Statistical Learning in Soil Sampling Design Aided by Pareto Optimization

Assaf Israeli
Dept. of Water Sciences, Tel-Hai College, and
Dept. of Precision Agriculture, MIGAL Research Institute
Upper Galilee, Israel
assafi@migal.org.il

Michael Emmerich
Leiden Institute of Advanced Computer Science
Leiden University
Leiden, The Netherlands
m.t.m.emmerich@liacs.leidenuniv.nl

Michael (Iggy) Litaor
Dept. of Water Sciences, Tel-Hai College, and
Dept. of Precision Agriculture, MIGAL Research Institute
Upper Galilee, Israel
litaori@telhai.ac.il

Ofer M. Shir
Dept. of Computer Science, Tel-Hai College, and
Dept. of Precision Agriculture, MIGAL Research Institute
Upper Galilee, Israel
ofersh@telhai.ac.il

ABSTRACT

Effective soil-sampling is essential for the construction of prescription maps used in Precision Agriculture for Variable Rate Application of nutrients. In practice, designing a field sampling plan is subject to hard limitations, merely due to the associated expenses, where only a few sample points are taken for evaluation. The accuracy of constructed maps is affected by the number of sampling points, their geographical dispersion and their coverage of the feature space. To improve the accuracy, ancillary data in the form of low-cost, high-resolution field scans could be used for inferring statistical measures for devising the high-cost sampling plan. The current study targets algorithmically-guided sampling plans using available ancillary data. We propose possible models for quantifying spatial coverage and diversity concerning the ancillary data. We investigate models as objective functions, devise Pareto optimization problems and solve them using NSGA-II. We analyzed the obtained sampling plans in an agricultural field, and suggest statistical tools for sample-size determination and plans' ranking according to additional information criteria. We argue that our approach is successful in attaining a practical sampling plan, constituting a fine trade-off between objectives, and possessing no discrepancies.

CCS CONCEPTS

• **Computing methodologies** → **Randomized search**; • **Applied computing** → **Agriculture**;

KEYWORDS

geostatistical learning, sampling information, Pareto optimization, precision agriculture, Solow-Polasky diversity, sampling design, spatial design, soil survey, Conditioned Latin Hypercube Sampling

GECCO '19, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6111-8/19/07…$15.00
https://doi.org/10.1145/3321707.3321809

1 INTRODUCTION

Precision Agriculture (PA) relies on soil sampling to produce prescription maps for variable rate application of fertilizers. Spatial coverage sampling strategies, such as grid sampling or stratified random sampling, provide a sound geographical distribution but require extensive sampling to generate effective variograms [27]. An alternative approach for sample design optimization is the so-called *Conditioned Latin Hypercube Sampling* (cLHS)[18], which accounts also for a predefined feature space. It aims to maximize the stratification of the feature space in the sample, and yet it may produce an uneven coverage of the geographical space, overall resulting in a low estimation quality. Hengl et al. [11] concluded that prediction accuracy may be improved by covering both the feature and the geographical spaces in a sample. Gao et al. [10] added a spatial measure to cLHS by aggregation into a single objective function, and achieved smaller mapping errors in comparison to strictly spatial or feature space methods. Lark [16] demonstrated the feasibility of multiobjective optimization of spatial sampling design using Simulated Annealing on a theoretical use-case, where the objective functions are total distance travelled for sampling and the variance of the sample mean. Following these studies, we propose a heuristic approach for obtaining a sample design by solving a bi-objective optimization problem by concurrently maximizing the feature space stratification *and* the geographical distribution of the sampling points. In doing so, we aim to provide sampling schemes that would render high accuracy maps portraying the entire information spectrum of the soil, supported by statistical analysis to assist in plan selection and scaling down the sample-size – a significant contributor to the operational costs.

Given the complex nature of soil attributes, geostatistical methods consider target variables as realizations of random fields [25]. These methods infer statistical parameters and calculate predictions based on partial observations of the random field realization. Assuming a normally distributed stochastic process with second-order stationarity (a constant *mean* and an *autocovariance* function dependent solely upon the distance between any two values), sampled data can be interpolated by the widely used Ordinary Kriging (OK) method [17], which provides a best linear unbiased prediction (BLUP) of values at unsampled locations. OK requires a positive

definite model of spatial variability, calculated as a function fitted to experimental variogram of the sample data. The reliability of the model depends on the capacity of sample information to capture the variability of the observed phenomenon, affected primarily by the number of observations [27].

The use of ancillary data can reduce prediction errors associated with reduction in sample size and facilitate cost-effective sampling [28]. By *ancilliary data* we denote any source of spatial information with some relation to soil properties and in the form of a digital soil map. Importantly, high-resolution data – e.g., multi-spectral aerial imaging, proximal sensing or yield maps – are available at a rather low cost. Although the exact relation of these data to soil attributes may be unknown, a spatial variability model can be formed to guide soil sampling design [2] and density [13], as well as to improve the approximation accuracy as a covariate in co-Kriging [21].

The quality of a sampling scheme can be evaluated *a priori* by the uncertainty of prediction map resulting from interpolation of known ancillary data values at sample locations, reflected by metrics such as Mean Ordinary Kriging Variance (MOKV), subject to the assumption that the attribute under study is a realization of a stationary Gaussian random function [12], or by the Root Mean Square Error (RMSE) between predicted and true values at all locations. We studied additional metrics to support agricultural soil-survey planning, focusing on practical questions concerning sample-sizing and composition. With this work, we introduce a preferential index for ranking candidate sampling schemes according to expected MOKV and statistical measures for model selection, calculated by the Kullback-Leibler Divergence (D_{KL}) [14] and Akaike Information Criterion (AIC) [1].

The current study targets the following *research questions*:

> Which model captures the effectiveness as well as the cost-efficiency of sampling-plans when accounting for both diversity and representation? Moreover, could an algorithm obtain practical sampling-plans?

The proposed contributions of the current study are:

(1) Modelling of objective functions quantifying sampling-plans designed for the efficient use of ancillary data;
(2) Formulation of multiobjective optimization problems for optimizing sampling strategy;
(3) Solving these optimization problems for agricultural farm using real-field data.

The paper has the following structure: In Section 2 we outline the geostatistical learning challenge – we specify our notation, describe existing approaches and motivate the multiobjective optimization perspectives. We then specify our methods for solving the prescribed challenge in Section 3, namely a selected Evolutionary Multiobjective Optimization Algorithm (EMOA) with problem-specific search operators. Our practical observations on the current case-study are reported in Section 4, where we also discuss the attained solutions. Finally, we summarize our work and findings in Section 5, where we also draw possible directions for future work.

2 PRELIMINARY: FROM ANCILLARY DATA TO SAMPLING-PLANS

We begin by specifying our notation. Let N denote the available number of sampling sites in the field, whose ancillary data are assumed to be acquired in dimensionality k (where each of the k coordinates is also known as a layer or a channel). That is, the ancillary data is represented by vectors in $\mathbb{R}^k$ at each of the N sites. Let $\mathcal{A}$ denote the corresponding $N \times k$-dimensional ancillary data matrix, and importantly, let it define the *feature space*. At the same time, every sampling-point is associated with spatial (x, y)-coordinates, $\{(x_i, y_i)\}_{i=1}^{N}$, subscribing to the so-called *geographical space*, which is defined by the field's boundaries and operational constraints. These N pairs of coordinates constitute the geographical data matrix $\mathcal{G}$:

$$\mathcal{A} = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,k} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{N,1} & \cdots & \cdots & \alpha_{N,k} \end{pmatrix}, \quad \mathcal{G} = \begin{pmatrix} x_1 & y1 \\ x_2 & y2 \\ \vdots & \\ x_N & y_N \end{pmatrix} \tag{1}$$

Pairwise distance calculations use the Euclidean metric. Depending on the context, distance calculations are conducted either within the geographical space $\mathcal{G}$ or within the feature space $\mathcal{A}$. They are denoted by $d_{i,j}^{(\mathcal{G})}$ for every $i, j \in \{1, \ldots, N\}$, or $d_{i,j}^{(\mathcal{A})}$ for the feature space. In our notation, we denote by $\mathcal{Z}$ the *augmented* data matrix of dimension $N \times (k + 2)$, encompassing the N sites' ancillary data and their geographical coordinates:

$$\mathcal{Z} = \left(\mathcal{A} \mid \mathcal{G}\right). \tag{2}$$

The ultimate target is to form a sampling-plan by locating $n \ll N$ sites, whose ancillary data vectors best represent the feature space's distribution, and at the same time, are spatially disperse concerning the geographical space. At this point, we assume that the user provides a value of n and postpone the discussion on setting this value to Section 3.3.

Formally, a candidate sampling-plan p is a mapping π indicating the subset selection of the n indices. Importantly, **a candidate sampling-plan p is associated with the following components**:

i An ancillary data matrix of dimension $n \times k$, denoted by $\mathbf{A}^{(p)}$, whose rows constitute a subset of $\mathcal{A}$'s rows adhering to the mapping π
ii A geographical data matrix $\mathbf{G}^{(p)}$, defined in an equivalent manner
iii An augmented matrix $\mathbf{Z}^{(p)}$:

$$\mathbf{A}^{(p)} = \begin{pmatrix} \alpha_{\pi(1),1} & \alpha_{\pi(1),2} & \cdots & \alpha_{\pi(1),k} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{\pi(n),1} & \alpha_{\pi(n),2} & \cdots & \alpha_{\pi(n),k} \end{pmatrix}, \quad \mathbf{G}^{(p)} = \begin{pmatrix} x_{\pi(1)} & y_{\pi(1)} \\ \vdots & \\ x_{\pi(n)} & y_{\pi(n)} \end{pmatrix},$$
$$\mathbf{Z}^{(p)} = \left(\mathbf{A}^{(p)} \mid \mathbf{G}^{(p)}\right). \tag{3}$$

Thus, the sampling planning process may be translated into obtaining the subset selection mapping π as a Combinatorial Optimization problem, or into locating $2n$ real-valued geographical coordinates

as a continuous search-problem, in which case the n geographical coordinates are matched with the closest points within $\mathcal{G}$.

2.1 Conditioned Latin Hypercube Sampling

Minasny and McBratney [18] devised a method for obtaining optimal sampling design in the presence of ancillary data, namely cLHS. This method solves sampling design as a single-objective optimization problem using dedicated combinatorial local search operators. It aims to maximally stratify the multivariate distribution of ancillary data layers by forming a Latin hypercube of their *quantiles*, while preserving the structure of their statistical correlation, for a representation of the full information spectrum. Importantly, this method accounts only for the feature space.

Given the ancillary data matrix $\mathcal{A}$, the idea is to compute n statistical quantiles per each of its k channels, then aim to place a single member of each quantile in the sample. Given a candidate sampling-plan of n sites, p, defined by a mapping π, let η hold histogram information of $\mathbf{A}^{(p)}$ with respect to $\mathcal{A}$ in the following manner: given the i^{th}-quantile of $\mathcal{A}$'s j^{th} channel, $q_j^{(i)}$, the element $\eta^{(p)}\left[q_j^{(i)} \leq \alpha_{\pi(i),j} < q_j^{(i+1)}\right]$ is the number of occurrences of $\alpha_{\pi(i),j}$ with values in the corresponding quantile. Accordingly, the first evaluation criterion within cLHS is defined by

$$\psi_1\left(\mathbf{A}^{(p)}\right) = \sum_{i=1}^{n}\sum_{j=1}^{k}\left|\eta\left[q_j^{(i)} \leq \alpha_{\pi(i),j} < q_j^{(i+1)}\right] - 1\right|. \tag{4}$$

Also, let $\mathbf{C}^{(\mathcal{A})}$ and $\mathbf{C}^{(\mathbf{A}^{(p)})}$ denote the *correlation matrices* of $\mathcal{A}$ and $\mathbf{A}^{(p)}$, respectively, both $k \times k$-dimensional. The second evaluation criterion is the following:

$$\psi_2\left(\mathbf{A}^{(p)}\right) = \sum_{i=1}^{k}\sum_{j=1}^{k}\left|\mathbf{C}_{i,j}^{(\mathcal{A})} - \mathbf{C}_{i,j}^{(\mathbf{A}^{(p)})}\right|. \tag{5}$$

For categorical data such as soil classification the sub-objective is to match the probability distribution for each of the classes. With strictly continuous ancillary data cLHS defines an objective function for evaluating p's quality as a weighted sum of the two criteria ($\omega_1, \omega_2 > 0$, for general application $\omega_1 = \omega_2 = 1$ [18]):

$$f_{\text{cLHS}}(p) = \omega_1 \cdot \psi_1\left(\mathbf{A}^{(p)}\right) + \omega_2 \cdot \psi_2\left(\mathbf{A}^{(p)}\right) \longrightarrow \min. \tag{6}$$

In terms of problem-solving, cLHS operates with a dedicated variation operator, which swaps a random site within the subset mapping π of the candidate sampling-plan p with one of the elements in its complement π^C, to obtain $\tilde{\pi}$ (defining $\tilde{p}$):

$$\{p, \pi\} \rightsquigarrow \left\{p', \pi'\right\} \text{ such that } \delta\left(\pi, \pi'\right) = 1, \tag{7}$$

where δ counts the differing subsets' attributes.

Overall, cLHS culminates at a perfect stratification of the feature space, yet it does not account for the geographic distribution of the sampling points, thus results in many impractical solutions. It was suggested to run cLHS for a subset of the points, followed by a space-filling algorithm for the remaining points [28]. In practice, we have found that this procedure still produces inefficient solutions, and suggest as a remedy to augment the cLHS objective function with a spatial dispersion objective function, as described in the following subsection.

2.2 max-min Diversity

The max-min diversity is one of the simplest notions for promoting dispersion. Despite the simplicity of this diversity indicator, finding maximally diverse subsets is an $\mathcal{NP}$-hard problem [15]. Here, it aims to maximize the minimal pairwise (geographical) distances among all sampling points:

$$f_{d_{\min}^{(\mathcal{G})}}(p) = \min_{\pi(i),\pi(j)}\left\{d_{\pi(i),\pi(j)}^{(\mathcal{G})}\right\} \longrightarrow \max \qquad i, j \in 1, \ldots, n,\ i \neq j. \tag{8}$$

2.3 Multiobjective Formulation

Evidently, competitions between spatial diversity and feature space coverage arise in the context of our research questions. We, therefore, provide herein the necessary multiobjective formulation as preparation for posing our specific optimization problems.

Given a multiobjective optimization problem with m objectives, let an objectives vector in $\mathbb{R}^m$ be denoted as,

$$\vec{f}(\vec{x}) = \left(f_1(\vec{x}), f_2(\vec{x}), \ldots, f_m(\vec{x})\right)^T,$$

and let all its coordinates assumed to be subject to *minimization*. A partial order is defined on the m-dimensional objective space, $\mathcal{F} = \vec{f}(\mathcal{X})$, by means of the *domination* concept: given any $\vec{f}^{(1)} \in \mathbb{R}^m$ and $\vec{f}^{(2)} \in \mathbb{R}^m$, we state that $\vec{f}^{(1)}$ weakly Pareto dominates $\vec{f}^{(2)}$, noted as $\vec{f}^{(1)} \preceq \vec{f}^{(2)}$, if and only if the following holds: $\forall i \in \{1, \ldots m\} : f_i^{(1)} \leq f_i^{(2)}$. We also consider the strict Pareto domination: $\vec{f}^{(1)} \prec \vec{f}^{(2)} \iff \vec{f}^{(1)} \preceq \vec{f}^{(2)} \wedge \exists i \in \{1, \ldots, m\} : f_i^{(1)} < f_i^{(2)}$. We then state that $\vec{f}^{(1)}$ and $\vec{f}^{(2)}$ are *incomparable* or *indifferent*, noted as $\vec{f}^{(1)} || \vec{f}^{(2)}$, if and only if $\vec{f}^{(1)} \not\preceq \vec{f}^{(2)} \wedge \vec{f}^{(2)} \not\preceq \vec{f}^{(1)}$. For any non-empty compact subset of $\mathbb{R}^m$, say $\mathcal{F}$, there exists a non-empty set of minimal elements for the partial order $\preceq$ [7]. Non-dominated points are the set of minimal elements for $\preceq$: $\mathcal{F}_N = \left\{\vec{f} \in \mathcal{F} | \nexists \vec{f}' \in \mathcal{F} : \vec{f}' \prec \vec{f}\right\}$. The goal of Pareto optimization is to obtain the *non-dominated set* for $\mathcal{F} = \vec{f}(\mathcal{X})$ entitled the Efficient Frontier, and its pre-image in $\mathcal{X}$, the so-called *Pareto optimal set*.

3 PROBLEM FORMULATION AND ALGORITHMIC APPROACH

We formulate here the concrete optimization problem that we target and then describe our solution approach.

3.1 Bi-Objective Formulation

As stated before, we are interested in investigating the competition between spatial diversity and feature space coverage. Accordingly, given the model functions presented in Section 2.1, we formulate a bi-objective optimization problem. For the sake of compatibility, we compute the multiplicative inverse, so that all objectives are subject to minimization:

$$\boxed{\begin{array}{l}\textbf{[P0]}\\ f_1 := f_{\text{cLHS}}(p) \longrightarrow \min\\ f_2 := 1/f_{d_{\min}^{(\mathcal{G})}}(p) \longrightarrow \min\end{array}} \tag{9}$$

To achieve dimensionless-scaling, the function f_{cLHS} is normalized (i.e., divided by n).

3.2 Algorithmic Approach

Here, we are especially interested in EMOAs, which have undergone considerable development in the past two decades [8]. In practice, we employed the renowned NSGA-II [5], utilizing the `ecr` R-package [3], with an elitist non-dominated sorting selector $(\mu + \lambda)$. We set the parental and offspring population sizes both to $\mu = \lambda = 10$, and the maximally available iterations to 50,000. A simple mutation operator tailored to the current domain *swaps* a random point in the sample set $\mathbf{Z}^{(p)}$ with a random member of its complementary set $(\mathbf{Z}^{(p)})^C$, as outlined by Algorithm 1.

```
mutatePlan(π, N)
  n ⟵ length(π)
  P^C ⟵ {1, ..., N} \ {π(1), ..., π(n)}
  i_rmv ⟵ uniformly randomly from {1, ..., n}
  i_add ⟵ uniformly randomly from P^C
  π' = (π(1), ..., π(i_rmv) ⇝ i_add, ..., π(n))
return π'
```

Algorithm 1: The utilized mutation operator, formulated as a function named `mutatePlan`. Upon receiving a mapping of an existing sampling-plan π as input, $\mathcal{P}^C$ is the subset of unsampled sites, calculated as the difference between the complete set of sites and π's sites. Then, the operator swaps a single site $\pi(i_{\text{rmv}})$ with a uniformly random unsampled site i_{add}. The modified mapping π' is returned as output.

The vector of objective functions is evaluated using Eq. 9. The cLHS fitness function, inspired by the `clhs` R-package [22], first segments each variable into n iso-probable quantiles (strata) according to its CDF (Cumulative Distribution Function) and calculates the correlation matrix $\mathbf{C}^{(\mathcal{A})}$.

The optimization procedure starts by initializing a population of μ individuals, each constituting a candidate sampling-plan comprising n random points. It then iterates over a serial execution of the variation (mutation) and the selection operations and terminates at a predefined maximal number of objective function evaluations. Results of the last generation are Pareto-sorted to exclude the dominated solutions, producing a portfolio of Pareto-optimal sampling-plans to consider. Technically, through CPU parallelization, 30 independent optimization runs are concurrently executed to ensure sufficient replications.

3.3 Sample-Size Identification

The choice of the sample-size n essentially reflects the economic concept of *marginal profit*, as every additional sampling-point presumably improves prediction accuracy at the cost of increased operational expenses. We seek to quantify the information gain by increments of sample-size, in a workflow of multiple optimization tasks (3.2) with n varying within a pragmatic budget range. The resulting Pareto-optimal solutions are evaluated by the following statistical measures: AIC, calculated for linear models, fitted with each ancillary data channel as a dependent variable; MOKV, resulting from Kriging interpolation of ancillary data values at sample points; and D_{KL}, derived from the ratio of ancillary data distributions in a sample, to those of the complete field, given by

$$D_{KL}\left(\mathcal{A} \,||\, \mathbf{A}^{(p)}\right) = -\sum \mathbf{Pr}(\mathcal{A}) \log\left(\frac{\mathbf{Pr}(\mathbf{A}^{(p)})}{\mathbf{Pr}(\mathcal{A})}\right). \tag{10}$$

3.4 Scheme Selection Criteria

Once the sample-size n is set, an EMOA may be deployed to solve **P0** and obtain an approximate Pareto frontier. However, decision-making needs to take place post-optimization, for selecting a concrete plan among the available on the frontier. To this end, the same information criteria as in 3.3 are used here to form a ranking order. It is then used to obtain a prioritized list of candidates for the expert's selection process, which potentially will account for additional practical/subjective aspects as well.

4 PRACTICAL OBSERVATIONS

In this section we report on our empirical findings on real-field settings and data. The utilized units' abbreviations read: [ha] for Hectare, [m] for meters.

As a part of an ongoing research on PA fertilizer management, we have tested the aforementioned methods to devise a soil-survey plan for a 37 ha plot in Jezreel valley, northern Israel, before winter wheat is sown. Ancillary data was collected using EM38-MK2 ground conductivity meter (*Geonics Ltd.*, Canada) to record apparent electrical conductivity (ECa) and apparent magnetic susceptibility (MSa) at geo-referenced points in two modes of operation: vertical and horizontal, measuring two depth ranges (0-1.5 m and 0-0.75 m, respectively). Normalized Difference Vegetation Index (NDVI) data were collected with a multi-spectral camera mounted on an unmanned aerial vehicle (UAV). The pre-processing stage involves ECa and MSa data compaction, log transformation, and Ordinary Kriging interpolation to a grid at a resolution of 1×1 m (using `gstat` R-package [20]), followed by a crop to field boundaries and normalization. NDVI layer was scaled down to the same resolution, cropped and normalized. Resolution of 1 m is a compromise between precision and computation efficiency. The field was divided into 4 management zones (MZs) using Fuzzy-c-means clustering [19] of the matrix $\mathcal{A}$, followed by smoothing with a median filter [4] to reduce zone fragmentation [23]. The feasible search space was then defined by exclusion of a 14 m buffer from field boundaries, and 7 m buffer from MZ boundaries, outlined with an edge-detection filter. Figure 1 provides a summary of ancillary data layers and the search space, where red areas represent high ECa and MSa values, green areas represent mid-range values while blue areas exhibit the lowest ECa and MSa values, probably due to coarse texture and low soil moisture content. It seems that the NDVI data add little to the clustering power because of relatively low spatial variations in the NDVI signal of bare soil.

The prescribed EMOA was executed in 30 parallel runs featuring $n \in \{10, 12, \ldots, 48, 50\}$ points to solve **P0** (Eq. 9), in an elitist configuration as described in Section 3.2. The Hypervolume Indicator measures the size of the subspace dominated by the evolving Pareto frontier, bound from above by an arbitrary reference point. The

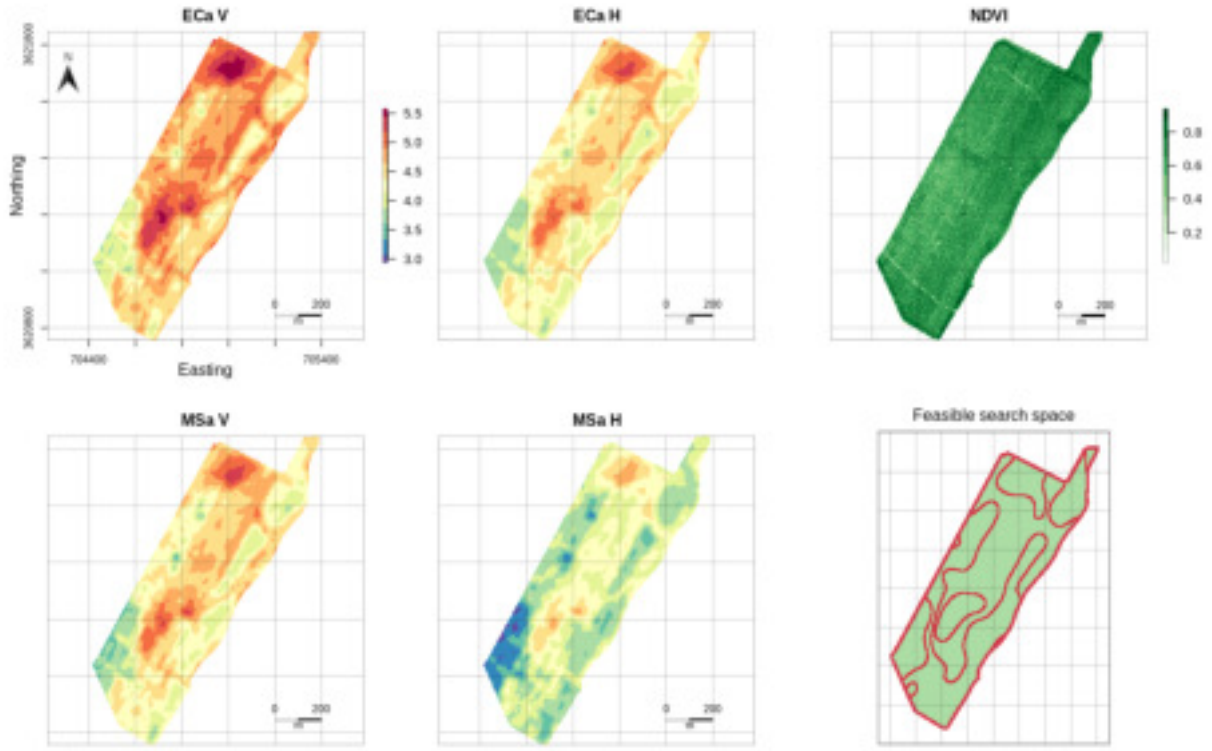

Figure 1: Ancillary data layers: ECa and MSa in vertical and horizontal modes, NDVI, and feasible search area (green).

evolution of the Hypervolume Indicator of 30 runs for sampling-plans with $n = 26$ points is depicted in Figure 2. The corresponding 300 solution points are displayed in Figure 3 over the objective space with their associated ranks, exhibiting the Pareto frontier approximation as well.

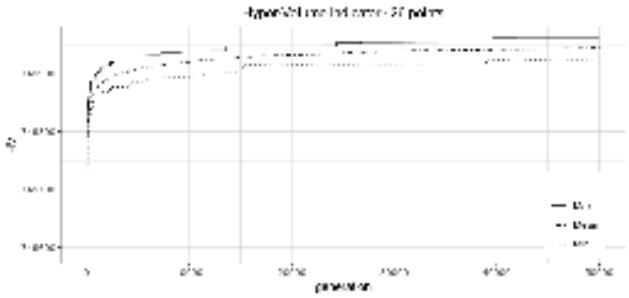

Figure 2: Evolution of the Hypervolume Indicator along 30 runs, each featuring 50,000 generations, with $n = 26$ points. HVI measures the space bounded by the attained Pareto front and a fixed point from above at each generation.

By analyzing the attained Pareto frontiers of different sample-sizes n with respect to the proposed information criteria (Figure 5), we identify a certain sample-size, $n^* = 22$, beyond which model improvements are locally deteriorating. Upon selecting this sample-size, additional two sets of solutions were generated to address an operational constraint of minimum 3 points per MZ, repeating the optimization task twice with 22 points.

Approximated Pareto-optimal solutions for $n = 22$ (Figure 4) were evaluated by the measures $MOKV$ and D_{KL} and ordered by rank accordingly (1 for the best performer, 2 for the first runner-up, and so on). Evidently, aggregated ranks (Figure 7) suggest *Sample 13* as the best candidate plan. Inspection of the plan revealed that it was not well spatially distributed, with a minimal distance between points of 27.6 m (compared to the maximal value attained in a plan of 132 m). As noticeable in Figure 4, this may be attributed to its location on one extreme of the frontier. The selection process proceeded to the first runner-up candidate, namely *Sample 21* (Figure 6), which met the requirements with a minimum distance of 88.8 m, and thus was selected as a *blueprint*.

Aftermath. Provided with this algorithmically generated sample-plan, the agricultural field has been precisely sampled according to its prescription. Currently, the research team is expecting the laboratory analyses results.

5 DISCUSSION AND SUMMARY

In this study we have demonstrated that bi-objective optimization with simultaneous targets of geographic dispersion and feature space stratification is a suitable approach for sampling design, that

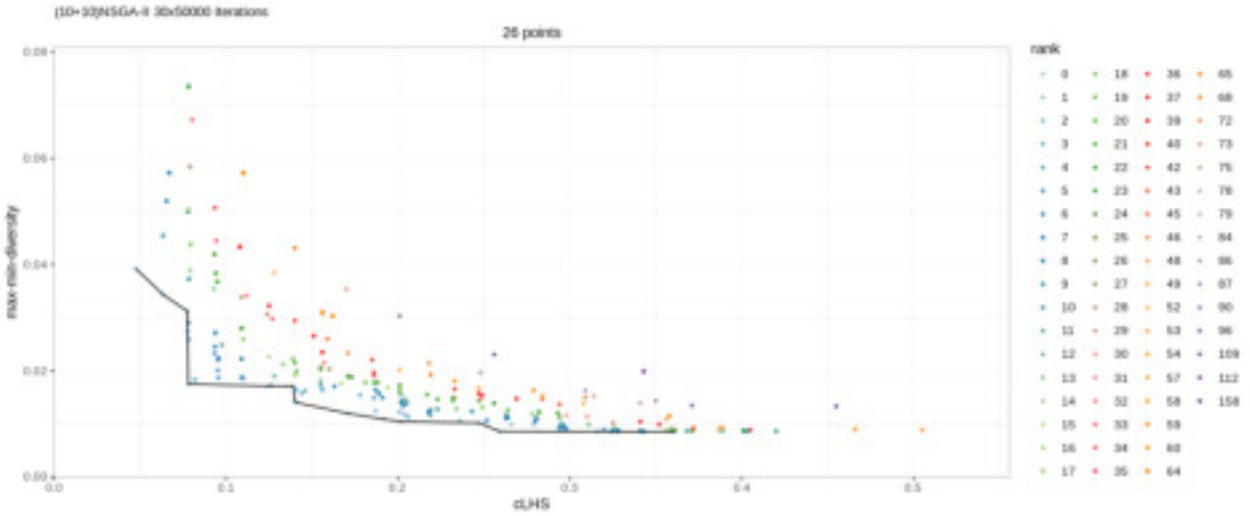

Figure 3: All solution-points, their rank and the Pareto front (rank 0, black line) obtained by optimization of cLHS and max-min-diversity with $n = 26$ points.

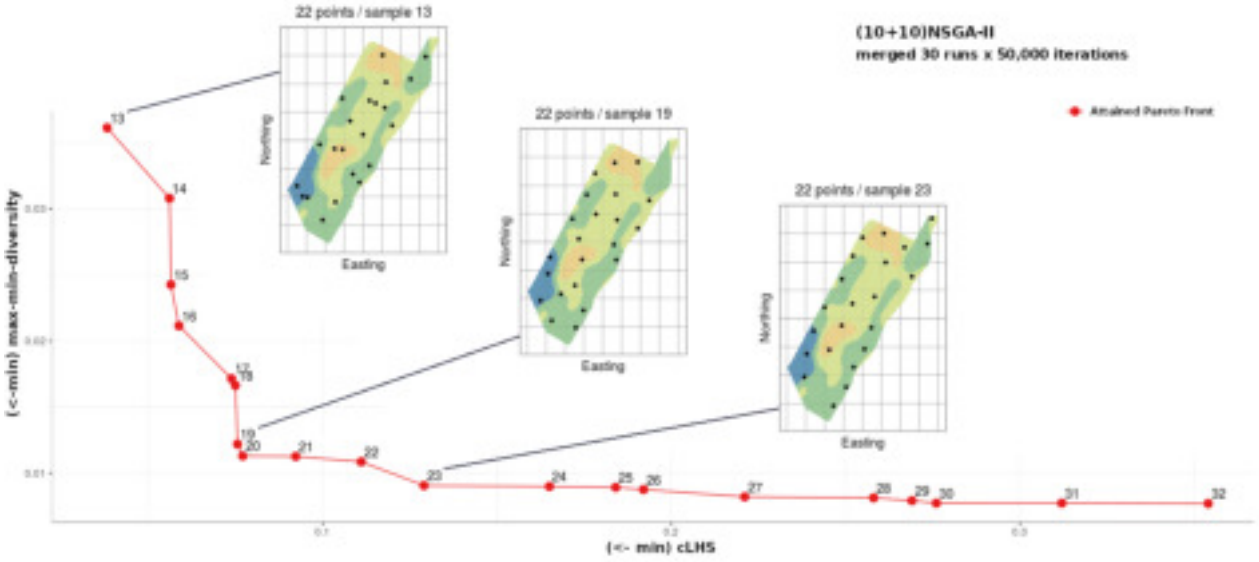

Figure 4: Attained Pareto front for samples with 22 points and some candidate solutions with their respective position.

reveals no discrepancies. In the use-case considered here, application of EMOA in a real farm with cLHS and max-min-diversity as objective functions, produced many feasible solutions, mostly found on the knee-point area of the approximated Pareto frontier – offering an apt compromise between the objectives.

A priori evaluation of the sampling-plan quality is a key for an informed process of sampling model selection. Several information criteria based on available ancillary data have been presented herein, alongside their application to qualify actual sampling schemes, providing a decision-support tool for soil-survey planning. Importantly, this approach may also be used to optimize the sampling-size n in future campaigns, leading to a more cost-effective practice of PA.

Clearly, this procedure can be improved, especially by introducing a more sophisticated variation operator, devising additional information criteria for model evaluation, and revising the selection process of candidate solutions to be included (e.g., not strictly from the frontier, but more broadly, using flexible notions such as Fuzzy Optimality [9]).

Next, we propose in detail a possible direction of future research.

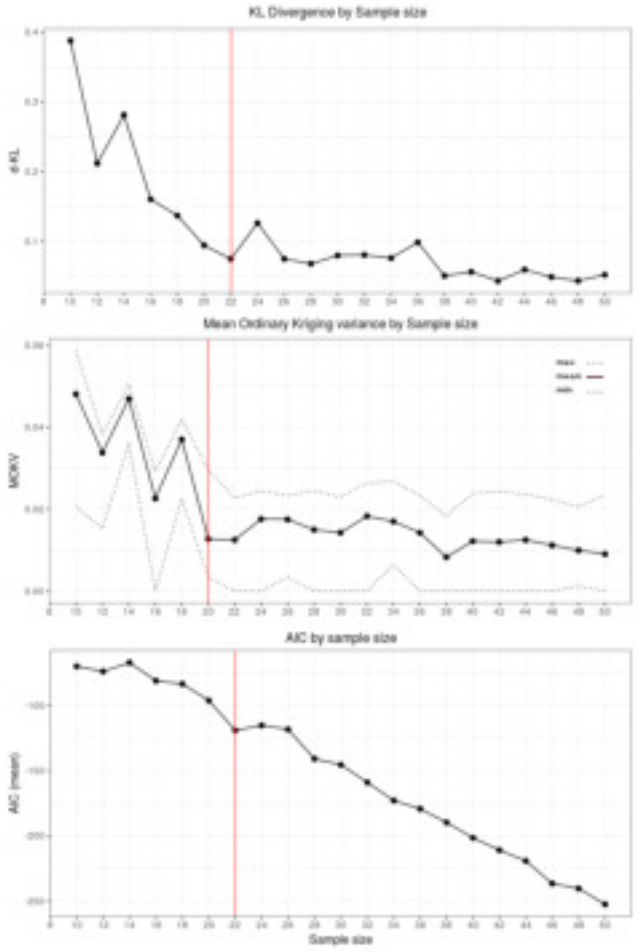

Figure 5: Statistical measures by sampling size. A knee-point is noticeable (red line) in all measures at 22 (D-KL, AIC), and 20 (MOKV), indicating a recommended sample-size of 22 points.

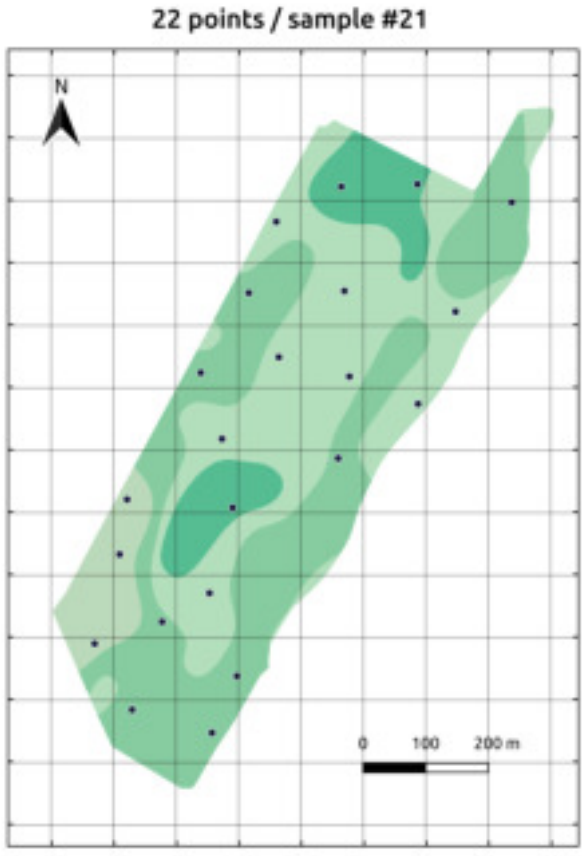

Figure 6: The selected soil sampling-plan in the field, superimposed on management zones.

5.1 Future Work: Solow-Polasky Diversity

We want to consider the so-called Solow-Polasky Diversity [24, 26] as another dispersion measure. It has been proposed in the field of biodiversity conservation as a statistical measure for the diversity of a population of individuals, given by a set of vectors in a metric space. Given pairwise distances between sites i and j, d_{ij} (either within the geographical or the feature space), let $\Psi := (\psi_{ij}) \in \mathbb{R}^{n\times n}$ be constructed with matrix elements $\psi_{ij} = \exp\left(-\gamma \cdot d_{ij}\right)$. Then, the Solow-Polasky Diversity is defined as:

$$D_{SP} = \vec{1}^T \Psi^{-1} \vec{1}, \tag{11}$$

i.e., the summation is over all the elements of Ψ^{-1}; γ is a domain-specific *normalization factor*. The Solow-Polasky Diversity strives to quantify the number of existing species within a given population. It obtains its minimum at 1, meaning that the community consists of only one species, and its maximum at n (the number of points), meaning that every aspect is a unique species. Thus, the larger this scalar, the more diverse the sampling plan is.

Importantly, in the current study, the Solow-Polasky Diversity can be applied both to the feature space as well as to the geographical space. Given a sampling plan p defined by a mapping π, two measures can be computed, using either $\left\{d^{(\mathcal{G})}_{\pi(i),\pi(j)}\right\}$ or $\left\{d^{(\mathcal{A})}_{\pi(i),\pi(j)}\right\}$, denoted as

$$D^{(\mathcal{G})}_{SP}(p),\ D^{(\mathcal{A})}_{SP}(p),$$

respectively. We then formulate another bi-objective optimization problem:

$$\boxed{\begin{array}{l}\textbf{[P1]}\\ f_3 := 1/D^{(\mathcal{A})}_{SP}(p) \longrightarrow \min\\ f_4 := 1/D^{(\mathcal{G})}_{SP}(p) \longrightarrow \min\end{array}} \tag{12}$$

Preliminary calculations indicate that this is a promising direction. At the same time, the objective functions exhibit sensitivity to the defining normalization factor γ, which requires further investigation. Another interesting approach would be to look into low discrepancy sampling methods [6], which provide the promise of small approximation errors when combined with regression models, but at the same time are computationally challenging.

Figure 7: Ranking of sampling plans with 22 points by cumulative performance of statistical measures (D-KL, MOKV).

ACKNOWLEDGMENTS

A.I. and O.M.S. thank Michal Horovitz for the fruitful discussions on Information Theory. This work was conducted at the Model Farm for Sustainable Agriculture, located at Neve Ya'ar, the northern branch of the Volcani Center, Israel's Agricultural Research Organization (ARO). The senior author would like to thank Tel-Hai College and MIGAL for the school stipend. Research was also supported by internal grants of MIGAL and by COST Action CA15140 "Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice" (ImAppNIO).

REFERENCES

[1] H. Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 6 (December 1974), 716–723. DOI: http://dx.doi.org/10.1109/TAC.1974.1100705

[2] E. Barca, A. Castrignanó, G. Buttafuoco, D. De Benedetto, and G. Passarella. 2015. Integration of electromagnetic induction sensor data in soil sampling scheme optimization using simulated annealing. *Environ Monit Assess* 187 (2015). DOI: http://dx.doi.org/10.1007/s10661-015-4570-y

[3] J. Bossek. 2017. Ecr 2.0: A Modular Framework for Evolutionary Computation in R. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion (GECCO '17)*. ACM, New York, NY, USA, 1187–1193. DOI: http://dx.doi.org/10.1145/3067695.3082470

[4] M. Córdoba, C. Bruno, J. Costa, N. R. Peralta, and M. Balzarini. 2016. Protocol for multivariate homogeneous zone delineation in precision agriculture. *Biosystems Engineering* 143 (03 2016), 95–107. DOI: http://dx.doi.org/10.1016/j.biosystemseng.2015.12.008

[5] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. 2002. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *Trans. Evol. Comp* 6, 2 (April 2002), 182–197. DOI: http://dx.doi.org/10.1109/4235.996017

[6] C. Doerr, M. Gnewuch, and M. Wahlström. 2014. *Calculation of Discrepancy Measures and Applications.* Springer International Publishing, Cham, Switzerland, 621–678. DOI: http://dx.doi.org/10.1007/978-3-319-04696-9_10

[7] M. Ehrgott. 2005. *Multicriteria Optimization* (second ed.). Springer, Berlin.

[8] M. T. M. Emmerich and A. H. Deutz. 2018. A tutorial on multiobjective optimization: fundamentals and evolutionary methods. *Natural Computing* 17, 3 (01 Sep 2018), 585–609. DOI: http://dx.doi.org/10.1007/s11047-018-9685-y

[9] M. Farina and P. Amato. 2003. Fuzzy Optimality and Evolutionary Multiobjective Optimization. In *Evolutionary Multi-Criterion Optimization.* Lecture Notes in Computer Science, Vol. 2632. Springer Berlin Heidelberg, 58–72.

[10] B. Gao, Y. Pan, Z. Chen, F. Wu, X. Ren, and M. Hu. 2016. A Spatial Conditioned Latin Hypercube Sampling Method for Mapping Using Ancillary Data. *Transactions in GIS* 20, 5 (2016), 735–754. DOI: http://dx.doi.org/10.1111/tgis.12176

[11] T. Hengl, D. Rossiter, and A. Stein. 2003. Soil Sampling Strategies for Spatial Prediction by Correlation with Auxiliary Maps. *Australian Journal of Soil Research 41 (2003) 8* 41 (01 2003). DOI: http://dx.doi.org/10.1071/SR03005

[12] G. B. M. Heuvelink and E. J. Pebesma. 2002. Is the ordinary kriging variance a proper measure of interpolation error?

[13] R. Kerry and M. A. Oliver. 2003. Variograms of Ancillary Data to Aid Sampling for Soil Surveys. *Precision Agriculture* 4, 3 (01 Sep 2003), 261–278. DOI: http://dx.doi.org/10.1023/A:1024952406744

[14] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *Ann. Math. Statist.* 22, 1 (03 1951), 79–86. DOI: http://dx.doi.org/10.1214/aoms/1177729694

[15] C. Kuo, F. Glover, and K. Dhir. 1993. Analyzing and Modeling the Maximum Diversity Problem by Zero-One Programming. *Decision Sciences* 24, 6 (1993), 1171–1185. DOI: http://dx.doi.org/10.1111/j.1540-5915.1993.tb00509.x

[16] R. M. Lark. 2016. Multi-objective optimization of spatial sampling. *Spatial Statistics* 18 (2016), 412 – 430. DOI: http://dx.doi.org/https://doi.org/10.1016/j.spasta.2016.09.001

[17] G. Matheron. 1965. *Les Variables régionalisées et leur estimation: une application de la théorie des fonctions aléatoires aux sciences de la nature.* Masson, Paris.

[18] B. Minasny and A. B. McBratney. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers and Geosciences* 32, 9 (2006), 1378 – 1388. DOI: http://dx.doi.org/10.1016/j.cageo.2005.12.009

[19] I. O. A. Odeh, D. J. Chittleborough, and A. B. McBratney. 1992. Soil Pattern Recognition with Fuzzy-c-means: Application to Classification and Soil-Landform Interrelationships. *Soil Science Society of America Journal* 56 (1992), 505–516. DOI: http://dx.doi.org/10.2136/sssaj1992.03615995005600020027x

[20] E. J. Pebesma. 2004. Multivariable geostatistics in S: the gstat package. *Computers & Geosciences* 30 (2004), 683–691.

[21] J. Reyes, O. Wendroth, C. Matocha, J. Zhu, W. Ren, and A. D. Karathanasis. 2018. Reliably Mapping Clay Content Coregionalized with Electrical Conductivity. *Soil Science Society of America Journal* 82 (05 2018). DOI: http://dx.doi.org/10.2136/sssaj2017.09.0327

[22] P. Roudier, D. Beaudette, and A Hewitt. 2012. *A conditioned Latin hypercube sampling algorithm incorporating operational constraints.* CRC Press, 227–232. DOI: http://dx.doi.org/10.1201/b12728-46

[23] E. Scudiero, P. Teatini, G. Manoli, F. Braga, T. H. Skaggs, and F. Morari. 2018. Workflow to Establish Time-Specific Zones in Precision Agriculture by Spatiotemporal Integration of Plant and Soil Sensing Data. *Agronomy* 8, 11 (2018). DOI: http://dx.doi.org/10.3390/agronomy8110253

[24] A. Solow and S. Polasky. 1994. Measuring biological diversity. *Environmental and Ecological Statistics* 1 (1994), 95–103. Issue 2. DOI: http://dx.doi.org/10.1007/BF02426650

[25] M. L. Stein. 2012. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer New York. DOI: http://dx.doi.org/10.1007/978-1-4612-1494-6

[26] Tamara Ulrich, Johannes Bader, and Lothar Thiele. 2010. Defining and Optimizing Indicator-Based Diversity Measures in Multiobjective Search. In *PPSN-XI.* 707–717. DOI: http://dx.doi.org/10.1007/978-3-642-15844-5_71

[27] R. Webster and M. A. Oliver. 2007. *Geostatistics for Environmental Scientists, Second Edition.* John Wiley and Sons Ltd., Chichester, England. DOI: http://dx.doi.org/10.1002/9780470517277.ch1

[28] Y. Zhao, X. Xu, K. Tian, B. Huang, and N. Hai. 2016. Comparison of sampling schemes for the spatial prediction of soil organic matter in a typical black soil region in China. *Environmental Earth Sciences* 75 (12 2016). DOI: http://dx.doi.org/10.1007/s12665-015-4895-4