How Much Is Too Much and How Little Is Too Little?

Národní výsledky evropského variantního testování stratifikovaného alokačního algoritmu pro optimalizace vzorkovacích sítí

Jan Skála a Daniel Žížala

VÚMOP
VÝZKUMNÝ ÚSTAV MONITORINGU A OCHRANY PŮDY, V.V.I.

Introduction

Soil Monitoring Law (SML)

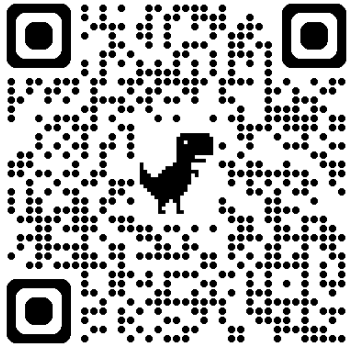
I. unified sample allocation
II. unified field sampling protocol
III. standardized set of indicators
IV. maximal possible harmonization of laboratory methods

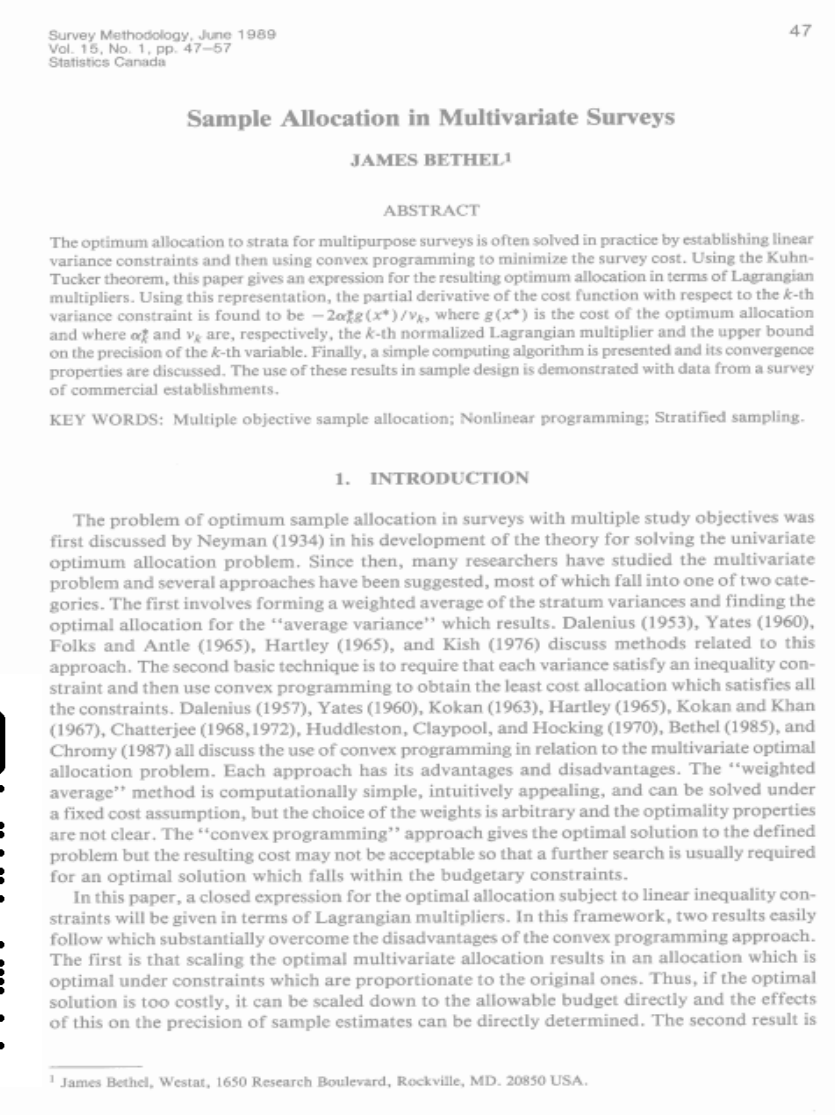
Part A of Annex II – obligatory criteria

I. stratified random sampling
II. stratified upon soil units/districts according Article 4(2).
III. required maximum estimation error within soil units
IV. best available information on the soil properties distribution

SML suggested methodology

- Bethel‘s algorithm (*Bethel, 1989*) = statistical determination of the minimum sample size and the allocation of those samples across different strata accounting the constrain of maximal require 5% error





Survey Methodology, June 1989
Vol. 15, No. 1, pp. 47–57
Statistics Canada

Sample Allocation in Multivariate Surveys

JAMES BETHEL[1]

ABSTRACT

The optimum allocation to strata for multipurpose surveys is often solved in practice by establishing linear variance constraints and then using convex programming to minimize the survey cost. Using the Kuhn-Tucker theorem, this paper gives an expression for the resulting optimum allocation in terms of Lagrangian multipliers. Using this representation, the partial derivative of the cost function with respect to the k-th variance constraint is found to be $-2\alpha_k^* g(x^*)/v_k$, where $g(x^*)$ is the cost of the optimum allocation and where α_k^* and v_k are, respectively, the k-th normalized Lagrangian multiplier and the upper bound on the precision of the k-th variable. Finally, a simple computing algorithm is presented and its convergence properties are discussed. The use of these results in sample design is demonstrated with data from a survey of commercial establishments.

KEY WORDS: Multiple objective sample allocation; Nonlinear programming; Stratified sampling.

1. INTRODUCTION

The problem of optimum sample allocation in surveys with multiple study objectives was first discussed by Neyman (1934) in his development of the theory for solving the univariate optimum allocation problem. Since then, many researchers have studied the multivariate problem and several approaches have been suggested, most of which fall into one of two categories. The first involves forming a weighted average of the stratum variances and finding the optimal allocation for the ‘‘average variance’’ which results. Dalenius (1953), Yates (1960), Folks and Antle (1965), Hartley (1965), and Kish (1976) discuss methods related to this approach. The second basic technique is to require that each variance satisfy an inequality constraint and then use convex programming to obtain the least cost allocation which satisfies all the constraints. Dalenius (1957), Yates (1960), Kokan (1963), Hartley (1965), Kokan and Khan (1967), Chatterjee (1968,1972), Huddleston, Claypool, and Hocking (1970), Bethel (1985), and Chromy (1987) all discuss the use of convex programming in relation to the multivariate optimal allocation problem. Each approach has its advantages and disadvantages. The ‘‘weighted average’’ method is computationally simple, intuitively appealing, and can be solved under a fixed cost assumption, but the choice of the weights is arbitrary and the optimality properties are not clear. The ‘‘convex programming’’ approach gives the optimal solution to the defined problem but the resulting cost may not be acceptable so that a further search is usually required for an optimal solution which falls within the budgetary constraints.

In this paper, a closed expression for the optimal allocation subject to linear inequality constraints will be given in terms of Lagrangian multipliers. In this framework, two results easily follow which substantially overcome the disadvantages of the convex programming approach. The first is that scaling the optimal multivariate allocation results in an allocation which is optimal under constraints which are proportionate to the original ones. Thus, if the optimal solution is too costly, it can be scaled down to the allowable budget directly and the effects of this on the precision of sample estimates can be directly determined. The second result is

[1] James Bethel, Westat, 1650 Research Boulevard, Rockville, MD. 20850 USA.

Bethel, J.W. (1989). Sample Allocation in Multivariate Surveys. Survey Methodology, Vol. 15, pp. 47-57

Implementation

- modern numerical solution for Bethel method using genetic algorithms exploring the universe of the possible alternative stratifications (*Ballin & Barcaroli, 2014*)

- implemented within R package „*SamplingStrata*“

SCAN ME for R package „SamplingStrata“

SCAN ME for manuscript

Ballin, M., & Barcaroli, G. (2014). Joint determination of optimal stratification and sample allocation using genetic algorithm. Survey Methodology, 39 , 369-393.

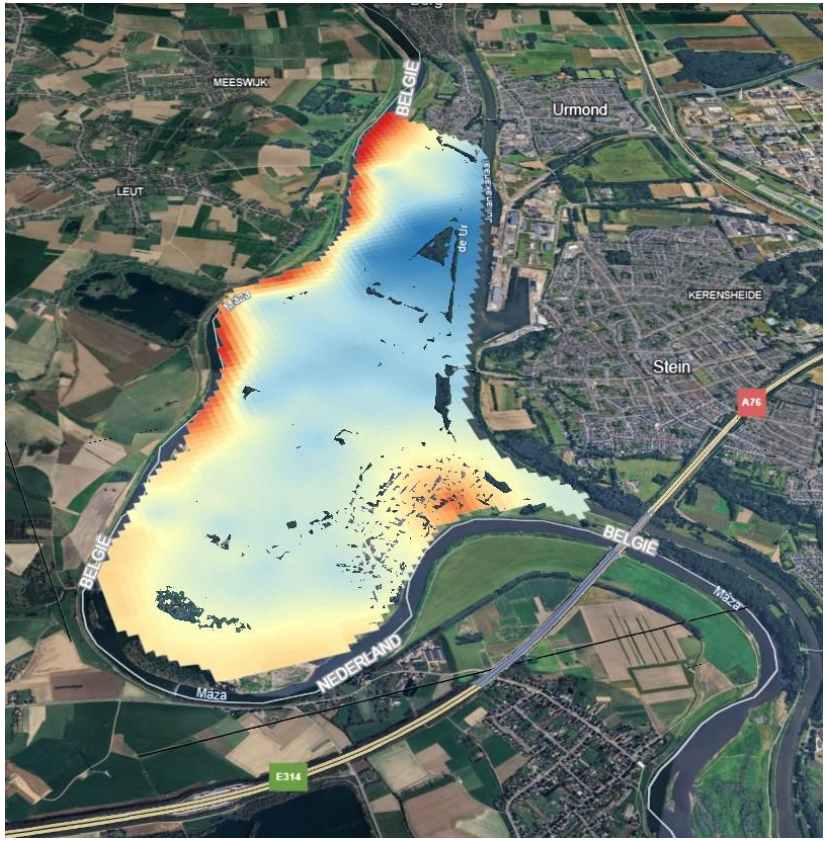
Initial conditions for test – Meuse dataset (NL)

- 1 domain
- stratification layers = 1 – 4 grids (Cd, Cu, Pb, Zn from universal kriging)

- optimal number of strata: k-means clustering
- sample allocation within strata: genetic algorithm

- required estimation error = as coefficient of variation =< 5 % in domain

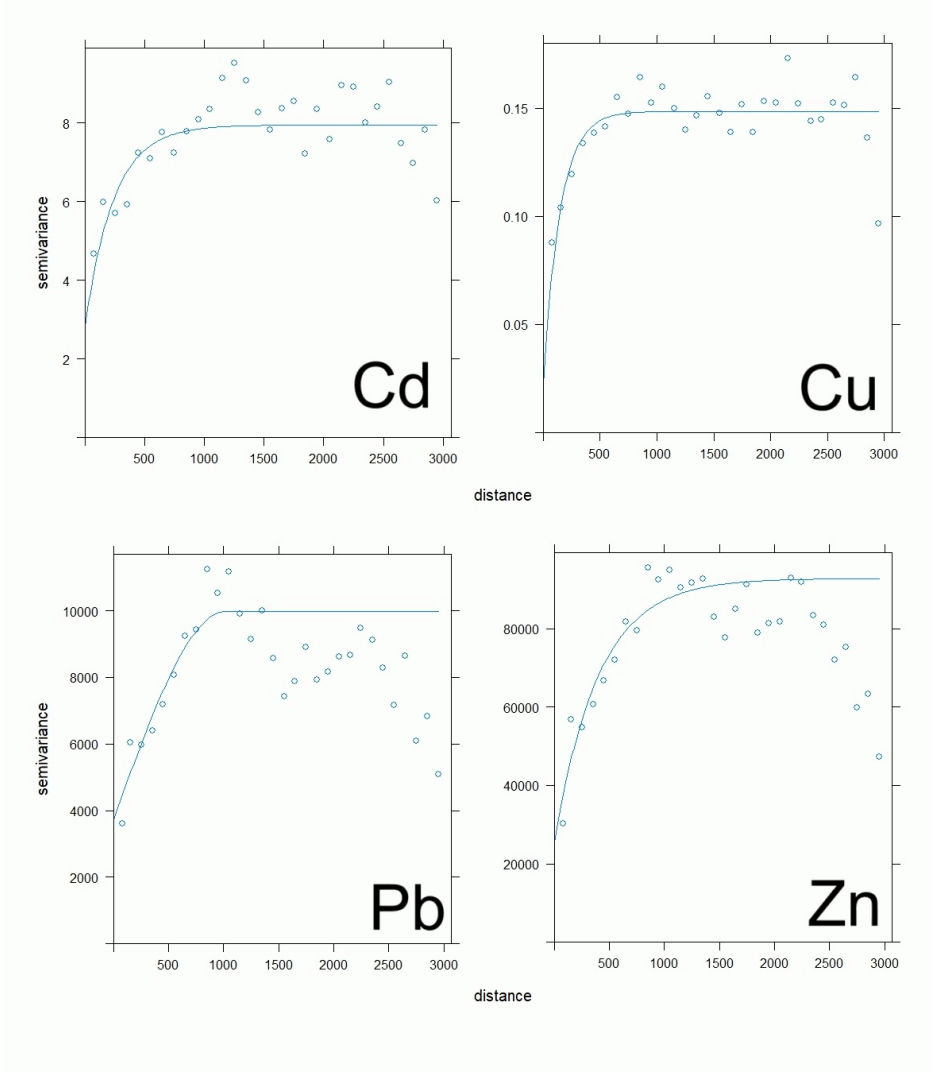
155 samples - trace elements (Cd, Cu, Pb, Zn)



Source Hengl T. (2014): plotKML: Scientific Visualization of Spatio-temporal Data. Journal of Statistical Software

Results

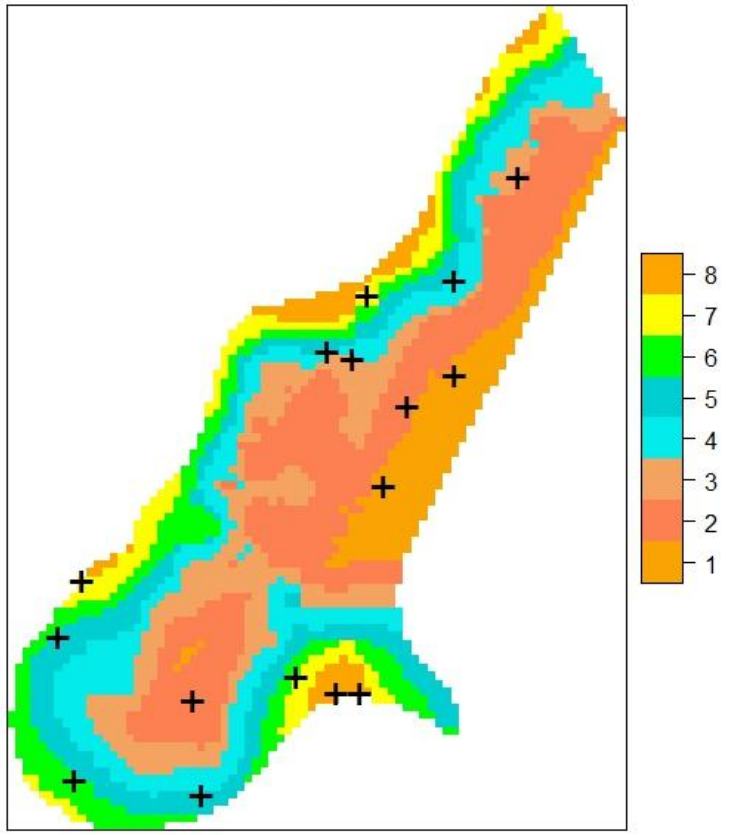
- 1 grid: Cu grid – 10 samples
- 2 grids: Cu and Zn grids – 15 samples
- 3 grids: Cu, Zn, and Pb – 16 sample
- 4 grids: Cu, Zn, Pb, and Cd – 16 samples



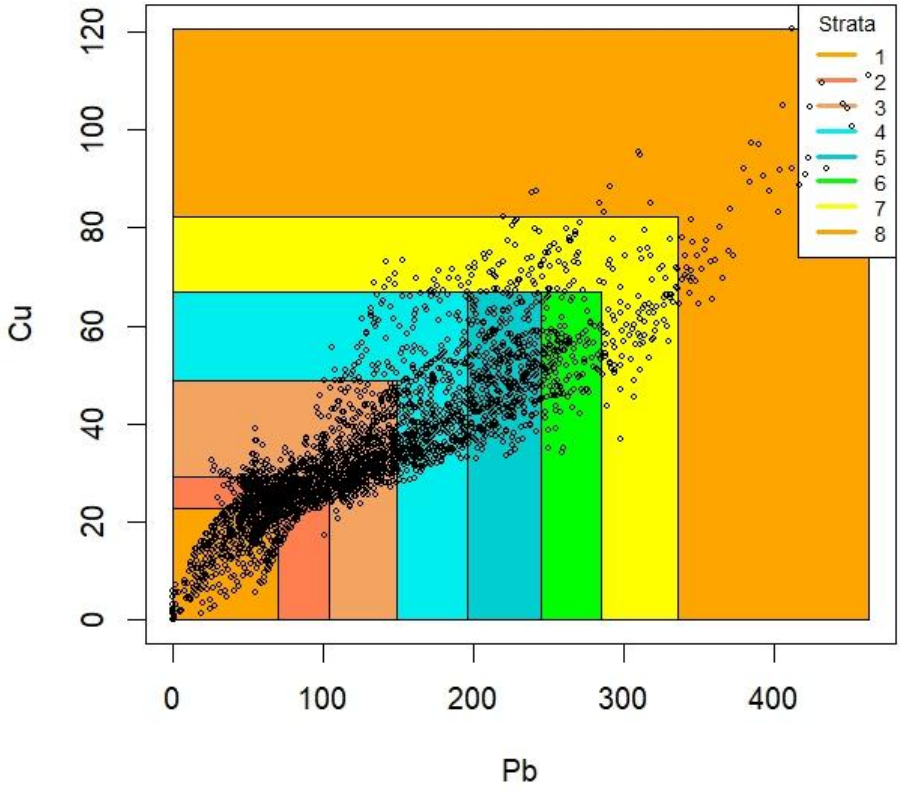
Results

Domain	Stratum	Population	Allocation	SamplingRate
1	1	294	2	0.006803
1	2	781	2	0.002622
1	3	469	2	0.004264
1	4	459	2	0.004645
1	5	414	2	0.006004
1	6	343	2	0.006602
1	7	230	2	0.008696
1	8	113	2	0.017699

Stratification using 4 grids and new samples



Strata boundaries in domain 1




```
> table_CV
      Cd.pred   Cu.pred   Pb.pred   Zn.pred
DOM1 0.0488414 0.0373385 0.0479819 0.0364119
>
```

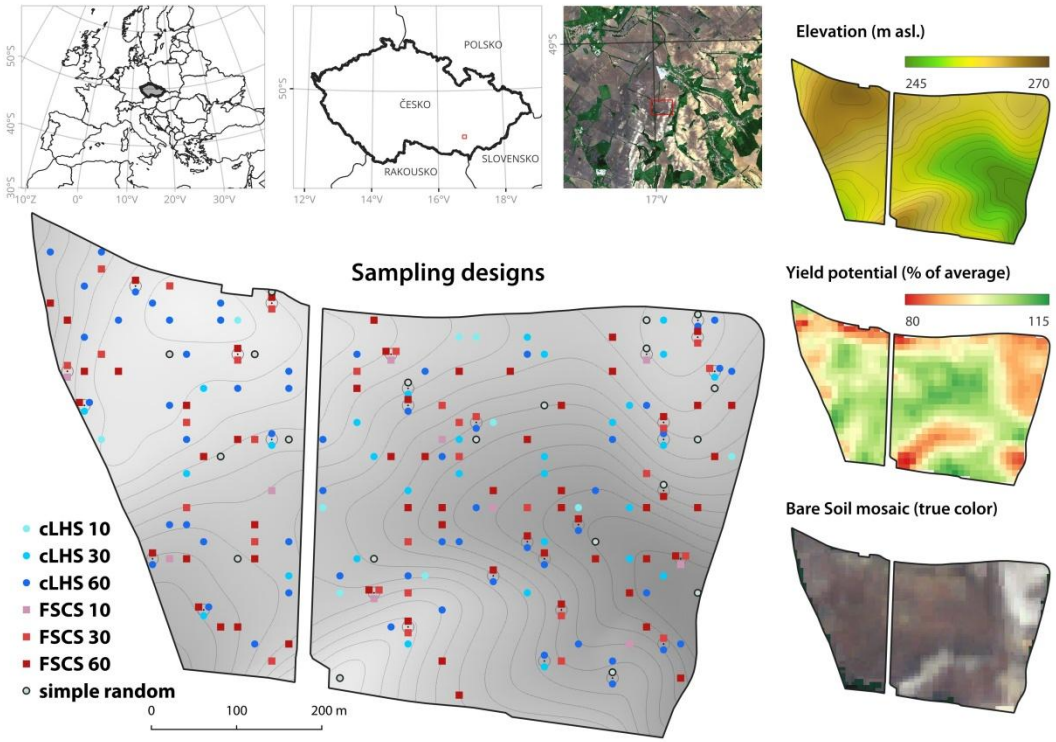
Initial conditions– Czech field study

- 1 domain
- stratification layers = 8 grids (kriging interpolation)

- optimal number of strata: k-means clustering
- sample allocation within strata: genetic algorithm

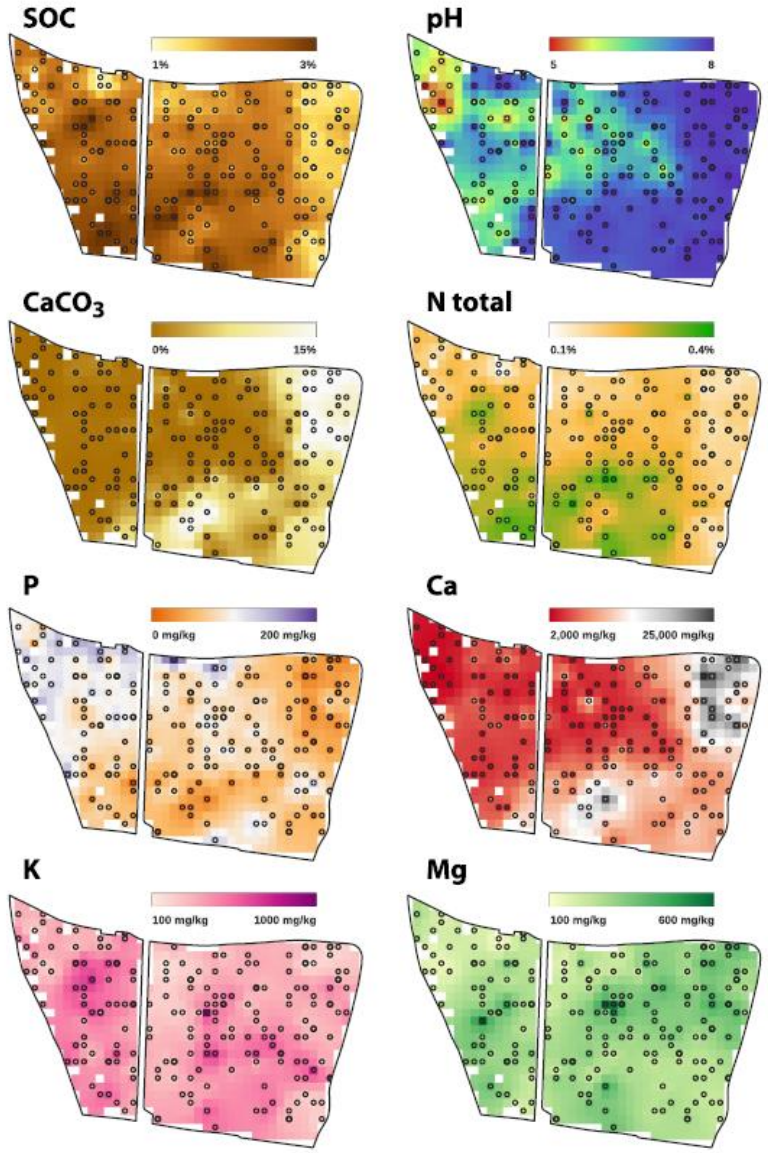
- coefficient of variation =< 5 % in domain

220 samples:
nutrients (N, P, K, Mg, Ca), SOC, pH, $CaCO_3$



Results – 39 samples

```
> table_CV
     P.pred    K.pred   Mg.pred Ca.pred   pH.pred Ntot.pred  Cox.pred
DOM1   0.05 0.0432898 0.0405533    0.05 0.0126584 0.0291314 0.0333388
> |
```



European test

- moving from local to national/continental scale = need for algorithm adaptation
- pan European testing of several scenarios of adaptation under coordination of the Belgian presidency (2024) and support of Joint Research Centre (JRC)

		CV	EU/REGIONAL	Layers Bold: descriptors part A and B, Purple: descriptors part C	Starting samples	CLHS Sample	Random sampling / Existing monitoring points + random sampling	Wheight of existing points in the total number of points	Resolution	Domain
	Test 1 (default):	(0,		, BD10, texture	(as domain detail, number should be high enough ; test for BE implied x points; to be stated in the text) - linked to the number able to repeat the variability	(as domain detail, number should be high enough ; test for BE implied x points; to be stated in the text)	Random		100x100	NUTS2 x CLC (which given aggregation used) For urban areas, a second step would be needed in order to adjust the sampling (replace CLC with local data, specific weight for sampling points)
A	**Test 2 :**	0,05	EU	OC, pH (H20), P, N, BD10, texture	same as test 1	same as t			**500X500**	same as test 1
A	**Test 3 :**	0,05	EU	OC, pH (H20), P, N, BD10, texture **+ CEC, AWC**	same as test 1	same as t			500X500	same as test 1
A	**Test 4 :**	0,05	EU	OC, pH (H20), P, N, BD10, texture	same as test 1	same as t			500X500	**NUTS1 + CLC + urban area option**
B	**Test 1 :**	0,05	Best ava ble	**Replace EU by national/subnational if available, even if national/subnational l don't target the same descriptors than i**		same as t		/	100X100	same as test 1
B	**Test 2 :**	0,05	Best av	**ational if bnational tors than i**		same as test 1	Random		500X500	same as test 1
B	**Test 3 :**	0,05	Best av	**ational if bnational layers ptors than initial set**	same as test 1	same as test 1	**Random + existing monitoring points**	**50%**		t 1
B	**Test 4 :**	0,05	Best available	**Replace EU by national/subnational if available, even if national/subnational layers don't target the same descriptors than initial set**	same as test 1					ndcover, ional Soil nal area
	"Best available" TEST	0,05	Best available	**Replace EU by national/subnational if available, even if national/subnational layers don't target the same descriptors than initial set**	same as test 1		sting monitoring points	**50%**		ndcover, ional Soil gional **Bioclimatic area**

Constant required error (CV=< 5%)

Various quality of data inputs (both stratification grids and data for soul unit definition)

Condition to retain existing national monitoring sites (for CZ BMP and ICP Forest)

Various lists of grids for stratification

Source of data: JRC vs best available national data

Numerical adaptation of the workflow using cLHS

Domain changes – various definitions of soil units

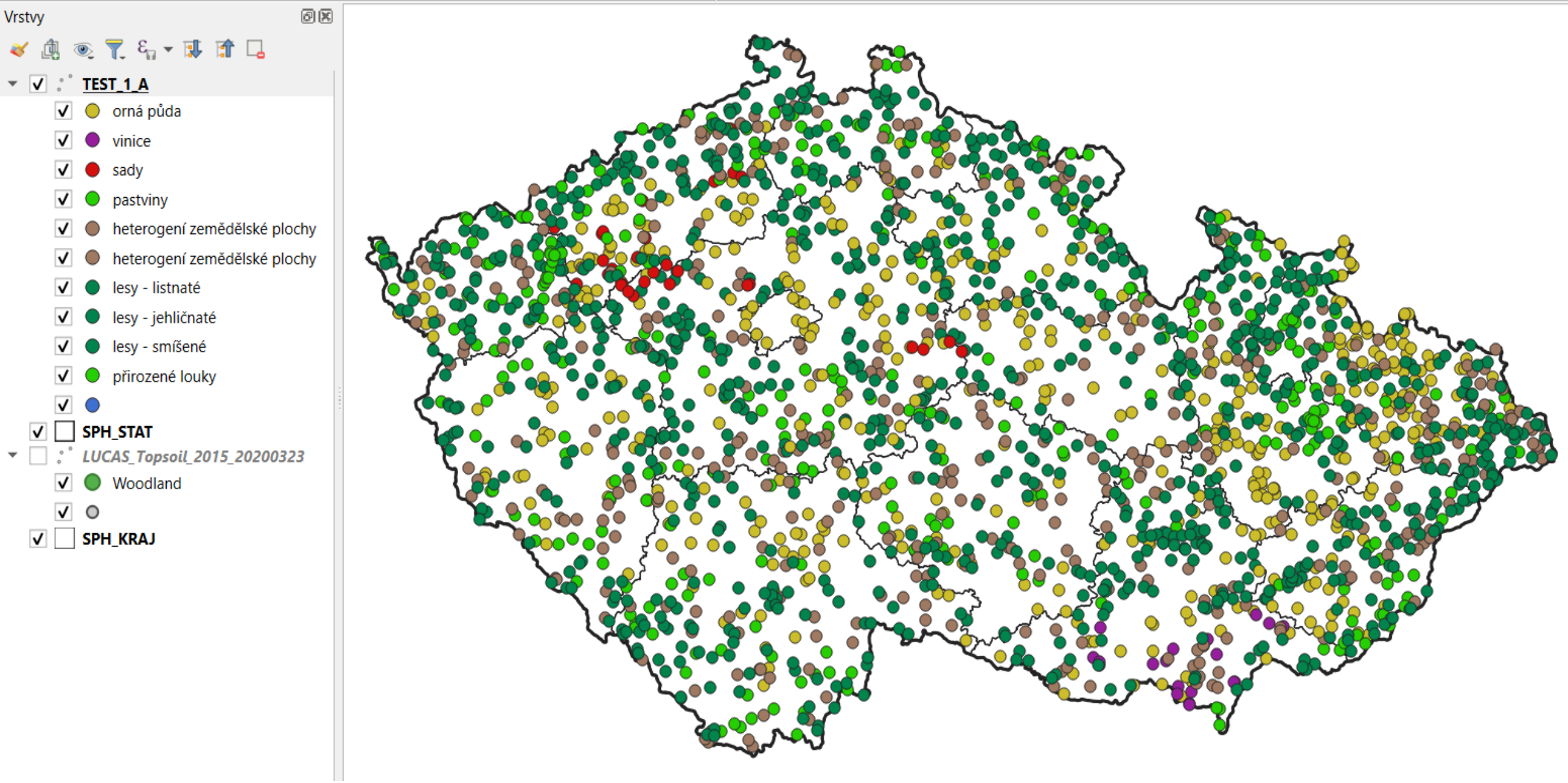
	A1	B5
AT	2665	4037
BE	1728	2940
CZ	2471	5049
DE	–	–
DK	1240	993
ES	886	–
EL	–	–
EE	718	2353
FR	4700	–
FI	1302	1943
IT	9199	2839
HU	3000	7522
IE	350	–
HR	–	–
MT	–	–
NL	3520	–
LU	211	195
LT	1427	2940
LV	722	–
PL	6051	–
SE	5305	–
RO	–	–
PT	–	–
SK	–	–
SI	520	–

Table 1. Total sample size for all EU-test scenarios from 2 rounds of simulation with changing input conditions

* JRC-based scenario with default LUCAS grids (A) Combination of best national data and JRC data LUCAS-based data (B)
** initial simulation conditions in terms of allocation of input random samples and cLHS-based filtering of data

	Test 1		Test 2		Test 3		Test 4		Test 5	
	A*	B*	A*	B*	A*	B*	A*	B*	A*	B*
210 000/15 000**	2471	2014	1311	2199	1357	2597	384	3185	2471	3888
400 000/30 000**	3361	2815	3330	2977	3400	2604	842	4816	3361	5049

Table 2. Total sample size for 2 scenarios (JRC data = A1 vs B5 = best national data) for EU-member states


Vrstvy
TEST_1_A
orná půda
vinice
sady
pastviny
heterogení zemědělské plochy
heterogení zemědělské plochy
lesy - listnaté
lesy - jehličnaté
lesy - smíšené
přirozené louky
SPH_STAT
LUCAS_Topsoil_2015_20200323
Woodland
SPH_KRAJ

Concluding notes

The main driving factors for final size and allocation of samples for Bethel implementation

1	• The definition of soil units – factors and the level of detail for their spatial grids	Very high influence
2	• The input grids for stratification (resolution, source, and quality of background model)	High influence
3	• The number of grids used for stratification – which soil descriptors should be used	Middle influence
4	• The algorithm adaptation (cLHS) for initial filtering of possible pixel for allocation	Middle influence
5	• Forcing the algorithm to retain existing monitoring points (BMP and ICP Forrest)	Lower effect
6	• The genetic algorithm tuning	Unknown?

Thank you for attention

T A
Č R

Tento projekt je spolufinancován se státní podporou Technologické agentury ČR a Ministerstva životního prostředí v rámci **Programu Prostředí pro život.**

www.tacr.cz www.mzp.cz

skala.jan@vumop.cz
zizala.daniel@vumop.cz

VÚMOP
VÝZKUMNÝ ÚSTAV MONITORINGU A OCHRANY PŮDY, V.V.I.

Podpořeno projektem č. SQ01010088 financovaným se státní podporou Technologické agentury ČR a Ministerstva životního prostředí v programu Prostředí pro život.