MCMC based Spatially Balanced selection algorithms

Roberto Benedetti

University of Chieti-Pescara, Italy

Federica Piersimoni

ISTAT, Italian National Statistical Institute, Rome, Italy

Francesco Pantalone

University of Perugia, Italy

Background

The decomposition lemma, states that (Knottnerus, 2003, p. 87):

$$\sigma^2_{\bar{y}} = V_s\left(\bar{\bar{y}}_s\right) + \frac{n-1}{n} E_s\left(S^2_{\bar{y},s}\right)$$

It can be seen that the HT estimator can be more efficient by setting the first-order inclusion probabilities in such a way that y_k/π_k is approximately constant and/or by defining a design $p(s)$ that increases the expected within sample variance. The intuitive explanation for this is that if a sample s contains as much information as possible, the uncertainty in the estimation process is clearly reduced to zero. This consideration suggests that we should find a rule that makes the probability $p(s)$ of selecting a sample s proportional, or more than proportional, to its variance S^2. This variance is unknown, because it is relative to the target, unobserved variable **y**. Thus, this is a purely theoretical topic unless we can find auxiliary information for s.

When dealing with spatially distributed populations, a promising candidate for this rule is the distance between units, as evidenced in spatial interpolation literature (Ripley 1981, Cressie 1993). This is because it is often highly related to the variance of variables observed on a set of geo-referenced units.

Some proposals

- Balanced Sampling by means of Simulated Annealing
- Spatially Balanced Sampling Proportional to the Within Sample Distance
- Doubly Balanced Sampling by means of Constrained Simulated Annealing

Balanced Sampling by means of Simulated Annealing

We propose an efficient algorithm to select balanced samples by using a MCMC method, such as Simulated Annealing (SA).

SA is a stochastic optimization method proposed by Kirkpatrick *et al.* (1983) and Černý (1985) for finding a global minimum of a function. The method is a generalization of the Metropolis-Hastings algorithm (Metropolis *et al*. 1953) in the optimization context and it represents one of the most popular and used optimization strategies to solve complex combinatorial problems.

The SA algorithm is an iterative procedure where each step consists in running a non-homogeneous Markov-Chain, with the temperature T being reduced at each step.

Selection Algorithm: Balanced Sampling

The proposed algorithm can be summarized as follows. The procedure starts at iteration t=0, with an initial point s(0), randomly selected from according to a SRS with constant inclusion probabilities. In a generic iteration t the elements of s(t) are updated in the following steps:

1. select at random two units included and not included in the sample in the previous iteration, say *i* and *j*. Respectively among the units within the sample, and another among the units outside the sample;
2. denote with $^{*}s^{(t)}$ the sample where the units in the position i and j exchange their status. Randomly decide whether or not to adopt this sample:

Selection Algorithm

$$s^{(t+1)} = \begin{cases} {}^{*}s^{(t)} & \text{with probability } p = \min\left\{1, \exp\{-[D({}^{*}s^{(t)}) - D(s^{(t)})]/T\}\right\} \\ s^{(t)} & \text{otherwise} \end{cases}$$

3. repeat steps (1) and (2) mq times (in our application and simulations we used m and q constantly equal respectively to N and 10).

The algorithm is very simple to be implemented and extremely flexible by simply modifying the criteria used in the first step to select the two candidate units i and j.

T is the temperature that decreases with the augmenting of the iterations, i.e. at the beginning there is a high probability to accept worsenings of the objective function; at the end, when t→0 only reductions of the objective function are accepted.

D(s) is the maximum percentage difference $\left(\sum_{k \in s} d_k x_{kj} - t_{x_j}\right) / t_{x_j}$ (among the j covariates) that we have in the linear constraints for sample s.

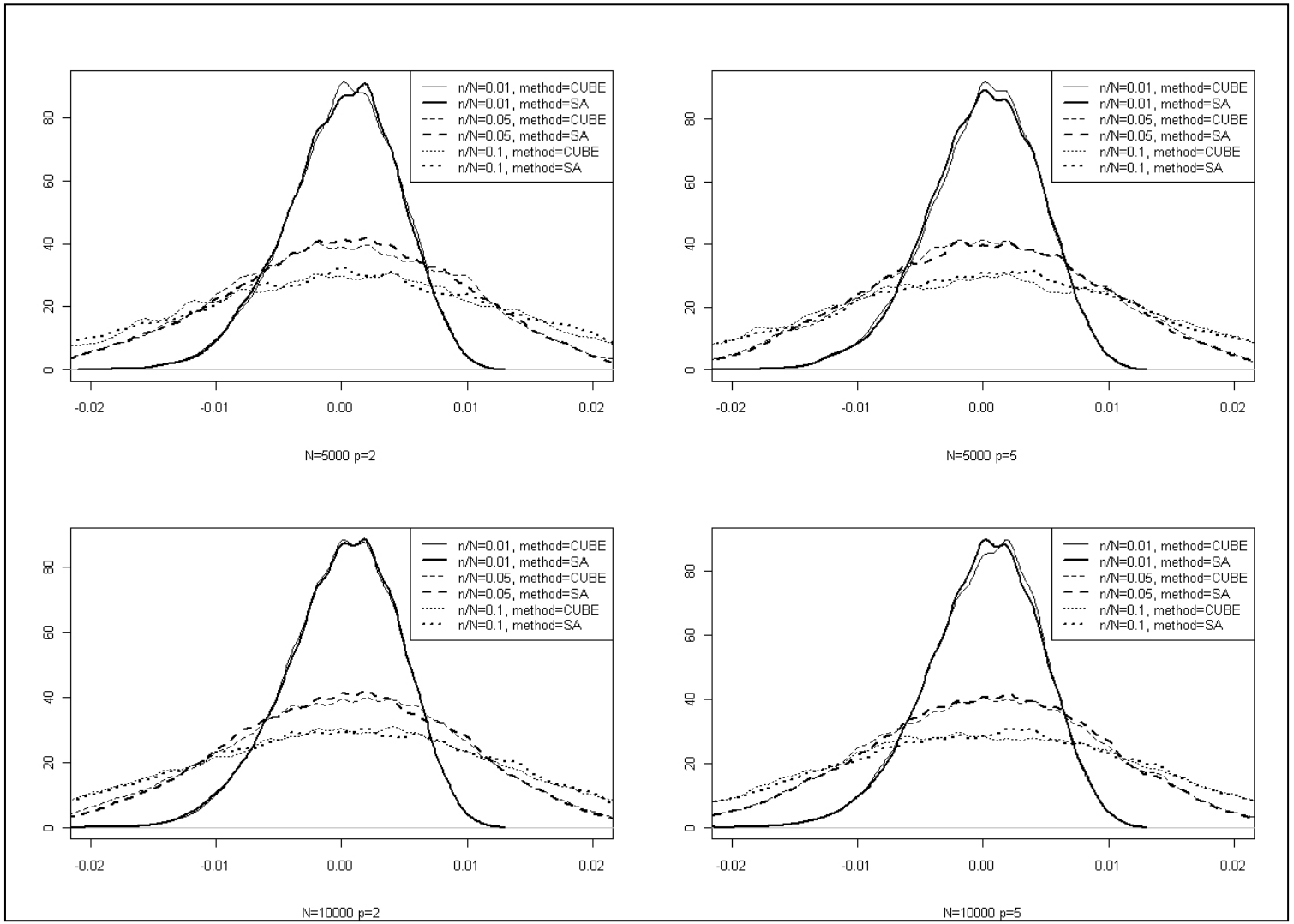
A Simulation Experiment

Evaluation the respect of the target inclusion probabilities and quality of balancing

We analize six experimental situations, in which values are generated using both an exponential variable and a uniform variable. For each situation a set of 500 independent balanced samples with equal probably and with unequal probability have been selected for sampling rates equal to 0.01, 0.05 and 0.1.
The unequal inclusion probabilities are generated by using both a uniform variable and an exponential variable. The comparison has been implemented between Cube method in its fast implementation (Chauvet and Tillé 2006) and the proposed algorithm.

A Simulation Experiment

Distribution of the differences between the inclusion probabilities and frequencies when the sampling design is a balanced sampling design with equal probability



A Simulation Experiment

Accuracy of the inclusion probabilities

N	Number of constrains	n/N=0.01		n/N=0.05		n/N=0.1	
		Cube	**SA**	**Cube**	**SA**	**Cube**	**SA**
		Equal Probability Sampling					
5000	2	0,196	0,193	0,196	0,190	0,180	0,177
	3	0,201	0,196	0,194	0,191	0,178	0,188
	5	0,199	0,194	0,184	0,190	0,182	0,179
10000	2	0,195	0,199	0,195	0,185	0,179	0,177
	3	0,198	0,200	0,192	0,194	0,182	0,181
	5	0,201	0,199	0,193	0,185	0,178	0,184
		Unequal Probability Sampling: Uniform					
5000	2	0,205	0,222	0,183	0,208	0,183	0,194
	3	0,195	0,212	0,189	0,198	0,174	0,186
	5	0,203	0,219	0,192	0,207	0,178	0,191
10000	2	0,197	0,210	0,187	0,194	0,180	0,186
	3	0,196	0,201	0,187	0,200	0,188	0,187
	5	0,199	0,207	0,192	0,197	0,178	0,192
		Unequal Probability Sampling: Exponential					
5000	2	0,203	0,225	0,195	0,216	0,175	0,207
	3	0,203	0,234	0,188	0,210	0,173	0,210
	5	0,197	0,236	0,186	0,214	0,179	0,206
10000	2	0,192	0,213	0,186	0,208	0,183	0,192
	3	0,199	0,216	0,190	0,205	0,182	0,194
	5	0,199	0,215	0,191	0,204	0,186	0,198

A Simulation Experiment

Efficiency of the balancing

		Uniform				Exponential			
		Equal Probability		Unequal Probability		Equal Probability		Unequal Probability	
N	n	Cube	SA	Cube	SA	Cube	SA	Cube	SA
2,000	100	0.92	0.67	0.60	0.75	1.83	0.70	0.96	0.56
2,000	200	0.42	0.73	0.27	0.81	1.04	0.80	0.39	0.74
2,000	300	0.33	0.87	0.22	0.72	0.72	0.78	0.25	0.86
5,000	100	0.72	0.76	0.48	0.70	1.87	0.76	0.81	0.61
5,000	200	0.38	0.92	0.25	0.78	1.45	0.73	0.44	0.71
5,000	300	0.25	0.83	0.20	0.86	0.66	0.81	0.21	0.86
10,000	100	0.84	0.81	0.55	0.61	1.61	0.71	0.93	0.66
10,000	200	0.45	0.74	0.33	0.85	0.75	0.79	0.31	0.80
10,000	300	0.25	0.90	0.18	0.86	0.66	0.69	0.32	0.80
50,000	1,000	0.06	0.91	0.08	0.87	0.17	0.84	0.10	0.87
50,000	2,000	0.04	0.92	0.03	0.88	0.11	0.92	0.05	0.93
50,000	3,000	0.03	0.83	0.02	0.68	0.06	0.96	0.04	0.94
50,000	4,000	0.03	0.87	0.01	0.81	0.04	0.95	0.02	0.88
50,000	5,000	0.02	0.81	0.02	0.73	0.03	0.85	0.01	0.85
100,000	1,000	0.10	0.95	0.07	0.96	0.23	0.89	0.06	0.92
100,000	2,000	0.05	0.93	0.03	0.68	0.10	0.90	0.03	0.93
100,000	3,000	0.03	0.90	0.01	0.89	0.06	0.95	0.02	0.93
100,000	4,000	0.02	0.95	0.01	0.76	0.05	0.97	0.02	0.83
100,000	5,000	0.02	0.94	0.01	0.71	0.03	0.95	0.02	0.85

Some Remarks

- The proposed sampling algorithm to select samples with fixed size for the estimation domains can be applied both with equal and with unequal probabilities sampling.
- The method may be easily extended to every methods that use auxiliary information for the sampling design.
- The simulation studies show a big improvement in the efficiency of the proposed algorithm.
- The simulation analysis shows good performances in the quality of balancing, which is uniformly the same for all the considered situations, and in the respect of the known inclusion probability, where there is no fundamental difference between the two algorithms.

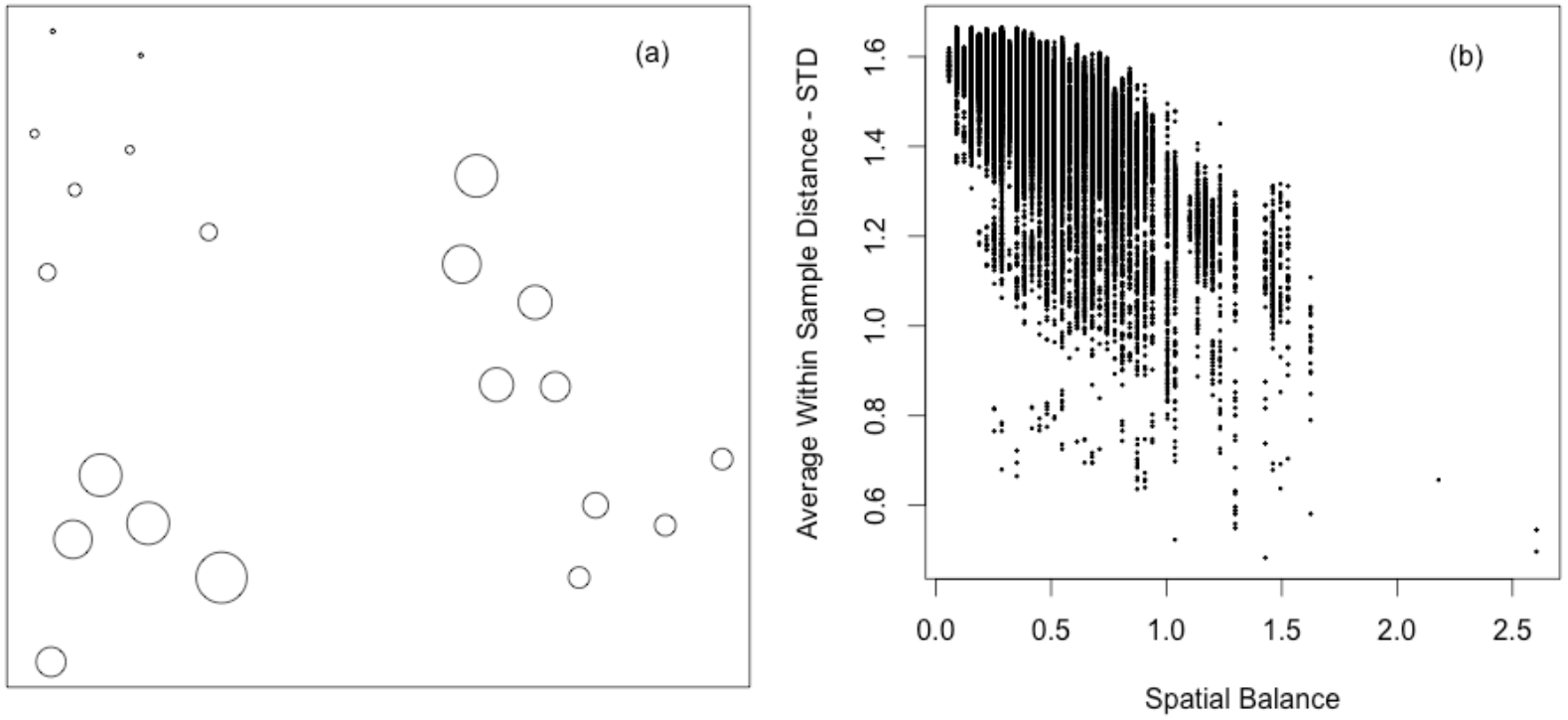
SBS Proportional to the Within Sample Distance

We propose a sampling algorithm that has the aim to select spatially balanced sampled. The design will assign higher probabilities to samples with higher variance and, thus, with higher distance.

Such a design p(S) can be obtained by setting each $p(s)=M(D_s)/\Sigma_s M(D_s)$ proportional to some synthetic index $M(D_s)$ of the matrix D_s, observed within each possible sample.

Note that the most common sample selection algorithms (for a review, see Tillé 2006) usually do not try to find a suitable choice for the probability p(S) of the sampling design, but its respect is at the most verified only *a posteriori*.

SBS Proportional to the Within Sample Distance



(a) Spatial distribution of a population of size N=21, the radius of each circle is proportional to the target variable y. (b) Scatterplot of the spatial balance index and of the average of the standardized distance matrix d_s within any possible sample of size n=6.

SBS Proportional to the Within Sample Distance

Gibbs-sampling was suggested as an efficient algorithm to draw a fixed size sample from a multivariate Bernoulli design (Traat et al., 2004). The proposed algorithm can be summarized as follows. The procedure starts at iteration t=0, with an initial point s(0), randomly selected from according to a SRS with constant inclusion probabilities. In a generic iteration t the elements of s(t) are updated in the following steps:

1. select at random two units included and not included in the sample in the previous iteration, say *i* and *j*. Respectively among the units within the sample, and another among the units outside the sample;
2. denote with ${}^{*}s^{(t)}$ the sample where the units in the position i and j exchange their status. Randomly decide whether or not to adopt this sample:

$$s^{(t+1)} = \begin{cases} {}^{*}s^{(t)} & \text{with probability } p = \min\left\{1, \left(\frac{M\left(D_{{}^{*}s^{(t+1)}}\right)}{M\left(D_{s^{(t+1)}}\right)}\right)^{\beta}\right\} \\ s^{(t)} & \text{otherwise} \end{cases}$$

3 repeat steps (1) and (2) mq times (in our application and simulations we used m and q constantly equal respectively to N and 10).

The algorithm is very simple to be implemented and extremely flexible by simply modifying the criteria used in the first step to select the two candidate units i and j or by changing the index M(Ds) or the parameter β.

SBS Proportional to the Within Sample Distance

$$M_1(D_s) = \sum_{i;s_i=1} \sum_{j;s_j=1} d_{ij} \qquad M_{1,i}(D_s) = \sum_{j;s_j=1} d_{ij}$$

$$M_0(D_s) = \prod_{i;s_i=1} \prod_{j\neq i;s_j=1} d_{ij}$$

$$M_{-\infty}(D_s) = \min_{i;s_i=1, j\neq i;s_j=1} \{d_{ij}\}$$

The parameter β plays an important role in the design as it controls our requirements on the within distance of the selected samples setting the *p(s)* as more than proportional to the distance as we need. In addition, we also found very useful to standardize the distance matrix to fixed row totals and, for the symmetry, column totals.

However, notice that an algorithm to select samples with fixed $\pi_{i,j}$ has been proposed by Bondesson (2012).

SBS Proportional to the Within Sample Distance

Inclusion Probabilities

When we use $M_1(D_s)$:

$$\pi_i = \frac{n-2}{N-2} + 2\frac{N-n}{N-2} d_{i0} \qquad \pi_{ij} = \frac{(n-2)(n-3)}{(N-2)(N-3)} + 2\frac{(n-2)(N-n)}{(N-2)(N-3)}\left(d_{i0} + d_{0j}\right) + 2\frac{(n-2)(n-3) + (N-3)(N-2n+2)}{(N-2)(N-3)} d_{ij}$$

When we use $M_0(Ds)$ (empirical):

$$\pi_i = k_1 \left(d_{i0}\right)^{k_2} \qquad \pi_{ij} = k_3 \left(d_{i0} d_{0j}\right)^{k_4} \left(d_{ij}\right)^{k_5}$$

When we use $M_1(D_s)^{\beta}$ (empirical):

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = k_1 + k_2 d_{i0} + \varepsilon_i \qquad \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = k_3 + k_4\left(d_{i0} + d_{0j}\right) + k_5 d_{ij} + \varepsilon_{ij}$$

SBS Proportional to the Within Sample Distance

Estimation and Variance Estimation

Estimation, and specifically variance estimation, can be problematic for some sampling schemes. This is particularly the case for most sequential sampling schemes such as the SCPS scheme. Unfortunately, explicit derivations of π_k and π_{kl} for each unit and pair of units in the population can be prohibitive for most summary distance indexes.

M samples have been independently selected from the population frame by repeating the same algorithm used to select *s*. An estimator of π_k and of π_{kl} that will always be positive is (Fattorini 2006; 2009):

$$\hat{\pi}_k = \frac{F_k + 1}{M + 1}, k \in U \qquad \hat{\pi}_{kl} = \frac{F_{kl} + 1}{M + 1}, k \neq l \in U$$

SBS Proportional to the Within Sample Distance

Estimation and Variance Estimation

Stevens provided exact expressions for the π_{kl} in a particular case of GRTS. However, these expressions unfortunately prevent the proper use of variance estimators based on the HT or Yates-Grundy-Sen estimators because they tend to be unstable if some π_{kl} are very close to zero. Steven and Olsen (2003) proposed a variance estimator that approximates the variance by averaging several *contrasts* over a local neighborhood of each sample point :

$$\hat{V}_{NBH}\left(\hat{t}_{HT,y}\right) = \sum_{k\in s}\sum_{l\in N(k)} wd_{kl}\left(\frac{y_k}{\pi_k} - \sum_{t\in N(k)} wd_{kt}\frac{y_t}{\pi_t}\right)^2$$

Where $N(k)$ is a local neighborhood of unit k. The wd_{kl}s are weights that decrease as the distance between unit k and l increases, and are constrained in such a way that $\sum_k wd_{kl} = \sum_l wd_{kl} = 1$

SBS Proportional to the Within Sample Distance

Estimation and Variance Estimation

A special case of this variance estimator is when continuous auxiliary variables are used and no equal distances exist. Then, there are only two units in each local neighbourhood, it simplifies to (Grafström and Schelin, 2014):

$$\hat{V}_{NBH}\left(\hat{t}_{HT,y}\right) = \frac{1}{2}\sum_{k\in s}\left(\frac{y_k}{\pi_k} - \frac{y_{c(k)}}{\pi_{c(k)}}\right)^2$$

Where $c(k)$ is the nearest neighbour to k in the sample. A similar estimator is recommended by Wolter (2007, p. 336), as one of the best general-purpose variance estimators for systematic sampling.

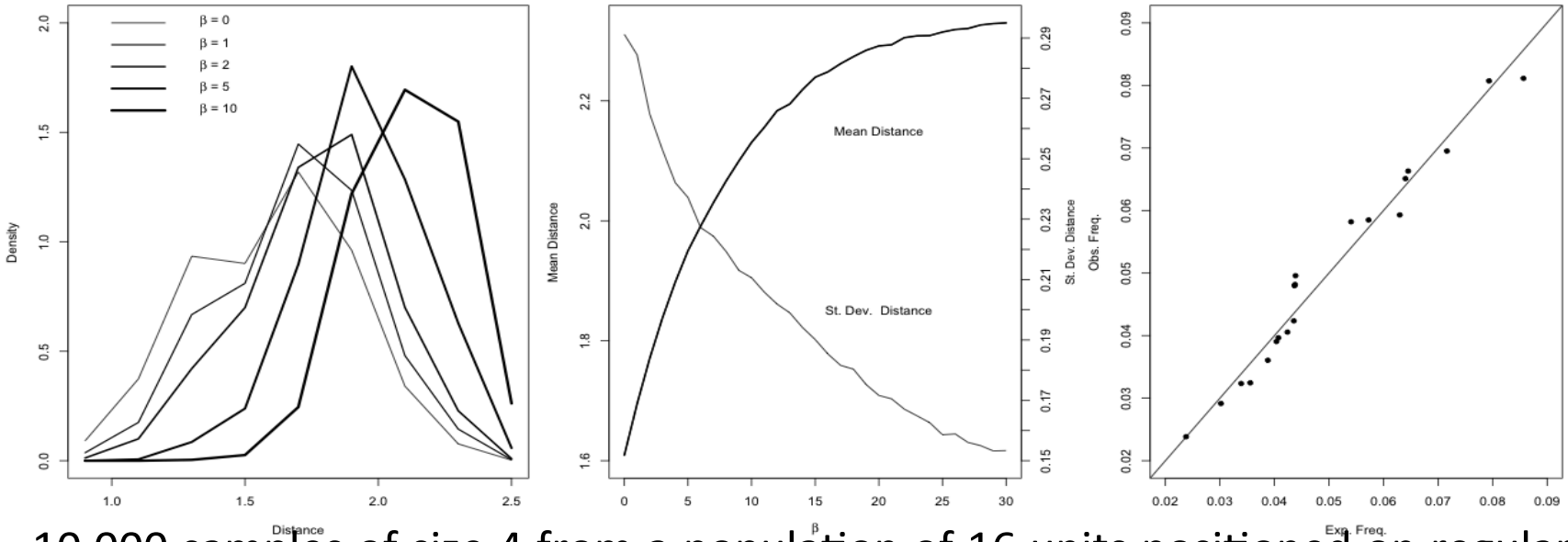
SBS Proportional to the Within Sample Distance

Estimation and Variance Estimation

- Benedetti, R. Espa, G. and Taufer, E. (2017) "Model-based variance estimation in non-measurable spatial designs", *Journal of Statistical Planning and Inference* 181, 52-61
- Fattorini L (2006). Applying the Horvitz–Thompson criterion in complex designs: a computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93: 269–278.
- Grafström, A., and Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, 41, 2, 277-290.
- Stevens DL Jr, Olsen AR (2003). Variance estimation for spatially balanced samples of environmental resources. *Environmetrics*, 14: 593–610.

SBS Proportional to the Within Sample Distance

A simulation experiment

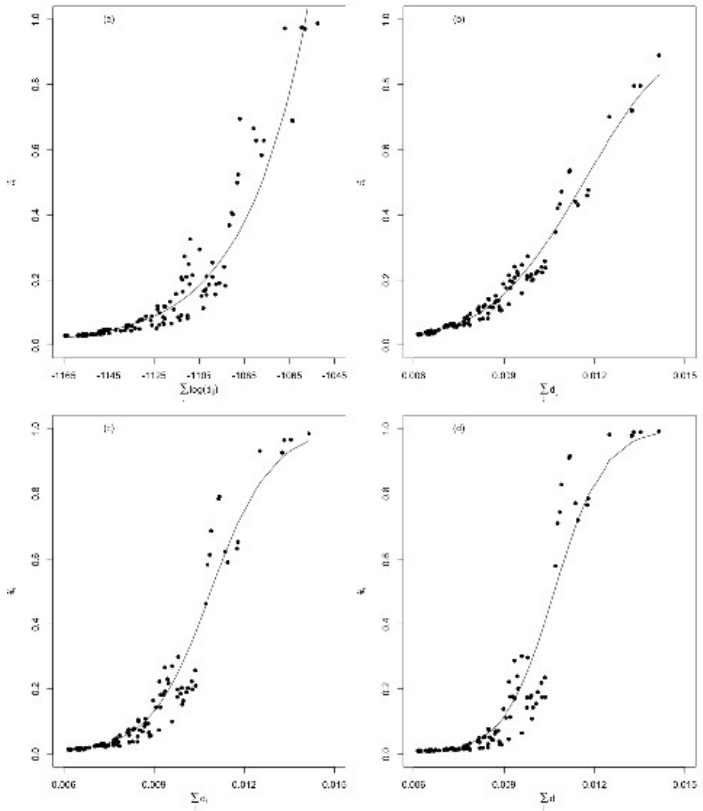


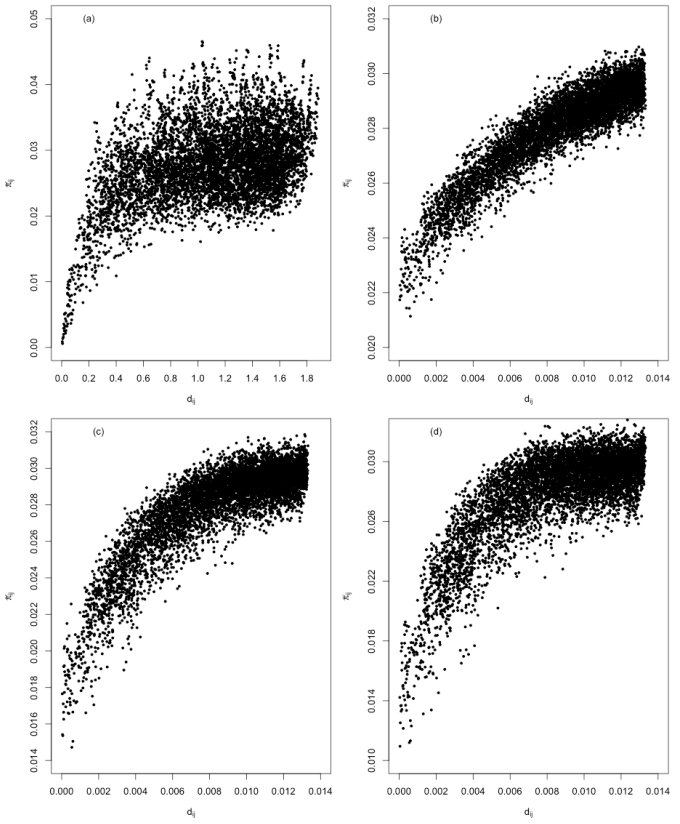
10,000 samples of size 4 from a population of 16 units positioned on regular 4x4 grid. Distribution of the samples for different values of the exponent β of the average euclidean distance (left); trend of the average and standard deviation of the within sample euclidean distance for different values of the exponent (center); expected and observed frequency for the quantiles p=0.05 for the samples proportional to the distance.

SBS Proportional to the Within Sample Distance

A simulation experiment

Scatterplot of the estimated inclusion probabilities with respect to: (a) the sum of logarithms of the distances when a PWD design is used, the sum of the distances when we used the design (b) SWD10 (c) SWD20 and (d) SWD30.





SBS Proportional to the Within Sample Distance

A simulation experiment

Regression parameters of the models for the π_i and the π_{ij} estimated on the EMAP data set for different designs.

Design	Type	k1	k2	R2
PWD	Double Log	exp(38.18)	0.03609	0.928
SWD10	Logistic	-7.404	636.111	0.974
SWD20	Logistic	-10.834	994.789	0.949
SWD30	Logistic	-12.918	1210.361	0.917
SWD40	Logistic	-13.918	1310.390	0.881
SWD50	Logistic	-14.416	1358.785	0.848

Design	Type	k3	k4	k5	R2
PWD	Double Log	exp(78.57)	0.4618	0.03514	0.892
SWD10	Logistic	-13.042	1478.208	496.946	0.974
SWD20	Logistic	-18.258	2550.393	740.002	0.933

SBS Proportional to the Within Sample Distance

A simulation experiment

Coefficient of variation of the estimated first order inclusion probabilities when we use a distance matrix standardized to constant row and column totals for the SWD or constant row and column products for the PWD.

PWD	SWD10	SWD20	SWD30	SWD40	SWD50
0.1160	0.0074	0.0122	0.0169	0.0220	0.0281

Mean and standard deviation of the index of spatial balance for different designs and for the PWD and the SWD for standardized and not standardized distance matrix.

			SWD							
D		PWD	b=10	b=20	b=30	b=40	b=50	GRTS	SCPS	LPM1	LPM2
NotSTD	μ	0.125	0.237	0.168	0.138	0.119	0.106	0.407	0.192	0.182	0.189
NotSTD	σ	0.046	0.087	0.068	0.062	0.057	0.054	0.146	0.062	0.050	0.060
STD	μ	0.189	0.287	0.259	0.243	0.232	0.223	-------	------	-------	-------
STD	σ	0.057	0.091	0.081	0.074	0.070	0.068	-------	------	-------	-------

SBS Proportional to the Within Sample Distance

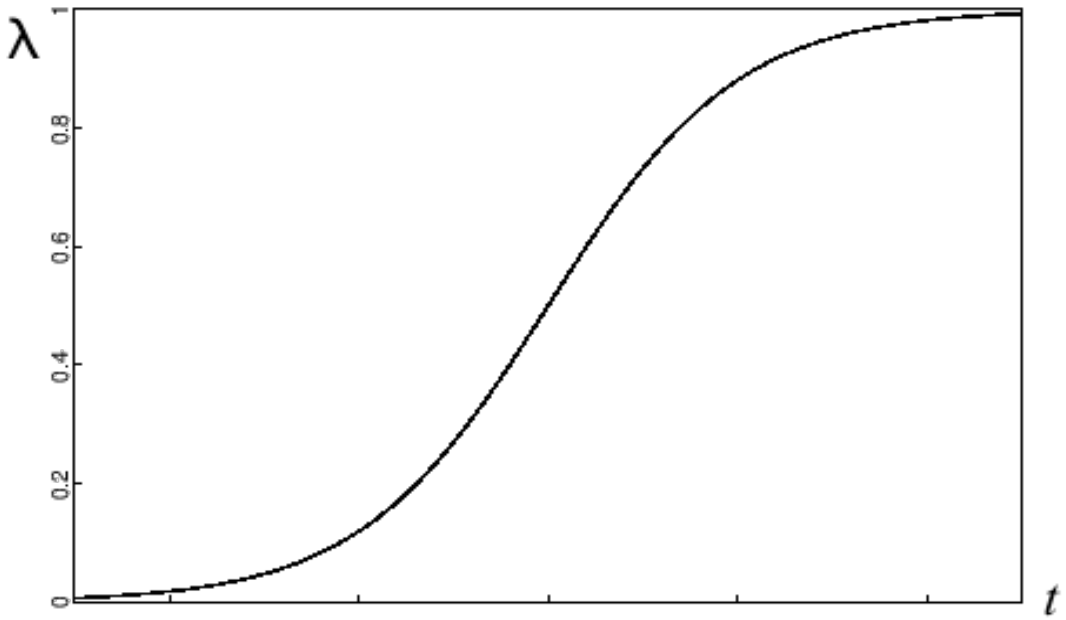
Some Feature

1. The first and second order probabilities are known when the index is the average of the within distances
2. The second order are never equal to 0 so the H-T variance estimation is feasible
3. To select a stratified sample proportional to $M(D_s)$ by simply starting with a stratified SRS with fixed sizes n_h in each stratum h and then selecting the candidate j not among all the population units but within the same stratum of the candidate i.
4. We can coordinate the sample selection between two or more different surveys, or the same survey but in different time, periods or phases of the survey. We can simply start with the previously selected sample and by restricting the selection of the candidate j within all the units which are not selected in the sample and that were not selected in other occasions and let the procedure run until the required number of units are replaced from the original sample.

Doubly SBS Proportional to the Within Sample Distance

$$s^{(t+1)} = \begin{cases} {}^{*}s^{(t)} & \text{with probability } p = \min\left\{1, \left(\frac{M\left(D_{{}^{*}s^{(t+1)}}\right)}{M\left(D_{s^{(t+1)}}\right)}\right)^{\beta} + \lambda^{(t)}\left(\sum_{k \in s} d_k x_{kj} - t_{x_j}\right)\right\} \\ s^{(t)} & \text{otherwise} \end{cases}$$

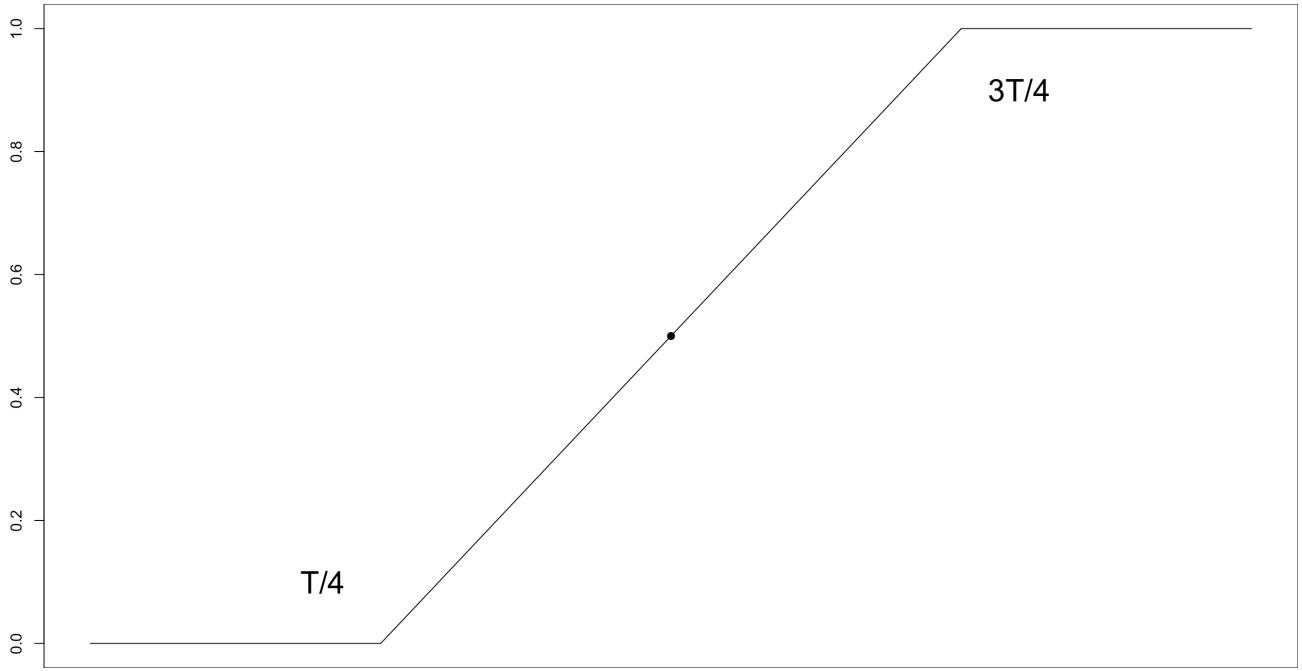
Where λ is a penalty that starts from 0 and increases with the number of iterations.



Doubly SBS Proportional to the Within Sample Distance

$$s^{(t+1)} = \begin{cases} {}^{*}s^{(t)} & \text{with probability } p = \min\left\{1, (1-\lambda^{(t)})\left(\frac{M\left(D_{{}^{*}s^{(t+1)}}\right)}{M\left(D_{s^{(t+1)}}\right)}\right)^{\beta} + \lambda^{(t)}\left(\sum_{k\in s} d_k x_{kj} - t_{x_j}\right)\right\} \\ s^{(t)} & \text{otherwise} \end{cases}$$

Where λ is a tuning parameter of a convex combination of the obj. func. and of the a penalty that starts from 0 and increases to 1 with the number of iterations.



A Very Simple and Quick Draw-By-Draw Heuristic

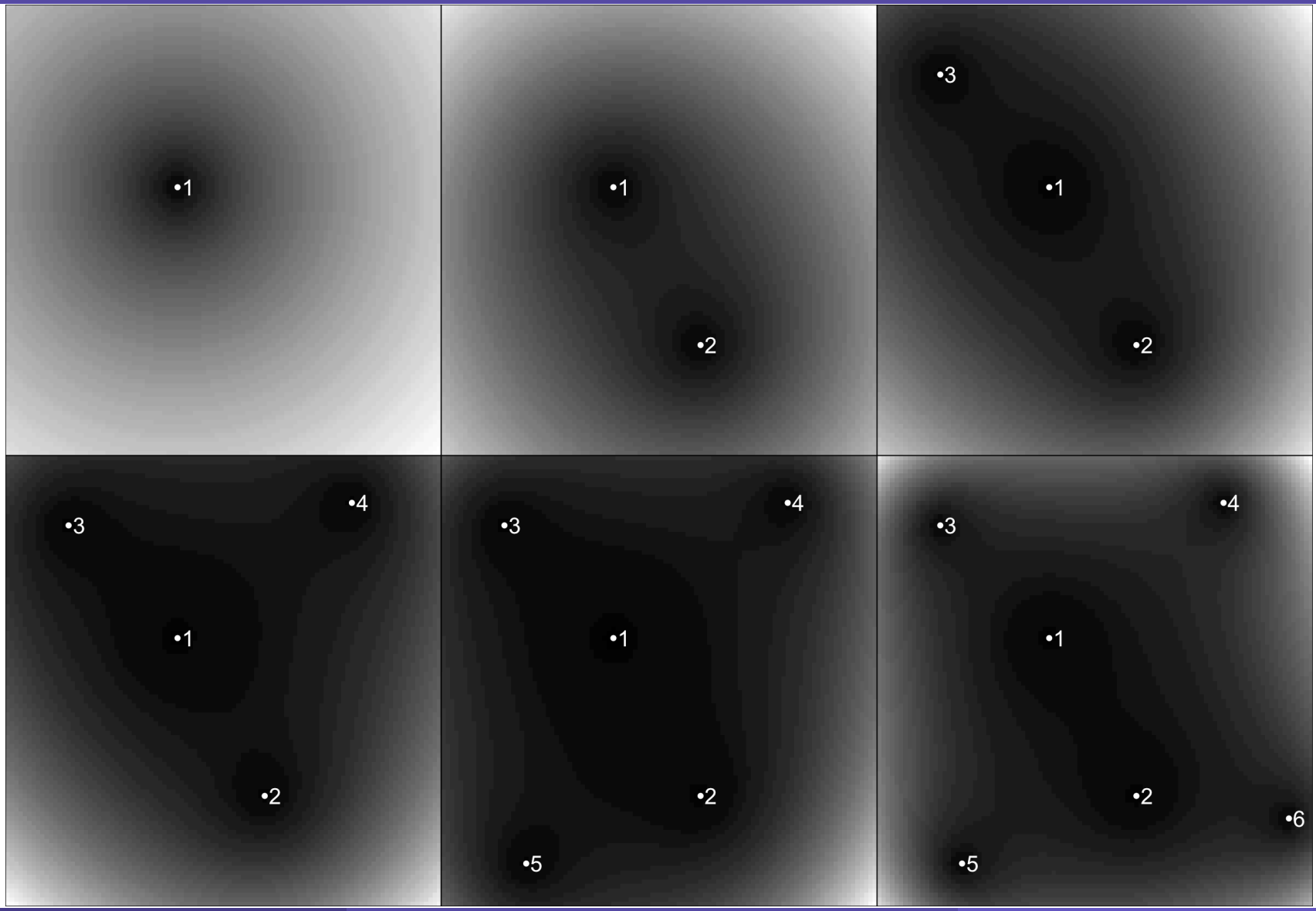
The algorithm starts by randomly selecting a unit *k*. Then, at every step *t*<*n*, the algorithm updates the selection probabilities of any other unit (*l*) of the population according to the rule

$$\pi_l^{(t)} = \frac{\pi_l^{(t-1)} \bar{d}_{kl}}{\sum_{l=1..N} \pi_l^{(t-1)} \bar{d}_{kl}}$$

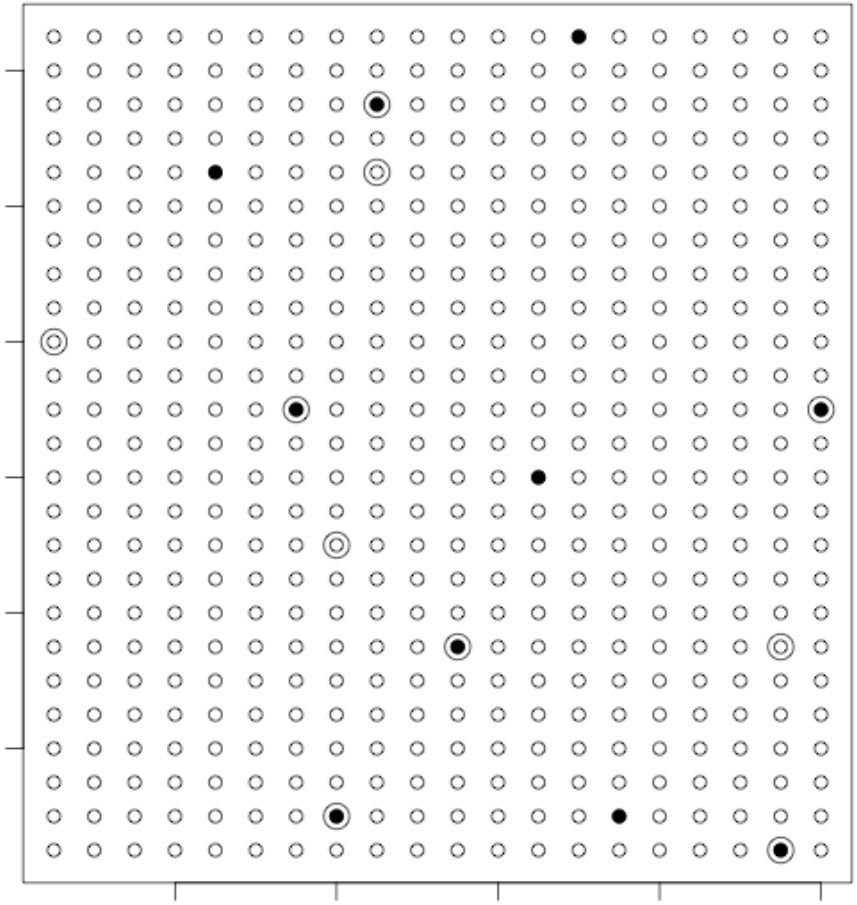
Where d_{kl} is a measure of distance between unit *k* and *l* and is standardized by row and columns products (log of row and column sums).

This algorithm require only n steps so it is very quick. It show empirically to have the same π_l of the PWD (thus to fix constant π_l the distance matrix should be standardized).

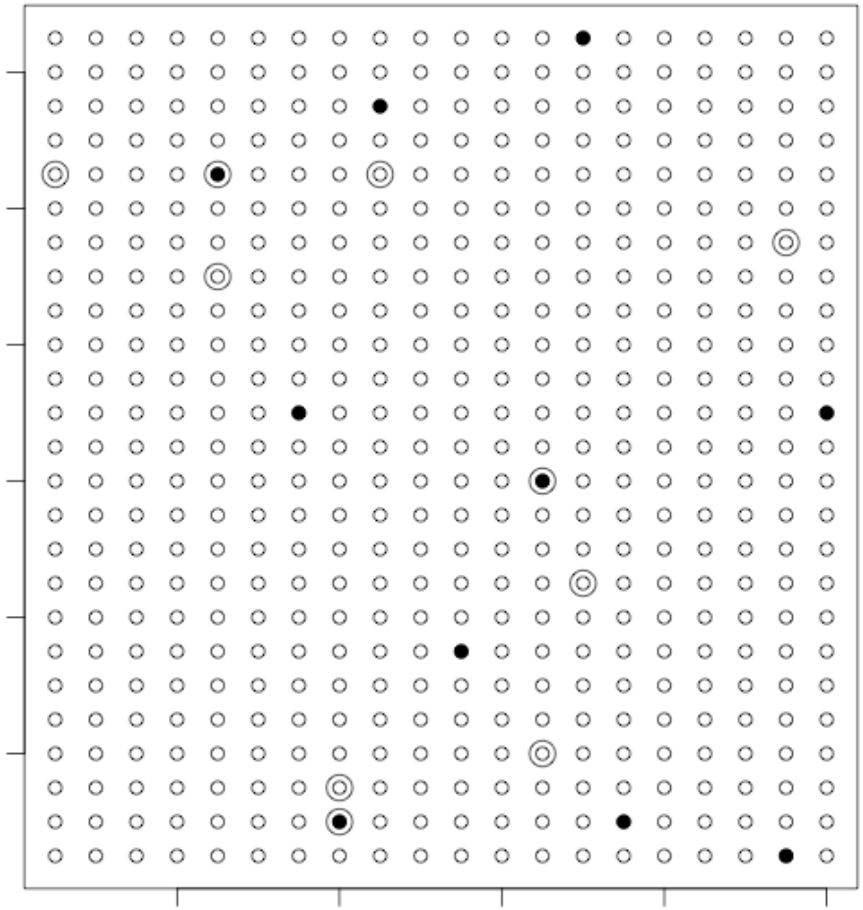
A Very Simple and Quick Draw-By-Draw Heuristic



Agenda: Panel (Rotation) of Spatially Balanced Samples (coordination of Sp. Bal. Samples)

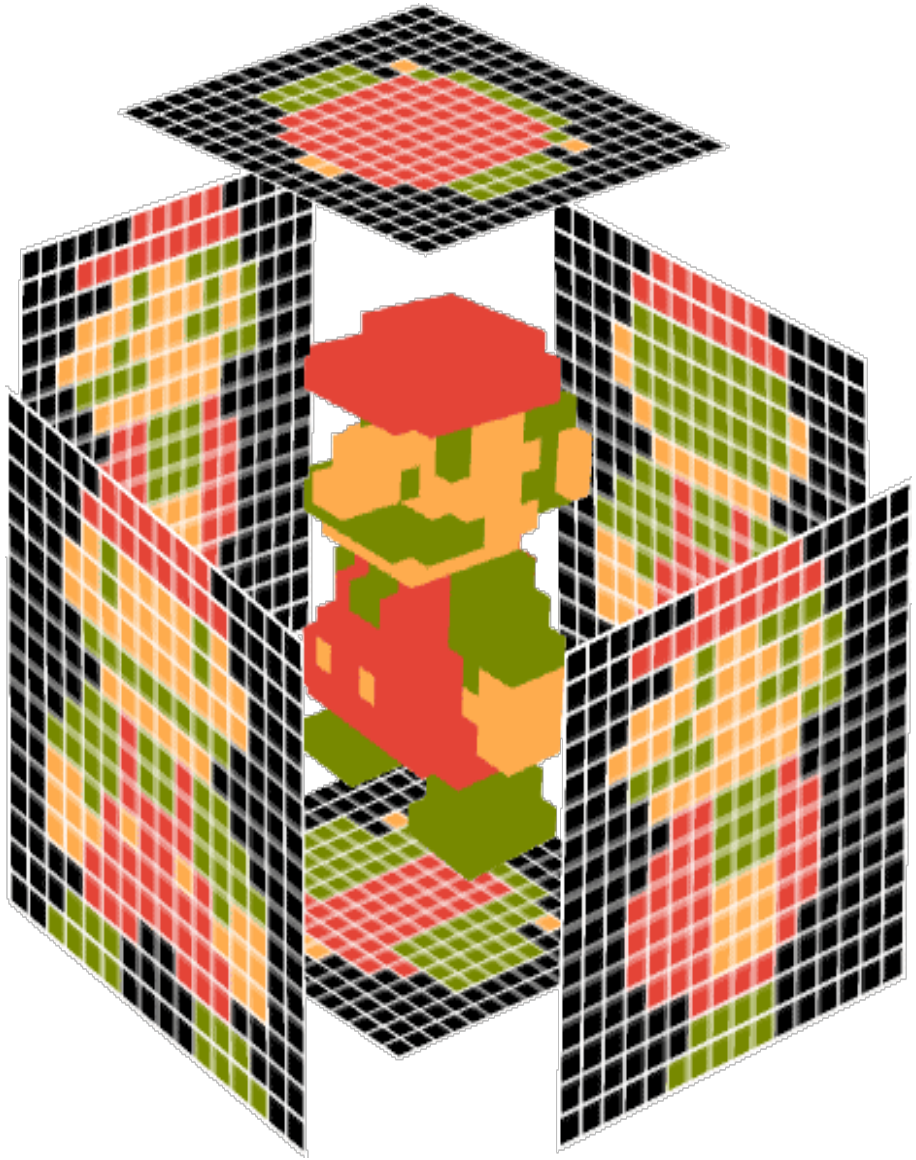


n=10, %rotation = 40 %, bal. Ind. 1 = 0.116, bal. Ind. 2 = 0.082

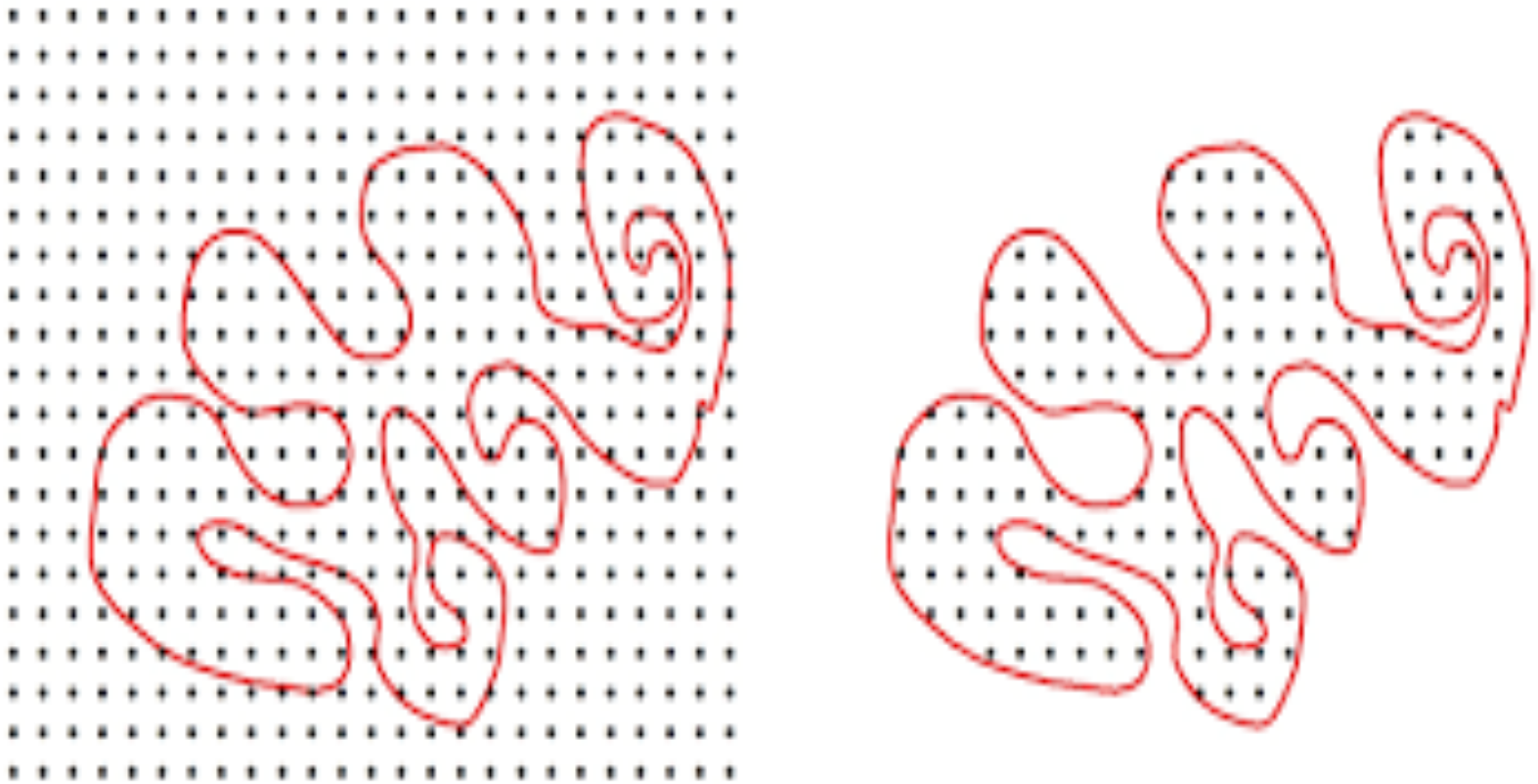


n=10, %rotation = 70 %, bal. Ind. 1 = 0.116, bal. Ind. 2 = 0. 139

Agenda: Camera position for 3d object reconstruction (with R.L. Chambers)

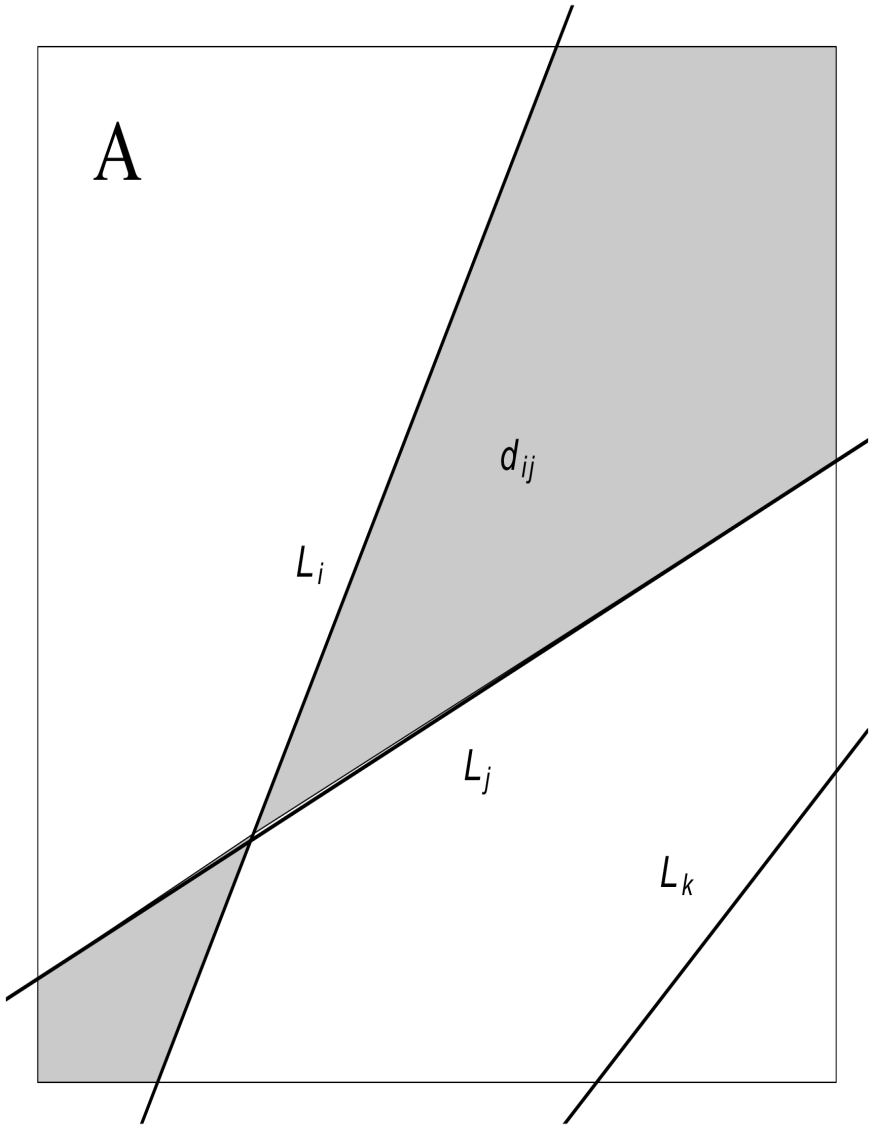


Agenda: Sampling from continuous spatial populations (with R.L. Chambers)



Design: 2 Stage – 1° systematic or TSS - 2° any design for finite populations (point transect sampling)

Agenda: Spatially Balanced Samples of Oblique Line Transects (with R.L. Chambers)



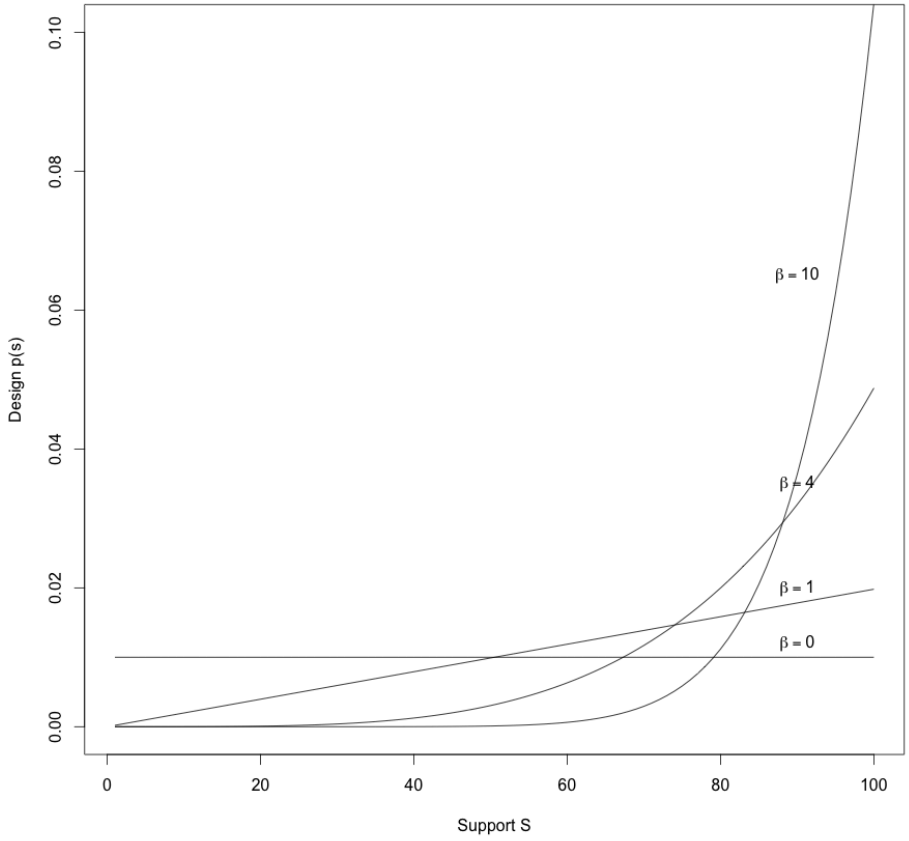
$$\pi_i = \frac{L_i}{\sum L_i} n$$

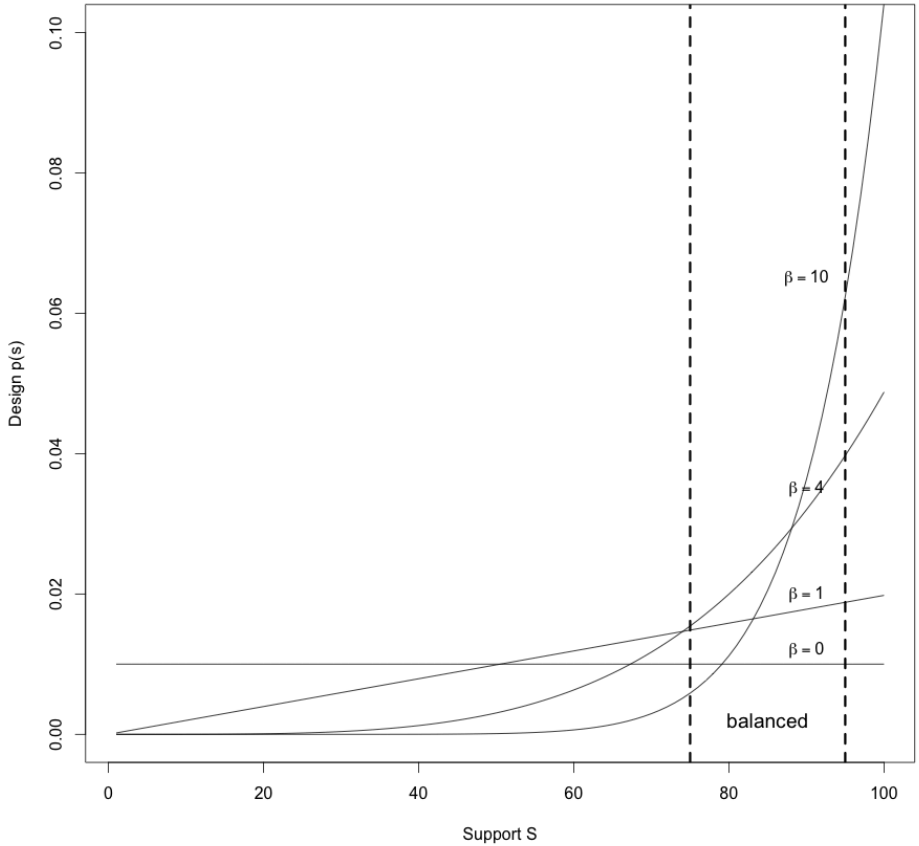
$$\pi_{ij} = f(d_{ij})$$

where f () is a monotone increasing function

$$P(S) \propto \mathrm{M}(d_{ij})$$

We have some doubts about what we are doing





we can reduce the randomization either by increasing the p(s) for samples with high variance or restricting the support (balancing), is it a reasonable practice ? How far we can go on this line ? The limit is an optimal (purposive) sample (i.e. back to 20 years ago)

Some references

- Benedetti, R, Piersimoni, F and Postiglione, P. (2015). Sampling spatial units for agricultural surveys. Advances in Spatial Science Series. Springer.
- Benedetti, R, Piersimoni, F and Postiglione, P. (2017). Spatially Balanced Sampling: A Review and A Reappraisal. International Statistical Review, 85, 3, 439–454.
- Benedetti, R, Piersimoni, F (2017). A spatially balanced design with probability function proportional to the within sample distance, Biometrical Journal, 59, 5, 1067-1084
- Benedetti, R, Piersimoni, F (2018). Fast Selection of Spatially Balanced Samples, Arxiv, https://arxiv.org/abs/1710.09116
- Bondesson, L. (2012). On sampling with prescribed second-order inclusion probabilities. *Scandinavian Journal of Statistics* **39**, 813–829.
- Traat, I., Bondesson, L., and Meister, K. (2004). Sampling design and sample selection through distribution theory. Journal of Statistical Planning and Inference, 123, 395 -413.

Thank you for your attention!