multiple times for an increasing number of primary samples, the error associated to the contamination estimates in the bulk sample decreases. Fig. 1 shows such decrease for 50 repeated samples at each sampling numerosity.

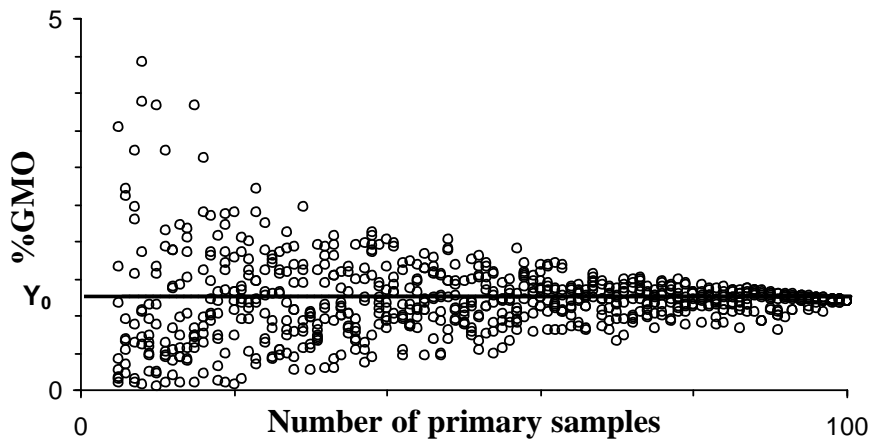


Figure 1

The presence of a common convergence pattern allows the definition of a model to estimate the maximum possible *SE* associated to any sampling numerosity. We found that a negative exponential model $SD = h e^{(-s\sqrt{x})}$ (*SD* = standard deviation, *x* = number of primary samples) best describes the decreasing trend of *SD* as a function of the increasing number of primary samples (Figure 2). The parameter *h* is indicative of lot heterogeneity and *s* is associated to primary samples characteristics. The level of lot heterogeneity will determine how large is the error and how rapidly *SD* converges to 0.

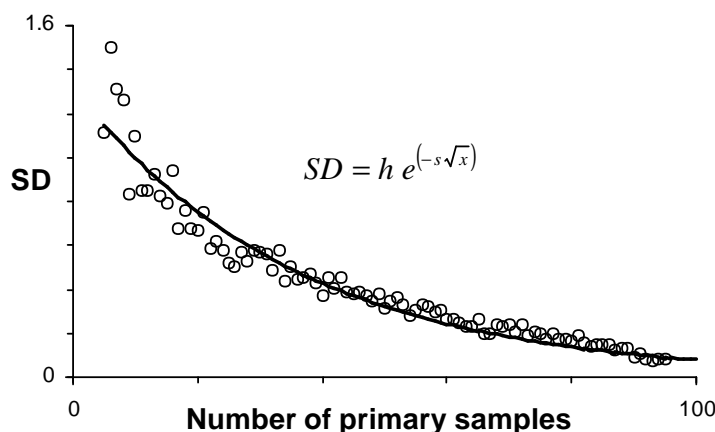


Figure 2. Standard deviation: exponential model

Similarly to *SD*, also the largest values of contamination estimates (dominant values) obtained performing repeated sampling at each sampling numerosity, show a constant pattern for different contamination and heterogeneity scenarios, converging asymptotically to the true lot

contamination value. Such pattern is described by the exponential model $y = y_0 + h e^{(-s\sqrt{x})}$, where $y_0 =$ true lot contamination (Fig. 3). Non-representative bulk samples can also arise in case of false-negative bulk samples, which are a consequence of insufficient sampling rate in case of heterogeneous material distribution. For any given sampling numerosity false-negative probability increases as function of lot heterogeneity. Our model allows estimating the probability distribution of false-negative bulk samples associated to any distribution and lot property scenario. For any given level of lot heterogeneity, such probability decreases increasing the sampling rate (Fig. 4).

The exponential SD curve is more robust compared to the exponential contamination curve given that it converges to 0, implying independency from the y_0 parameter, and it does not require selection of dominant values at each sampling numerosity. In addition, provided that lot characteristics were *a priori* defined in our simulations, we could impose the known value of y_0 and estimate parameters h and s for both the SD and contamination curve, for a series of different contamination and heterogeneity scenarios.

In our simulation primary samples characteristics were maintained constant. As a result, we could use the average value of s (calculated over a broad range of heterogeneity and contamination conditions) to re-estimate h . This improves the precision of h estimates because h and s show correlation in both exponential models. Once the two series of h parameters are estimated, h_Y can be expressed as a function of h_{SD} (Figure 5) and, using the average value of s_Y , y_0 can be estimated with the greatest precision.

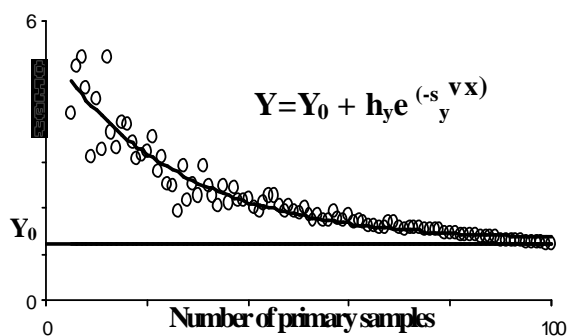


Figure 3. Dominant values: exponential model

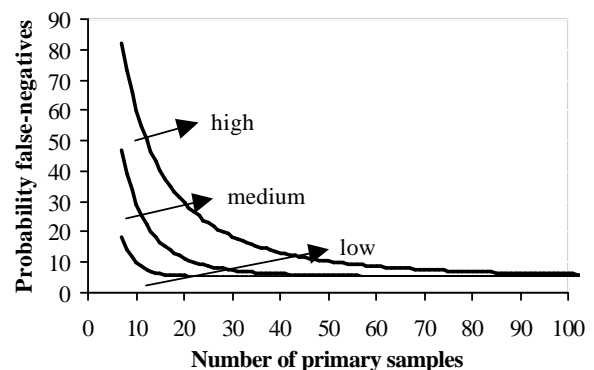


Figure 4. Probability false-negative bulk samples for 3 heterogeneity levels

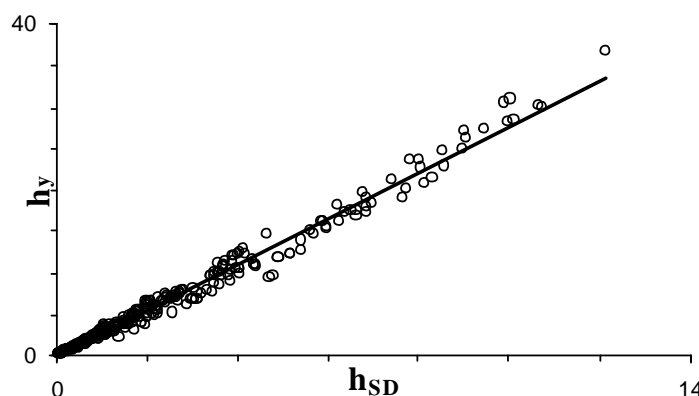


Figure 5. h_Y vs h_{SD} Proceedings GMCC Conference page119-122
November 13-14 2003, Denmark

Perspectives

In this paper we present the heterogeneity model and the preliminary results of our investigation on the stability of heterogeneity parameters. Also, the estimation of the probability distribution of false-negative bulk samples associated to any distribution and lot features scenario is explored; such risk will never be null. This implies that users must set a risk threshold they are willing to accept.

The possibility of expressing h_y as function of h_{SD} allows estimating y_0 with greater precision because of the better stability of parameter estimates achievable with the more robust SD exponential model. Indeed, h_{SD} is a crucial parameter of our model because it allows the best possible estimate of y_0 (i.e. the lot contamination).

Our next goal is to define the minimum number of primary samples necessary to maximize h_{SD} estimates robustness. This will have a remarkable impact upon the definition of sampling protocols, as it will ensure a proper sampling numerosity if non-random distribution of contaminants is observed or expected, as in the case of kernel lots.

References

- Paoletti C., Donatelli M., Kay S. & Van den Eede G. 2003. Simulating kernel lot sampling: the effect of heterogeneity on the detection of GMO contamination. *Seed Science and Technology* 31(3): 629-638.
- Lischer P. 2001. Sampling procedures to determine the proportion of genetically modified organisms in raw materials. Part I: Correct sampling, good sampling practice. *Mitt. Lebensm. Hyg.* 92: 290-304.
- Kay, S. Paoletti, C. 2001. Sampling strategies for GMO detection and/or quantification. European Commission Report, Code EUR20239EN, Joint Research Centre.
- Remund, K., Dixon D., Wright D., Holden L. 2001. Statistical considerations in seed purity testing for transgenic traits. *Seed Science Research*, 11:101-120.