Spatial sampling and entropy

Daniela Cocchi, Linda Altieri
daniela.cocchi@unibo.it

Department of Statistical Sciences,
University of Bologna, Italy

.

Ephastat

ITACOSM 2019

FLORENCE 5-7 June 2019 | 6th ITAlian COnference on Survey Methodology

Background

Viewpoints on environmental sampling

- reference to the population
- spatial sampling (at least two dimensions added)

Role of auxiliary information in survey sampling

- modify the inclusion probabilites
- estimation via the inclusion probabilities (and/or further auxiliary variables)

Data spatial correlation may influence the variance of estimated standard error

Link between sampling and entropy

Why search for sampling plans with high entropy?

The **entropy** of a sampling plan is seen as a measure of **RANDOMNESS**

Conclusion:
(Conditional) Poisson sampling enjoys the maximum entropy property

But careful:

- sampling entropy
- spatial entropy

A good sampling plan for spatially correlated data ought to

- produce similar estimates for different spatial configurations of the variable under study,
- i.e. the estimate should not be affected by the underlying spatial structure.

Inclusion probabilities

They are the basis of the design based inference, where HT-type estimators are proposed.

Role of inclusion probabilities:
They weigh the values of the variables under study in the sampled units

Methods have been developed for sequentially modifying the (population) inclusion probabilites by means of FURTHER pre-sampling **weights**.

Such further weights consider DISTANCES.

The aim is trying to obtain sampling plans that are well spread.

Parallel and independent work

Modifications of classical entropy measures that keep the POPULATION spatial structure into account

For a given variable X, different spatial structures deliver the same entropy value, unless ...

Briefly, some modifications are introduced

Aim of this work

To explore the consequences of using some components of the spatial entropy measures as weights

There are differences according to the consideration of

- "labels" spread (sampling entropy)
- "values of the variable" spread (spatial entropy)

Spatial entropy

$$H(X) = E[I(p_X)] = \sum_{i=1}^{I} p(x_i) \log\left(\frac{1}{p(x_i)}\right). \tag{1}$$

Consider the two variables Z and W:

Z is the variable corresponding to unordered pairs of realizations of X over the observation area; it may present $R = \binom{I+1}{2}$ categories.

W classifies the Euclidean distances within the observation window according to a set of distance classes w_m, with $m = 1, \dots, M$.

A set of distance breaks $d_0, \dots, d_M$ is fixed, with $d_0 = 0$ and d_M being the maximum possible distance inside the window; then, each class is $w_m =]d_{m-1}, d_m]$.

Use of the bivariate properties of entropy

The following well known relationship of entropy theory holds

$$H(Z) = MI(Z,W) + H(Z)_W. \tag{2}$$

Rename the first term as Spatial Mutual Information

$$SMI(Z,W) = \sum_{m=1}^{M} p(w_m) SPI(Z|w_m) \tag{3}$$

and its weighted components as Spatial Partial Information

$$SPI(Z|w_m) = \sum_{r=1}^{R} p(z_r|w_m) \log\left(\frac{p(z_r|w_m)}{p(z_r)}\right). \tag{4}$$

Sampling entropy

$$H(S) = \sum_{j=1}^{J} p(s_j) \log \left(\frac{1}{p(s_j)} \right)$$

Spatially correlated Poisson Sampling (SCPS)

Idea: the variable is correlated so sampled units should be well spread

Sequential method: units are visited according to spatial order

- initial inclusion probabilities sum to the expected sample size n
- an indicator function I_k is defined for each population unit, taking value 1 if the unit is sampled
- the sampling outcome is decided for unit 1

How does the procedure work

- the remaining units' inclusion probabilities are updated accordingly, following the rule

$$\pi_l^{(k)} = \pi_l^{(k-1)} - (I_k - \pi_k^{(k-1)}) w_k^{(l)}$$

- repeat for unit 2, 3, ..., N
- at step N, the final vector of inclusion probabilities is $(\pi_1^{(N)}, \ldots, \pi_N^{(N)})' = (I_1, \ldots, I_N)'$

Weights decide how the sampling of a unit is affected by the previous ones; they depend on a distance function $d(k, l)$ and give negative correlation to close units

Spatial entropy in SCPS

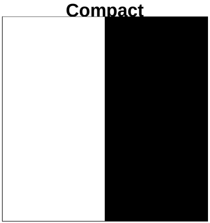
Innovation: new weighting system for SCPS, which exploits the theory of spatial entropy.

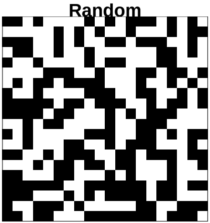
Weights are built in order to take the spatial correlation of the variable X into account (novelty!), via SMI. The stronger SMI, the smaller our interest in sampling neighbouring units.

SMI becomes the auxiliary variable for building a well founded weighting system for spatially balanced sampling.

Simulation study - data

- binary variable X
- $N = 400$ realizations: 200 $x_0 = 0$ and 200 $x_1 = 1$
- realizations arranged according to two spatial configurations





Simulation - weights

The two configurations produce different SPI values

Distance classes: $w_1 = [0, 1]$, $w_2 =]1, 2]$, $w_3 =]2, 5]$, $w_4 =]5, 20\sqrt{2}]$

Spatial mutual information - partial terms

	$[0, 1]$	$]1, 2]$	$]2, 5]$	$]5, 20\sqrt{2}]$
Compact	0.574	0.485	0.289	0.010
Random	$<$0.001	0.001	$<$0.001	$<$0.001

- a unit k is visited for sampling
- SPI values are rescaled to sum to 1 and assigned to units $l = k + 1, \dots, N$ according to their distance from k
- they become the weights for updating the remaining inclusion probabilities

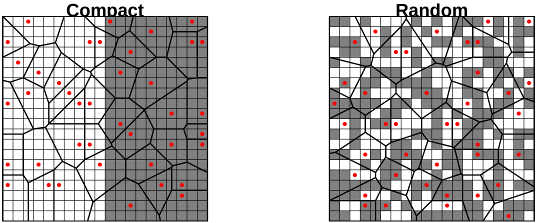
RANDOM CONFIGURATION: probabilities remain almost constant
COMPACT CONFIGURATION: probabilities for close units decrease

Simulation - sampling

100 samples of size $n = 40$ are drawn from each dataset.

The initial inclusion probabilities are constant: $\pi_k = n/N$ for all units k.

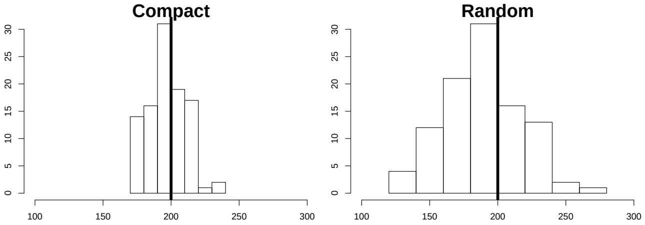
Example (one simulated sample):



The Voronoi tessellation is also plotted

Results - HT estimates

The thick vertical line marks the true total $Y = 200$



MSE:

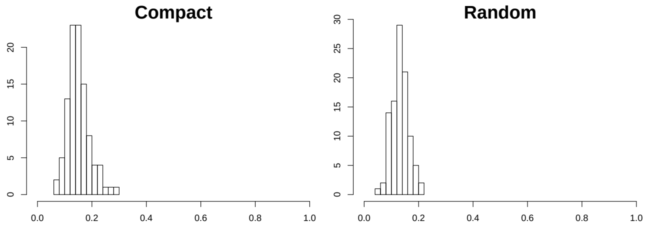
- compact configuration: $MSE_c = 215$
- random configuration: $MSE_r = 928$

Results - variance of Voronoi polygons

It is a measure of how well spread samples are (the smaller the better)

$$v^2(v) = \frac{1}{n}\sum_{h=1}^{n}(v_h - 1)^2$$

v_h: sum of the inclusion probabilities of all units in the hth Voronoi polygon
$E(v_h) = 1$



Future work

- We look for the communication between seemingly separated worlds:
  spatial entropy
  and
  sampling entropy
- Deepen the means of considering the important contribution contained in partial spatial information
- Explore ways for estimating the probabilites entering spatial entropy

Thank you!