Improving vineyard sampling efficiency via dynamic spatially explicit optimisation

J.M. MEYERS[1], G.L. SACKS[2], H.M. VAN ES[3] and J.E. VANDEN HEUVEL[1]

[1] Department of Horticulture and [2] Department of Food Science and Technology, Cornell University, 630 W. North Street, Geneva, NY 14456, USA

[3] Department of Crop and Soil Sciences, Cornell University 235 Emerson Hall, Ithaca, NY 14853, USA

Corresponding author: Dr James M. Meyers, fax +914 373 4112, email jmm533@cornell.edu

Abstract

Background and Aims: Environmental variables within vineyards are spatially correlated, impacting the economic efficiency of cultural practices and accuracy of viticultural studies that utilise random sampling. This study aimed to test the performance of non-random sampling protocols that account for known spatial structures ('spatially explicit protocols') in reducing sampling requirements versus random sampling.

Methods and Results: Canopy microclimate data were collected across multiple sites/seasons/training systems. Autocorrelation was found in all systems, with a periodicity generally corresponding to vine spacing. Three spatially explicit sampling models were developed to optimise the balance between minimum sample sizes and maximum fit to a known probability density function. A globally optimised explicit sampling (GOES) model, which performed multivariate optimisation to determine best-case sampling locations for measuring fruit exposure, reduced fruit cluster sample size requirements versus random sampling by up to 60%. Two stratified sampling protocols were derived from GOES solutions. Spatially weighted template sampling (STS) reduced sampling requirements up to 24% when based on probabilistic panel weighting (PW), and up to 21% when preferentially selecting specific locations within canopy architecture (AW).

Conclusions: GOES, PW STS and AW STS each reduced required sample size versus random sampling. Comparative analyses suggested that optimal sampling strategies should simultaneously account for spatial variability at multiple scales.

Significance of the Study: This study demonstrates that dynamically optimised sampling can decrease sample sizes required by researchers and/or wineries.

Keywords: *autocorrelation, canopy management, cluster exposure, heuristic algorithm, sampling strategy*

Introduction

Viticulturists collect quantitative information of vine characteristics and fruit composition to understand and guide cultural practices. However, vineyards are heterogeneous and many key viticultural variables vary significantly throughout a domain (e.g. a block). Vine-level spatial autocorrelation in microclimatic canopy variables have been measured by enhanced point quadrat analysis (EPQA; Meyers and Vanden Heuvel, 2008), and these variables correlate with flavour compounds in winegrapes (Meyers et al. 2009). Block-level spatial patterns in vineyards have been identified and quantified for a variety of variables, including canopy fill (Bramley and Hamilton 2004), yield (Bramley and Hamilton 2004, Taylor et al. 2005) and vine water status (Acevedo-Opazo et al. 2010).

To address inherent variability within a block or vineyard, viticulturists typically pool together multiple berries or clusters to improve accuracy, and test samples in replicate to evaluate the precision of their measurements and facilitate statistical comparisons. Both commercial growers and researchers aim to minimise the number and size of field samples to reduce labour and material costs while still achieving acceptable accuracy and precision. Many sampling protocols have been suggested for collecting fruit from vineyards, with the majority of approaches involving random selection of clusters from either the whole vineyard or a vineyard subsection. Some sampling protocols involve the use of spatial patterns to ensure that fruit is sampled from all regions of an individual vine (Rankine et al. 1962), or balanced across sides of a row (Iland et al. 2004).

From a statistical perspective, the ideal protocol should result in a representative sample of the population, such that variables are sampled with a distribution similar to that of the population. The justification for random sampling is that variables contributing to the variance of a population are expected to be mutually independent with respect to their location in time and space, i.e. independent and identically-distributed (*i.i.d.*) random variables. Under these circumstances, any arbitrary set of sampled field measurements is equivalent to any other set (of equal size) in its ability to estimate population parameters, hence the standard use of random sampling in viticultural experiments.

The realisation that random variables are not *i.i.d.* for most agricultural plots (Student 1938, Jeffreys 1939) has led to the introduction of blocking and replication to compensate for potential spatial patterns. In modern agricultural experiments, this compensation often takes the form of randomised complete block designs (RCBD) and their variants. For the most part, these techniques are applied to experimental designs without attempting to gain prior knowledge of existing field patterns

doi: 10.1111/j.1755-0238.2011.00152.x

(van Es et al. 2007), despite the demonstration that compensation for known systematic variability can improve precision in hypothesis testing (Kirk et al. 1980, Tamura et al. 1988; Panten et al. 2010). Considering that spatiotemporal dynamics of canopy architecture create localised ecophysiological variability (Xu et al. 2008), sampling methods should ensure that variability is captured. In viticultural literature, some authors have suggested using systematic approaches to sample berries from all parts of the cluster as opposed to randomly selecting berries (Roessler and Amerine 1958, Rankine et al. 1962, Iland et al. 2004) to compensate for variability in ripeness, but vineyard sampling protocols that account for known spatial heterogeneity (i.e. 'spatially explicit') have not been thoroughly explored.

An alternative to RCBD, the spatially balanced complete block (SBCB) design, was developed to assist researchers in minimising the effects of unknown block-level spatial trends (van Es et al. 2007) through the use of static block-layout templates. Because they are designed to maintain spatial balance among treatments in factorial experiments, SBCB designs are superior to RCBD in their ability to protect against the adverse effects of unknown field trends such as those reported in vineyards (Bramley and Hamilton 2004, Taylor et al. 2005; Meyers and Vanden Heuvel 2009; Acevedo-Opazo et al. 2010). The concept of spatially explicit design can be extended to field sampling practices to maximise the fit of a sample distribution to population statistics, such that a smaller spatially explicit sample should achieve the same accuracy and precision as a larger random sample, particularly when the population probability density is known.

For example, a grower or researcher studying the effects of cluster exposure on fruit composition may wish to perform extensive chemical analyses on fruit sample. Ideally, sampling protocols should minimise sample size, but also preserve the population probability density function (PDF) to allow for non-linear or discontinuous relationships among experimental variables. TDN (TDN = 1,1,6-trimethyl-1,2-dihydronaphthalene), for example, has been shown to respond to cluster exposure in Riesling only at levels above 20% of ambient light (Gerdes et al. 2002). Thus, a researcher studying physiological responses of aroma compounds may be better served to preserve PDFs when sampling while, similarly, a winemaker who wishes to control chemical profiles within tight tolerances may be better served by selectively harvesting towards particular PDF targets rather than mean values.

In principle, a large sample size could be used to overcome high variation in the population, but in practice, most sampling protocols recommend a modest number of clusters out of practical considerations (e.g. logistics associated with sample transport and sample preparation). If cluster light environment in a vineyard has been previously characterised for a much larger number of clusters, then this information can be used to define a purposive model-based sampling method (as described by De Gruijter et al. 2006) to ensure that any subsequent sample is optimally representative of the population.

Stratified sampling, the process of balancing sampling frequency among mutually exclusive strata, has been demonstrated to improve sampling efficiency by reducing standard error of samples (Lark and Marchant 2009). However, this approach is generally used to estimate a population mean and standard deviation and is not intended to preserve an existing, potentially non-normal, probability density function. In the example presented, estimation is not required because the population is known. In this case, the sample can be fully optimised without the uncertainty of estimation. As advances in technology continue to improve density and precision of vineyard monitoring (e.g. three-dimensional canopy imagery), parameter estimation may become less necessary while the resulting increase spatial resolution will facilitate more precise methods of characterisation and cultural control.

For our initial investigations into optimised spatially explicit sampling protocols focused on microclimatic variables, as canopy microclimate is widely reported to influence many aspects of both grape and wine composition. Using natural variation in fruit cluster sunlight exposure in Riesling (*Vitis vinifera* L.) and Vignoles (*Vitis* sp.) populations from the Finger Lakes region of the state of New York, we determined the maximum potential benefit in reducing sample size using spatially explicit globally optimised (i.e. cluster-specific) sampling methods versus randomised sampling methods. In the interest of convenient application, we considered the use of univariate 'sampling template' strategies where clusters were preferentially sampled from specific regions within each panel, or from particular panels within the vineyard. We then compared the performance of these sampling template strategies against the global optimum and against randomly sampled clusters.

Materials and methods

Vine material

Three commercial New York vineyards (Finger Lakes region, east side of Seneca Lake) designated as domain 'A', 'B' and 'C' were used for this study. From domain A, 66 Riesling vines (22 three-vine panels) trained to two-tier flatbow (as described by Reynolds and Vanden Heuvel 2009) with vertical shoot positioning were selected for consistency (i.e. no missing vines or obviously young replants) from an 11-row/77-panel subplot. From domain B, 72 Scott Henry-trained (as described by Reynolds and Vanden Heuvel 2009) Riesling vines (18 four-vine panels) were selected for consistency from a 7-row/72-panel subplot. From domain C, 96 Vignoles (*Vitis* sp.) vines (24 four-vine panels) trained to high-wire umbrella kniffen (as described by Reynolds and Vanden Heuvel 2009) were selected at random from an 8-row/96-panel subplot. Selected panels from all domains were generally not contiguous. Vine spacing was 280 cm (row) × 200 cm (vine), 280 cm × 200 cm and 280 by 180 cm for domains A, B and C, respectively. Data were collected within each domain for two consecutive seasons (2008 and 2009 for domains A and B; 2007 and 2008 for domain C) using the sample panels in consecutive seasons.

Canopy density and cluster exposure characterisation

EPQA and calibrated exposure mapping (CEM; Meyers and Vanden Heuvel 2008) were performed at the onset of veraison (18 July 2007 and 15 July 2008 for Vignoles; 25 August 2008 and 20 August 2009 for Riesling). A tape measure was used as a guide for insertions, which were made at 20 cm intervals along the length of the panels at the height of the fruiting wire, resulting in 30, 39 and 35 insertions per panel at domains A, B and C, respectively. A Decagon AccuPAR LP-80 photosynthetically active radiation sensor (Decagon Devices, Pullman, Washington, USA) was used to measure percent photon flux values used in canopy calibration. The EPQA metrics occlusion layer number (OLN), cluster exposure layer (CEL) and cluster exposure flux availability (CEFA) were calculated as measures of canopy biomass density, un-calibrated cluster exposure and calibrated cluster exposure, respectively, using Microsoft Office Excel version 12.0.6514.5000 SP2 (Microsoft Corporation, Redmond, Washington, USA) and EPQA-CEM Tools version 1.6.2 (available on request from jmm533@cornell.edu). A separate dataset was computed and maintained for each domain.

Quantification of canopy autocorrelation

Autocorrelation in OLN, CEL and CEFA was calculated in 20 cm lag distance increments along the length of each panel using the *autocorr* function from the MATLAB add-on Econometrics Toolbox version 1.1 (The Mathworks, Natick, Massachusetts, USA). Results were averaged across all panels in each domain-year combination and plotted to compare the autocorrelation patterns between seasons and among the three tested EPQA metrics. To quantify potential repetitive spatial patterns in canopy architecture, Fourier series and signal periods were computed for each EPQA metric, as a function of distance along canopy row. Primary signal periods were computed for the purpose of exploring the relationships among periodic cultural practices (i.e. trunk spacing) and spatial structure. Secondary signal periods were computed to explore the potential influence of effects of structural features with periods either shorter or longer than trunk spacing. Discrete Fourier transforms were computed using MATLAB's discrete Fourier transform function, *fft*, and additionally processed via custom software to determine signal periods.

Simulated sampling using real field data

A fruit cluster contact database for each domain-year combination was created by exporting the calculated EPQA-CEM values as measured at veraison (i.e. OLN, CEL and CEFA) along with location information (vineyard row number, panel number and EPQA insertion position) to a text file containing a unique record for each cluster. The six resulting cluster inventories (three domains in each of 2 years) contained 706 and 819 clusters from 66 vines at domain A in years 2008 and 2009, respectively; 1178 and 967 clusters from 72 vines at domain B in years 2008 and 2009, respectively; and 591 and 490 clusters from 96 vines at domain C in years 2007 and 2008, respectively. These databases were imported into custom-written MATLAB software (Version 7.10.0.499, The Mathworks; custom software available on request from jmm533@cornell.edu) designed to perform methods described below. The cluster inventory for each domain assumed the role of that domain's experimental population for the remainder of the experiment.

Determination of arbitrary sample fitness

The term 'fitness', in this context, is borrowed from the field of computer science and the optimisation methods employed. A type of optimisation algorithm known as a genetic algorithm (GA; Holland 1975) simulates a Darwinian evolutionary system. As the simulation runs, potential solutions to the optimisation problem are evaluated for their 'fitness' (i.e. a quantitative measure of suitability), and the fittest solutions are allowed to survive to influence the structure of downstream potential solutions. For these simulations, a scoring system for sample fitness was designed to preserve the PDF.

A numerical measure of sample fitness was computed as follows: (i) cluster exposure maps, defined as a set of CEFA values binned in fixed 1% increments, from 0 to 100% sunlight exposure, were calculated to establish a histogram representing the discrete probability density of population cluster exposure for each domain's cluster inventory, $P(CEFA_{Pop})$, and formatted as a vector of 101 histogram bins; (ii) a similar vector representing the CEFA probability density of a particular sample, $P(CEFA_{Sample})$, was computed for every sample generated during a simulation; (iii) sample fitness was determined by subtracting the $P(CEFA_{Sample})$ vector from the $P(CEFA_{Pop})$ vector and computing Euclidean length of the resulting vector as shown in Equation 1.

$$Sample\ Fitness = \left\| P(CEFA_{Pop}) - P(CEFA_{Sample}) \right\|_2 \quad (1)$$

Determination of random sample fitness

A baseline model of random sample fitness was established through simulation for the purpose of later comparison with spatially explicit sample fitness scores. The baseline was determined by computing the fitness scores of simulated random samples ranging in size from 1 to 100% of the population equally spaced in 1% increments. To account for variation in results, the simulation was repeated in 30 trials and the final fitness score for each sample size was determined as the averaged score across all trials.

Determination of globally optimal samples

Maximally fit samples (i.e. those sampling patterns determined by simulation to have the best fitness score, and thus, be optimally representative of the population) were calculated to establish the best-case scenario of spatially explicit sampling performance. These globally optimised explicit samples (GOES) were computed by searching for the maximally fit combination of clusters (i.e. the sample with the best sample fitness score) from the known population while constrained only by a target sample size (e.g. 5% of population) and without regard for balancing the sample among individual vines or panels. No assumptions were made regarding the consistency of spatial variability within the block, and no spatial extrapolation was performed (i.e. sampling of uncharacterised areas in the vineyard was not performed).

Because of the combinatorial complexity of the problem space (e.g. exploring every sampling combination of 80 clusters in a population of 800 would require computing the sample fitness value of over 10^{100} sampling combinations), heuristic methods were used to identify the optimal sampling locations for each sample size. Two heuristic optimisation methods, Tabu Search (TS; Glover 1990) and GA (Holland 1975) were implemented to perform a minimisation of sample fitness score (with 0 as a perfect score) while searching the global sample space for samples of the specified size. A preliminary comparison of algorithm performance, using a subset of the experimental data, indicated that TS and GA optimisations converged on functionally equivalent solutions, but that TS found its minimum for our fitness function in about 10% of the computing time required by GA. Thus, TS was chosen for use in this experiment and the complete set of simulation trials, which analysed approximately 1.2 billion sample combinations, was computed in <48 hours processing time using a personal workstation.

Our custom simulation software was used to identify the optimal GOES solutions for samples ranging in size from 1 to 100% of the population equally spaced in 1% increments. The simulation was repeated in 30 trials and the final fitness score for each sample size was determined as the averaged score across all trials. Cluster sampling location information from the 30 trials was analysed to determine the probability of each cluster being included in the optimal sampling solution. Results were plotted, one plot for each domain-year combination, to illustrate the spatial patterns and their relationship to sample size.

Spatially explicit sampling models

In the interest of developing simple field methods that could more easily be used by growers and researchers, and illuminating the relationships between domain variability and optimal sampling solutions, two strategies for spatially weighted

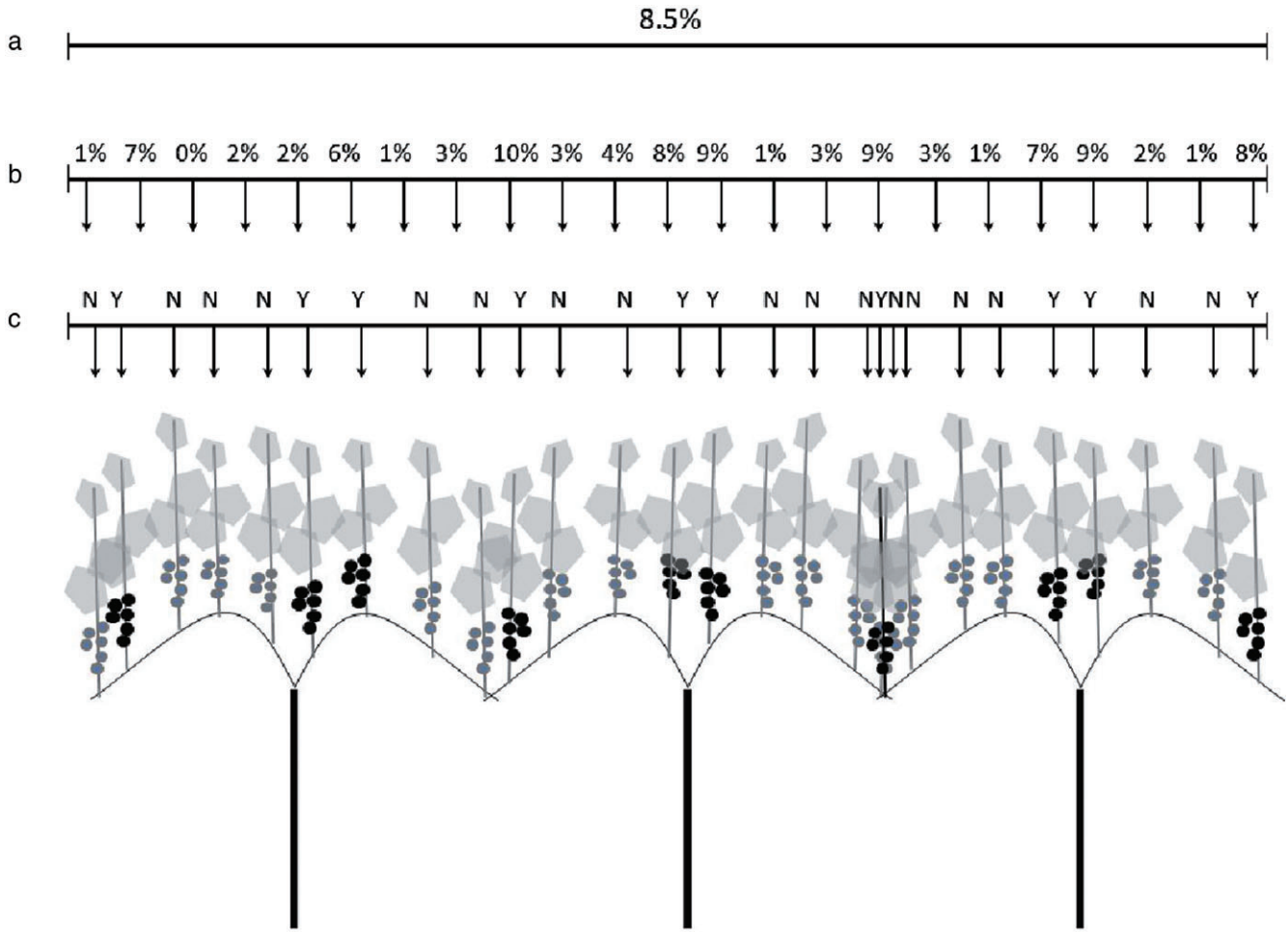

Figure 1. Example of a panel-weighted (a), architecture-weighted (b) and globally optimised (c) sampling template generated from EPQA dataset. The panel-weighted template indicates what percentage of the overall vineyard sample should be collected from this panel (i.e. 8.5% of the total sample – other panels will have different weights and all weights add up to 100%). The architecture-weighted sampling template indicates what percentage of the overall vineyard sample should be collected from this location along the length of each panel (applied consistently to each panel). The locations are based on 20 cm increments and the weights add up to 100% for each panel. The globally optimised template indicates precisely which clusters should be sampled from each panel (Y = sample, N = do not sample). Each panel has a unique template.

template sampling (STS) were developed, which employed a form of pseudo-random sampling that concentrated sampling frequencies within specific panels or to specific panel locations in quantities proportionate with the globally optimised probability densities. For each domain-year combination, GOES solutions were analysed to determine the inter-panel and intra-panel probability densities of optimal cluster sampling locations. These densities were used to define two types of weighted sampling templates for re-sampling the panels: a panel-weighted (PW) template, which determined the number of clusters to be harvested from each panel within the domain; and a canopy architecture-weighted (AW) template that assigned a weighted number of clusters to each 20 cm increment along the length of a panel (Figure 1). More specifically, the sampling strategy of PW templates directed sampling towards preferentially weighted panels, which were then sampled at random locations, while the sampling strategy of AW templates chose panels randomly, which were then sampled at preferentially weighted locations. For GOES solutions, all panels and clusters were included in the sampling selection, although as a consequence of the probabilistic weightings PW eliminated any panels that failed to appear in the GOES solution and AW eliminated any panel locations that failed to appear in the GOES solution.

PW templates consider variability as a function of vineyard location while AW templates consider variability as a function of vine organ distribution as influenced by vine spacing and training system. In PW sampling, each panel represented a stratum and each stratum was randomly sampled at a relative frequency corresponding to its appearance in the globally optimised solution. Similarly, in AW sampling each panel position (i.e. 20 cm increment) defined a stratum and each stratum was randomly sampled, among all of the panels in the dataset, at a relative frequency corresponding to its appearance in the globally optimised solution. Custom software simulated the template application for samples ranging in size from 1 to 100% of the population equally spaced in 1% increments. The simulation was repeated in 30 trials and the final fitness score for each sample size was determined as the averaged score across all trials.

Efficiency of GOES and STS versus random sampling

Fitness values versus sample size (1 to 100% of population) were determined for random, GOES and STS sampling protocols as described in previous paragraphs. For each domain-year combination, the resulting sample size versus fitness scores were

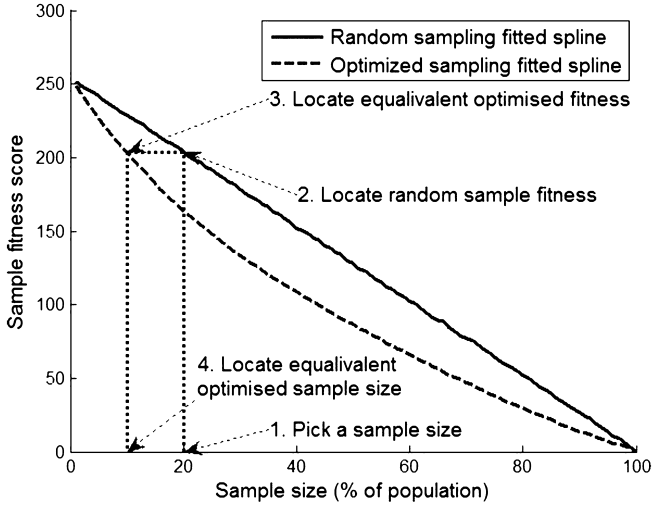

Figure 2. Illustration of the method used to interconvert sample size equivalents. A piecewise cubic spline was fitted to plots of fitness score versus random sampling and fitness score versus optimised sampling. Equivalent sample sizes are determined by locating the point on each curve where the fitness values are equal. Fitness score represents the 2-norm of the difference between population and sample discrete probability densities of the target vineyard variables.

fitted as a series of piecewise cubic polynomials, as determined by the MATLAB *spline* function. The fitted curves were then used to compute random-sample-equivalent sample sizes for each GOES and STS sample. Figure 2 illustrates the method of using the fitted curves to convert from a random sample size to an optimised sample size. Differences in sample size were plotted for each domain-year combination.

Results

Quantification of canopy autocorrelation

Spatial autocorrelation was present in all canopies, for all measured EPQA metrics and varied with lag distance (Table 1). Aggregated by EPQA metric, absolute mean autocorrelation for OLN, CEL and CEFA were 0.22 (±0.09), 0.13 (±0.02) and 0.12 (±0.02), respectively. Aggregated by training system, absolute mean autocorrelation for U, SH and 2TFB were 0.17 (±0.10), 0.16 (±0.09) and 0.14 (±0.02), respectively. Autocorrelation, expressed as aggregated absolute mean, varied among reported EPQA metrics with values of 0.22 (±0.09), 0.13 (±0.02) and 0.12 (±0.02) for OLN, CEL and CEFA, respectively.

Periodic spatial structure

Most autocorrelation plots revealed a visible hole-effect pattern, modulating between positive and negative values (illustrated in Figure 3 for domain C; domains A and B not shown). Fourier analysis of the underlying EPQA metrics at domain A determined that the primary OLN and CEL value cycles occurred at a period of 170 and 190 canopy row centimetres (i.e. 8.5 and 9 lags) for 2007 and 2008, respectively – a distance within 10 cm of the vine spacing at that domain. Primary CEFA value cycles occurred at 170 cm in 2007 and 84 cm (approximately 1/2 of the vine spacing) in 2008. The primary and secondary signal periods for domain C are annotated on Figure 3, and the complete Fourier dataset is presented in Table 1.

Table 1. Summary statistics for canopy autocorrelation and Fourier analysis of signal periods within domains A, B and C in the 2008 and 2009 season.

Domain	**OLN**						**CEL**						**CEFA**					
	A		**B**		**C**		**A**		**B**		**C**		**A**		**B**		**C**	
Year	**2008**	**2009**	**2008**	**2009**	**2007**	**2008**	**2008**	**2009**	**2008**	**2009**	**2007**	**2008**	**2008**	**2009**	**2008**	**2009**	**2007**	**2008**
Autocorrelation analysis																		
Min autocorrelation	−0.27	−0.25	−0.21	−0.6	−0.69	−0.47	−0.41	−0.24	−0.23	−0.36	−0.25	−0.32	−0.27	−0.27	−0.2	−0.15	−0.17	−0.27
Max autocorrelation	0.46	0.58	0.54	0.66	0.55	0.58	0.19	0.4	0.44	0.25	0.17	0.25	0.36	0.39	0.31	0.27	0.42	0.19
Mean autocorrelation	−0.05	−0.02	0.07	0	−0.01	0.01	−0.03	−0.03	0.07	−0.02	−0.02	−0.01	−0.04	0.01	0.03	0.03	−0.02	−0.01
Absolute mean autocorrelation	0.16	0.15	0.12	0.33	0.32	0.26	0.11	0.13	0.13	0.16	0.1	0.15	0.12	0.16	0.03	0.1	0.1	0.11
Fourier analysis																		
Vine spacing (cm)	200		200		180		200		200		180		200		200		180	
Primary signal period (cm)	290	310	*380*	**190**	**170**	**190**	290	**207**	**190**	**190**	**170**	**190**	48	67	84	**190**	**170**	*84*
Secondary signal period (cm)	**193**	**207**	253	380	52	253	53	78	109	95	76	84	290	150	*380*	48	227	58

Autocorrelation values were computed for the first 20 lag distances (i.e. at 20 cm intervals, which is approximately half of the panel length) within each panel. The reported mean values are the averages at each lag distance for all panels in a domain-year combination. Bolded signal values are within 10 cm of the vine spacing (i.e. 190 cm). Italicised values represents periods consistent with double- or half-vine spacing. CEL, cluster exposure layer; OLN, occlusion layer number.

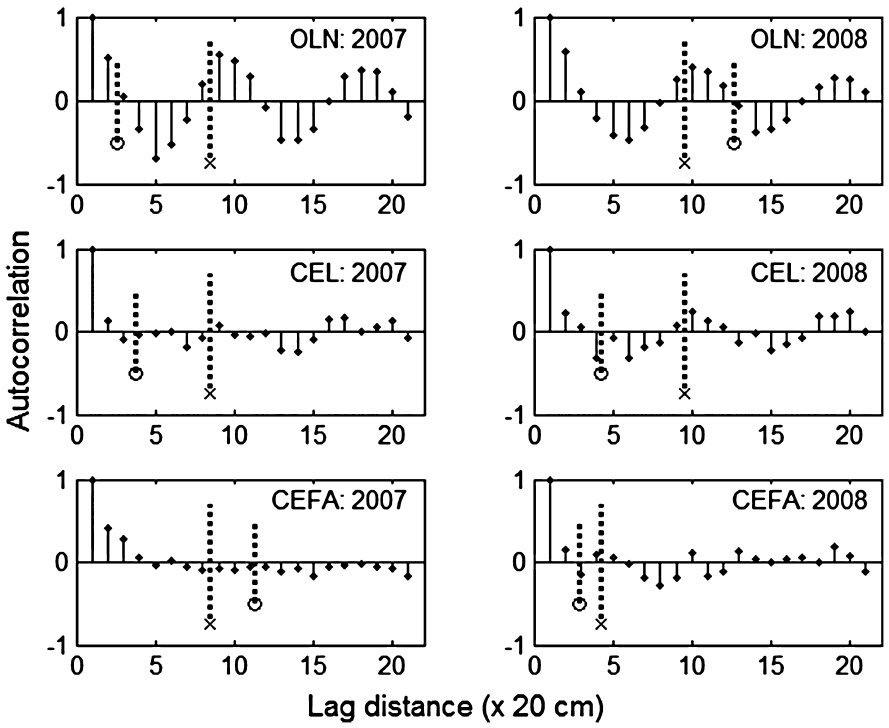

Figure 3. Average spatial autocorrelation (range –1 to 1) of EPQA metrics in Umbrella trained Vignoles vines (domain C). By definition, a lag distance of zero has an autocorrelation value of '1'. The 'hole-effect' pattern, attributed to the repetitive spatial structure of the canopy row along its longitudinal axis, is most evident in occlusion layer number, (OLN) which is a measure of canopy biomass. The dashed line located by the 'x' indicates the most prominent signal period as determined by Fourier analysis of the underlying EPQA metrics. The dashed line located by the 'o' indicates the signal period of next highest power. Vine spacing in the field is 180 cm (nine lags). CEFA, Cluster Exposure Flux Availability; CEL, Cluster Exposure Layer.

Ten of the 18 domain-year-metric combinations analysed revealed primary signal periods within 10 cm of the vine spacing (domain A: OLN in 2008 and 2009, CEL in 2008 and 2009, CEFA in 2009; domain B: CEL in 2009; domain C: OLN in 2007 and 2008, CEL in 2007 and 2008, CEFA in 2007). Four of the remaining eight combinations revealed either a primary signal that was reflective of double-vine spacing (domain A: OLN in 2008), half-vine spacing (domain C: CEFA in 2008) or a secondary signal period within 10 cm of the vine spacing (domain B: OLN in 2008 and 2009).

Spatial structure and sampling efficiency in GOES solutions

The globally optimised and randomised cluster sampling strategies over a range of sample sizes are shown in Figure 4 (domain A, 2009 data is presented; other domain-year combinations not shown). The optimal sampling strategy, as determined by GOES simulations, resulted in preferential selection of specific clusters at low sample sizes. The preferred clusters are associated with the darker bands in Figure 4b, and are more representative of the population distribution of CEFA. Conversely, non-preferred clusters, which are poorly representative of the CEFA of the population (i.e. outliers), appear as white bands in Figure 4b. Optimal cluster sampling locations, as determined by GOES simulations, indicated that certain cluster positions were preferential across all sample sizes, domains, years and training systems.

When compared to random samples of equal fitness scores, GOES achieved a reduction in sample size versus random sampling at all sample sizes in all domain-year combinations (Figure 5), although the extent of the improvement varied among treatments. For operationally practical sample sizes below 20% of the population, GOES resulted in reductions in required sample size ranging from 25 to 60% compared to random sampling (Figure 5b).

Spatial structure and sampling efficiency in STS solutions

Similar to the GOES optimisation, the use of a panel-weighted template strategy (PW STS) resulted in some panels being

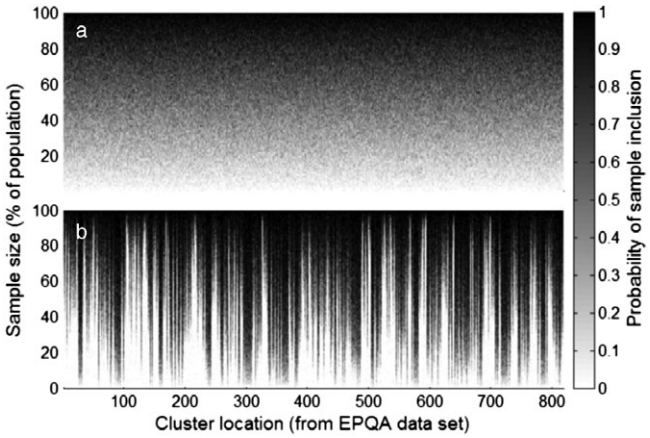

Figure 4. Comparison of sampling patterns generated by random sampling (a) and globally optimised spatially explicit sampling (b) at domain A in 2009. Banded pattern in the optimised sample indicates that some clusters (i.e. the dark bands) are preferred over others (i.e. the light bands) when choosing a minimum sample that best represents the targeted population variables (i.e. the probability distribution of cluster exposure flux availability in the population). Each cluster location on the *x*-axis corresponds to a specific cluster in an enhanced point quadrat analysis dataset. Each row on the *y*-axis represents a sample of size Y%.

preferably sampled over others in the optimal strategy, indicated by the visual banding pattern in Figure 6b. Unlike the GOES optimisation, a banding pattern is also visible when clusters are selected at random because of variation in cluster count among panels and thus cluster sampling frequencies (illustrated for domain A 2009 in Figure 6a; other domains not shown). Similar banding patterns appear for AW STS, reflecting higher cluster count numbers at some positions along the panel than other positions (Figure 7a). Visual banding in Figure 7b represents the variation in sampling frequencies after being adjusted to the optimal AW frequency.

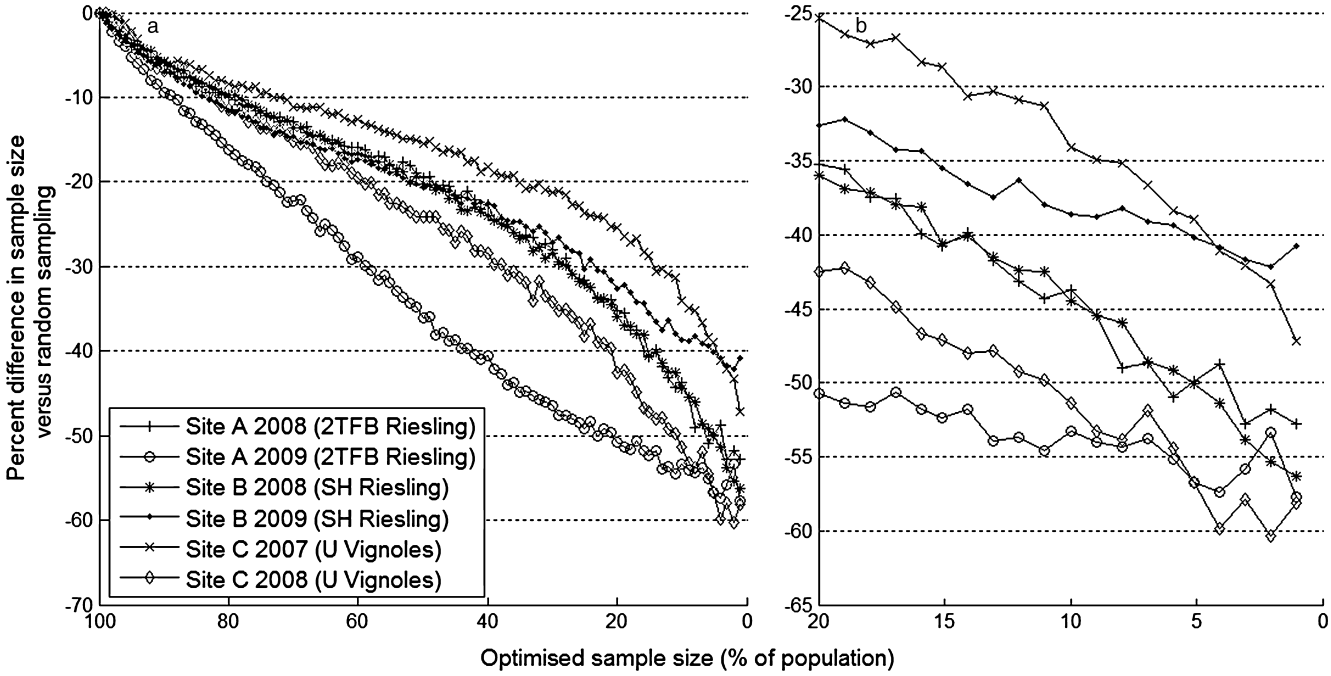

Figure 5. Reduction in sample size achieved through spatially explicit global optimisation of sampling. The sample sizes indicated on the *x*-axis are Y% (i.e. the value on the *y*-axis) smaller than a statistically comparable random sample. 2TFB, two-tier flatbow; SH, Scott Henry; U, Umbrella. Sample optimisation was performed for cluster exposure flux availability (CEFA). Panel A shows all tested sample sizes. Panel B shows detail of sample sizes between 1% and 20% of population. Site = sampling domain.

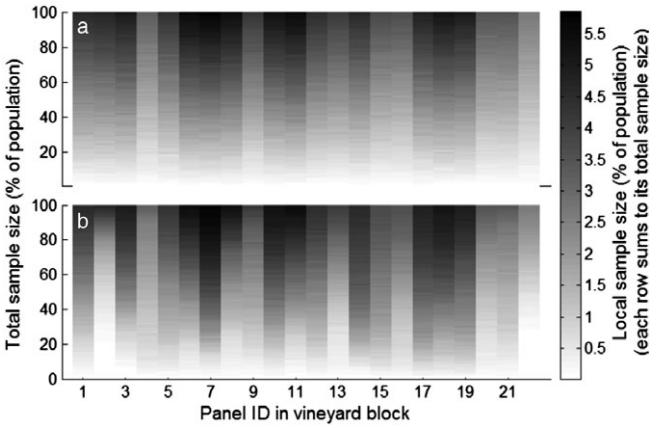

Figure 6. Sample density in random (a) and panel-weighted sampling templates (b; PW STS). Optimised panel-weighted sampling frequencies demonstrating the impact of globally optimised sampling on the probability that a cluster is sampled from each panel (data from domain A, 2009). Visual banding in the random sample plot is caused by variability in cluster count among panels that result in some panels naturally being sampled more than others. Banding differences in the optimised samples are a result of weighting adjustments introduced to improve sample fitness with respect to population variables.

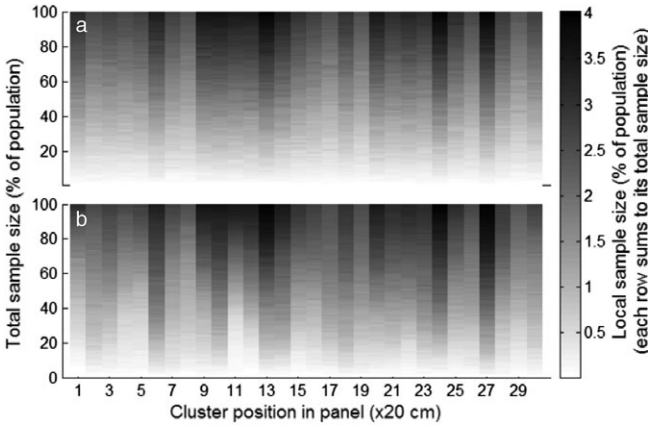

Figure 7. Sample density in random (a) and architecture-weighted sampling templates (b; AW STS). Optimised vine-architecture-weighted sampling frequencies demonstrating the impact of globally optimised sampling on the probability that a cluster is sampled from each position within a canopy panel (data from domain A, 2009). Visual banding in the random sample plot is caused by variability in cluster count among panel locations that result in some locations being sampled more than others. Banding differences in the optimised samples are a result of weighting adjustments introduced to improve sample fitness with respect to population variables.

When compared to random samples of equal fitness scores, PW STS achieved a reduction in sample size at random-sample sizes below 95% of population in all domain-year combinations (Figure 8) and reductions varied among treatments. For random sample sizes below 20% of population, PW STS achieved sample reductions ranging from 1 to 24% (plus one outlier of 47%) compared to random sampling. Similarly, AW STS achieved reduction in sample size at random sample sizes below 65% of population in all domain-year combinations (Figure 9) and reductions varied among treatments. For random sample sizes below 20% of population, AW STS achieved sample reductions ranging from 2 to 21%.

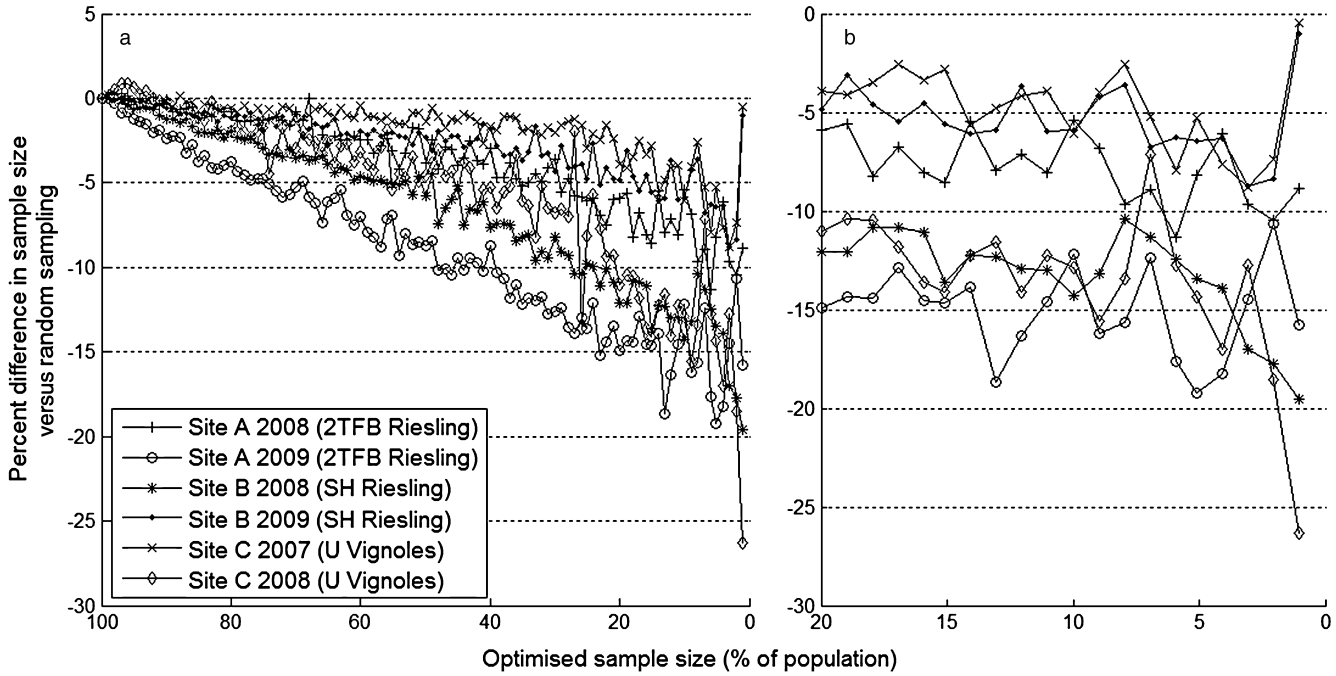

Figure 8. Difference in sample size achieved through panel-weighted template sampling. The sample sizes indicated on the *x*-axis are Y% (i.e. the value on the *y*-axis) smaller than a statistically comparable random sample. 2TFB, two-tier flatbow; SH, Scott Henry; U, Umbrella. Sample optimisation was performed for cluster exposure flux availability (CEFA). Values below zero represent a reduction in sampling requirements (vs random sampling), while values above zero indicate an increase in sampling requirements. Panel A shows all tested sample sizes. Panel B shows detail of sample sizes between 1 and 20% of population. Site = sampling domain.

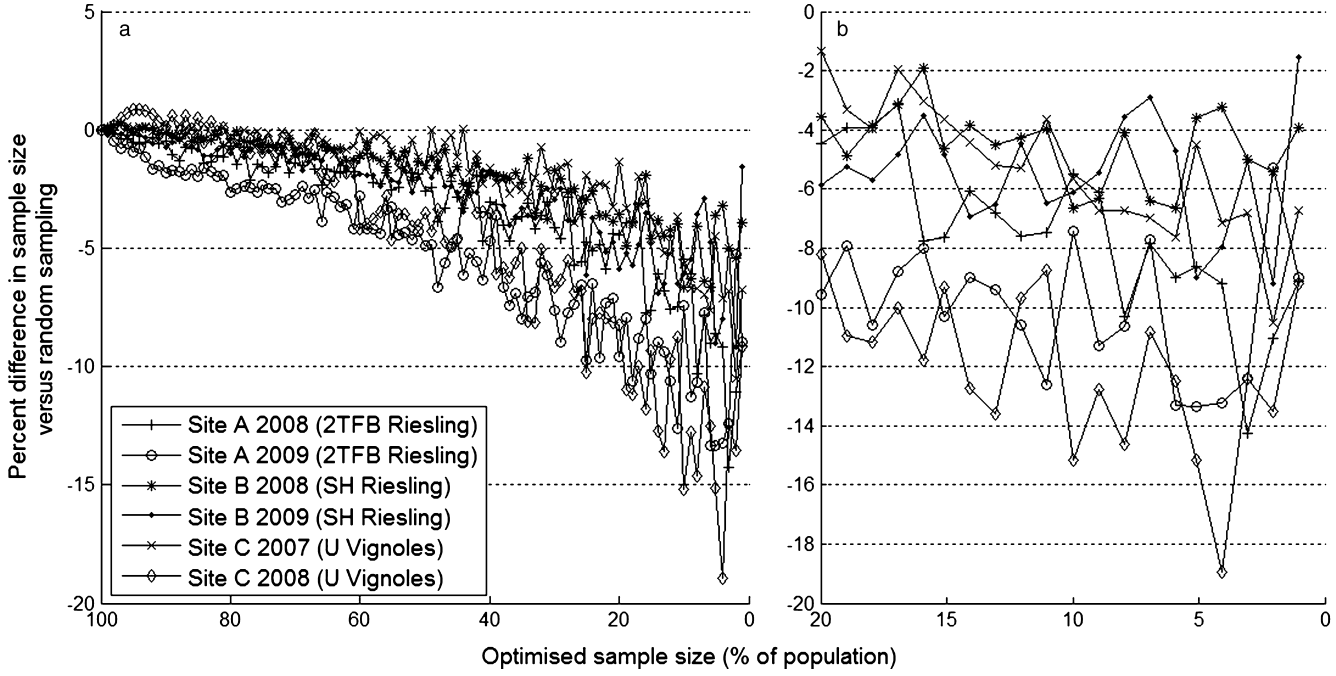

Figure 9. Difference in sample size achieved through architecture-weighted template sampling. The sample sizes indicated on the *x*-axis are Y% (i.e. the value on the *y*-axis) smaller than a statistically comparable random sample. 2TFB, two-tier flatbow; SH, Scott Henry; U, Umbrella. Sample optimisation was performed for cluster exposure flux availability (CEFA). Values below zero represent a reduction in sampling requirements (vs random sampling), while values above zero indicate an increase in sampling requirements. Panel A shows all tested sample sizes. Panel B shows detail of sample sizes between 1 and 20% of population. Site = sampling domain.

Discussion

Consistency of structural patterns

Autocorrelation among the reported EPQA metrics (Figure 3) suggests that measures of biomass density (OLN) exhibit more recognisable periodic spatial patterns than measures of cluster exposure (CEFA, CEL), perhaps because of higher number of leaf contacts versus cluster contacts per unit of row length, and the efficacy of cultural practices (i.e. leaf pulling) intended to improve exposure consistency. As expected, the vine spacing (180 or 200 cm) was similar to either the primary or secondary periods from Fourier analysis (Table 1) in nearly all treatments, suggesting that vine spacing is a major factor in determining within-panel spatial patterns in cluster exposure.

Limits of univariate sampling templates

Based on the presence of spatial patterns, and the potential for decreasing sampling sizes by up to 60% (Figure 5), we expected that an AW sampling template based on selecting clusters at regular locations along the panel (AW STS) should be successful in reducing cluster sample sizes because certain positions would be more representative of the population than others. While AW STS resulted in a reduction in sample size as compared to random selection (up to 21%), the smaller reductions enabled by AW STS compared to both PW STS (up to 24% reduction) and GOES (up to 60% reduction) suggests that the localised three-dimensional structure of a vine row, as influenced by vine spacing, training system and vine morphology can only partially explain the patterns.

Similarly, PW STS resulted in modest improvements as compared to random sampling, indicating that some panels are more representative of the population then other panels. However, the additional sample size reductions achieved with GOES as compared to PW STS suggests that those inter-panel patterns do not, in isolation, explain the majority of the spatial structure within the block. Moreover, the additive sample size reductions achieved with both sampling templates (AW and PW) was still less that the global optimum. This suggests that there are additional dimensions, some likely temporal in nature, that must be simultaneously balanced to achieve a reduction in sample size or maximise precision when measuring and describing cluster exposure.

Seasonal stability of spatial patterns

The reduction in sample size requirements for all optimised sampling methods versus random sampling varied between seasons (Figure 5), as did autocorrelation in vineyard variables and signal period (Table 1), suggesting that optimal sampling patterns would differ season to season. While it seems likely that temporal variability in seasonal weather patterns had an effect on season-to-season variability, cultural variability in cane pruning at all three domains may have been a more important factor, particularly for the sampling template based on position along a vine (AW STS). Cane positioning, with respect to trunk location at the vineyard floor, was inconsistent both within a vineyard row and between seasons (data not reported). Individual vines in vineyards in the Finger Lakes region often have several trunks that are trained without deliberate vertical positioning. Recording the location along the panel where a cane (or cordon) meets the fruiting wire would allow for new AW sampling templates based on position with respect to the trunk, and potentially improve season-to-season portability of vineyard maps and sampling solutions. In support of this work, EPQA-CEM Tools (Meyers and Vanden Heuvel 2008; available from first author upon request) has been updated to allow for the recording of this location, as designated by the letter 'T', which can be added to any EPQA insertion string to denote the location where the cane/cordon meets the fruiting wire.

Economically scalable operational models

The results of the optimised sampling methods demonstrated here suggest a substantial potential for improvement in sampling efficiency, but the practicality of implementation must also be considered. The optimum solution defined by GOES requires locating individual clusters during sampling. While this type of selective harvesting may be justifiable for a researcher who wishes to limit sample size because of limited availability or constraints associated with sample preparation, cluster-scale GOES methods are not likely to translate to a commercial vineyard, where sampling is often done by seasonal labour. Ultimately, the increased application of robotic automation in vineyards (Morris 2007, Cunha et al. 2010) may make GOES approaches commercially viable, but the current state of commercial grape production requires a more straightforward and less labour-intensive approach towards improving sampling protocols.

PW STS represents an example of an operationally customised approach to spatially explicit vineyard operation. Choosing clusters for sampling at the panel-scale preserves some of the spatial information computed for the best-case GOES solution without the challenge and cost of locating specific clusters. Once CEFA or another variable of interest (e.g. average estimated berry temperature) is characterized, a commercial vineyard could then use selective sampling of particular blocks to direct vineyard operations and improve sampling efficiency. Moreover, by using a cluster-scale GOES model as an ideal target, the spatial unit of work within vineyard operations can be scaled up (i.e. moving performance towards logistical simplicity) or down (i.e. moving performance towards the 'best-case' sampling scenario) as current labour cost, equipment capability, resource planning maturity and other economic factors dictate. In a heavily monitored vineyard where canopy density information is already available, the incremental cost of computing the protocols would be very low, making even a modest improvement in sample quality economically justifiable. Architecture-based sampling strategies (AW STS) also offered improvements over random sampling, and would be simple to translate into verbal instructions, e.g. 'take clusters at 80, 140 and 200 cm from the left edge of each panel'. However, the improvements were more modest than those achieved with the block-level-based sampling strategies.

Multivariate applications

Although the demonstrated cluster sampling optimisation was based on a single ecophysiological variable (i.e. fruit cluster exposure), heuristic optimisation methods naturally facilitate optimisations across multiple operational objectives. Rather than simply finding the best sampling strategy for representing the population cluster exposure, a multi-objective optimisation strategy could balance the influence of additional vineyard variables. Although collecting the necessary spatial vineyard values for multiple variables requires an investment in labour and/or technology, continued advances in vineyard sensing technology are likely to improve the economics of data collection through increases in sensor density, sampling frequency and variety of measurable variables (Cunha et al. 2010). Although current canopy imaging methods cannot produce fruit exposure maps with a spatial resolution and sensitivity equal to manual data collection, vineyard imagery will eventually surpass the

precision of manual collection. As this shift occurs, spatial interpolation of vineyard variables will likely become decreasingly necessary, while globally optimised data analysis methods will become increasingly beneficial in maximising operational efficiency, particularly when optimising around highly dimensional decision models. While the increased application of traditional data processing tools is an obvious response to increasingly larger datasets, the adoption of more sophisticated information processing and operational decision support methods such as those presented here could fundamentally transform vineyard management.

Conclusions

All tested spatially explicit models reduced required sample size to achieve similar performance as random sampling. Reduction in the required sample size was observed for both panel-weighted (PW STS) and architecture weighted (AW STS) sampling templates, although neither reduction was as great as the best-case globally-optimised sample (GOES). Measurements of cluster exposure have been shown to exhibit varying scales of spatial patterns within a domain. Although some patterns can be loosely predicted based on deliberate repetitive cultural practices (e.g. vine spacing or shoot positioning), some are based on less visibly obvious field conditions (e.g. soil structure or slope). Autocorrelation in canopy biomass and fruit exposure was quantitatively linked to fixed vine spacing, but the smaller efficiency gains associated with AW and PW sampling suggest that optimal sampling strategies should simultaneously account for spatial variability at multiple scales. However, even when the underlying cause for the patterns is not definitively known, data from measured spatial patterns can be used to improve operational efficiency by guiding vineyard activities (e.g. sampling or selective harvesting) towards the locations that will most effectively achieve a desired goal. Finally, the methods presented here should be readily applicable to optimising sampling protocols for other targeted variables, including multivariate models.

Acknowledgements

The authors thank Steve Lerch for technical assistance in the vineyard; Aditya Barua for helpful programming discussions; and Atwater Vineyards, Wagner Vineyards, and Lamoreaux Landing for use of their vineyards.

References

Acevedo-Opazo, C., Tisseyre, B., Ojeda, H. and Guillaume, S. (2010) Spatial extrapolation of the vine (*Vitis vinifera* L.) water status: a first step towards a spatial prediction model. Irrigation Science **28**, 143–155.

Bramley, R.G.V. and Hamilton, R.P. (2004) Understanding variability in wine grape production systems 1. Within-vineyard variation in yield over several vintages. Australian Journal of Grape and Wine Research **10**, 32–45.

Cunha, C.R., Peres, E., Morais, R., Oliveira, A.A., Matos, S.G., Fernandes, M.A., Ferreirad, P.J.S.G. and Reis, M.J.C.S. (2010) The use of mobile devices with multi-tag technologies for an overall contextualized vineyard management. Computers and Electronics in Agriculture **73**, 154–164.

De Gruijter, J., Bierkens, M., Brus, D. and Knotters, M. (2006) Sampling for natural resource monitoring (Springer-Verlag GmBH: Heidelberg, Germany).

van Es, H.M., Gomes, C.P., Sellmann, M. and van Es, C.L. (2007) Spatially-balanced complete block designs for field experiments. Geoderma **140**, 346–352.

Gerdes, S.M., Winterhalter, P. and Ebeler, S. (2002) Effect of sunlight exposure on norisoprenoid formation in white Riesling grapes. Carotenoid-Derived Aroma Compounds. Chapter 19, Washington, DC, USA, American Chemical Society, 262–272.

Glover, F. (1990) Tabu search – a tutorial. Interfaces **20**, 74–94.

Holland, J.H. (1975) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control and artificial intelligence (The University of Michigan Press: Ann Arbor, MI).

Iland, P., Bruer, N., Ewart, A., Markides, A. and Sitters, J. (2004) Monitoring the winemaking process from grapes to wine: techniques and concepts (Patrick Iland Wine Promotions, Pty Ltd: Adelaide, SA, Australia).

Jeffreys, H. (1939) Random and systematic arrangements. Biometrika **31**, 1–8.

Kirk, H.J., Haynes, F.L. and Monroe, R.J. (1980) Application of trend analysis to horticultural field trials. Journal of the American Society for Horticultural Science **105**, 189–193.

Lark, R.M. and Marchant, B.P. (2009) Using advanced methods to reduce the cost of soil sampling; ideas for the future. R&D Conference 'Precision in arable farming: current practice and future potential', October 28–29, Grantham, Lincolnshire, UK, pp. 18–25.

Meyers, J.M. and Vanden Heuvel, J.E. (2008) Enhancing the precision and spatial acuity of point quadrat analyses via calibrated exposure mapping. American Journal of Enology and Viticulture **59**, 424–431.

Meyers, J.M. and Vanden Heuvel, J.E. (2009) Spatial correlation in vine biomass density suggests need for new design and sampling protocols. American Journal of Enology and Viticulture **60**, 553A.

Meyers, J.M., Sacks, G.L. and Vanden Heuvel, J.E. (2009) Naturally occurring spatial variability in canopy biomass impacts flavour and aroma compounds in Riesling. American Journal of Enology and Viticulture **59**, 394A–395A.

Morris, J.R. (2007) Development and commercialization of a complete vineyard mechanization system. HortTechnology **14**, 411–420.

Panten, K., Bramley, R.G.V., Lark, R.M. and Bishop, T.F.A. (2010) Enhancing the value of field experimentation through whole-of-block designs. Precision Agriculture **11**, 198–213.

Rankine, B.C., Cellier, K.M. and Boehm, E.W. (1962) Studies on grape variability and field sampling. American Journal of Enology and Viticulture **13**, 58–72.

Reynolds, A.G. and Vanden Heuvel, J.E. (2009) Influence of grapevine training systems on vine growth and fruit composition: a review. American Journal of Enology and Viticulture **60**, 251–268.

Roessler, E.B. and Amerine, M.A. (1958) Studies on grape sampling. American Journal of Enology and Viticulture **9**, 139–145.

Student (1938) Comparison between balanced and random arrangements of field plots. Biometrika **29**, 363–379.

Tamura, R.N., Nelson, L.A. and Naderman, G.C. (1988) An investigation on the validity and usefulness of trend analysis for field plot design. Agronomy Journal **80**, 712–718.

Taylor, J., Tisseyre, B., Bramley, R. and Reid, A. (2005) A comparison of the spatial variability of vineyard yield in European and Australian production systems. In: Precision Agriculture '05: Proceedings of the 5th European Conference on Precision Agriculture. Eds. J.V. Stafford and A. Werner (Wageningen Academic Publishers: Wageningen, the Netherlands) pp. 907–915.

Xu, X., Murray, R.A., Salazar, J.D. and Hyder, K. (2008) The temporal pattern of captan residues on apple leaves and fruit under field conditions in relation to weather and canopy structure. Pest Management Science **64**, 565–578.

Manuscript received: 28 September 2010
Revised manuscript received: 30 March 2011
Accepted: 1 April 2011