Spatial Statistics
August 2020, Volume 38, Pages 100439 (13p.)
https://doi.org/10.1016/j.spasta.2020.100439
https://archimer.ifremer.fr/doc/00619/73140/

Sequential process to choose efficient sampling design based on partial prior information data and simulations

Kermorvant Claire [1, *], Coube Sébastien [1], D'amico Frank [1], Bru Noëlle [1], Caill-Milly Nathalie [2]

[1] CNRS/Univ Pau & Pays Adour, Laboratoire de Mathématiques et de leurs Applications de Pau-Fédération MIRA, UMR 5142, 64600 Anglet, France
[2] Ifremer, Laboratoire Environnement Ressources d' Arcachon, France

* Corresponding author : Claire Kermorvant, email address : claire.kermorvant@univ-pau.fr

Abstract :

Issues on sampling procedure definition led numerous study results to be biased and object of controversy. Choosing relevant sampling design and number of samples is a difficult task when wanted to set up or optimize a survey. The survey design choice is very important to avoid bias and increase the survey cost-efficiency. It can have a strong effect on the sample size needed to achieve some targeted accuracy on results. And on the final cost of the procedure.

The sequential process we expose here melt design based and model based sampling theories. Its objectives is helping practitioners defining a sampling design and a number of samples for their survey when inference to the whole population is wanted. The main idea is to mathematically reconstruct the distribution of the surveyed population. Then assess and compare cost-effectiveness of various sampling designs on this population. This process allows setting predetermined level(s) of accuracy to be reached in the targeted estimates and to take into account previous relevant data. Results are an optimal sampling design and an associated optimal sample size for a desired accuracy in the results. This accuracy is so achieved without excess sampling. Strength of this process is that it is based on simulations. This allows trying a high number of combinations between sampling design, sample size and desired level of accuracy. Sampling design performances can thus be compared.

The user can decide which combination is the best for his survey and apply it for real. We discuss how to use available data to improve the survey, from the case were several historical data are provided to the case where no data are available.

Keywords : accuracy, cost-effectiveness, survey, simulations, survey optimization, virtual ecology

1. INTRODUCTION

Monitoring programs are tools used in environmental science in three main tasks: to detect a change into a system, to measure success or failure of management actions and to identify effects of perturbations or disturbances (Legg and Nagy, 2006). Likens and Lindenmayer (2018) reviewed the term "monitoring" in the ecological literature between 1985 and 2016 and more than 131 000 articles and numerous books were returned. Monitoring information is essential answering most ecological and environmental questions (Albert et al., 2010). For example they can be the basis for restoration programs or for endangered species conservation.

Sampling is very common because exhaustive information cannot be collected for almost all the cases. But its theory is complex and several environmental scientists are not trained to it. Thus, programs suffer from lack of details of problematic definition, hypothesis formulation, adapted sampling design and so data quality (Legg and Nagy, 2006). Poor method has numerous undesirable effects that can lead to the failure of a monitoring program (Legg and Nagy, 2006). Issues with poor designs used in ecological studies often have led to significant controversy (Hayward et al., 2015). It also means that it becomes difficult to evaluate management actions and results are not very useful for decision making (Vos et al., 2000). For example it has not been possible to evaluate the effectiveness of US 15 billion projects of rivers restoration all around US (37 099 projects) because of poor experimental design and lack of rigorous monitoring (less than 10 percent of them indicated a form of assessment or monitoring of project efficiency) (Bernhardt et al., 2005). Roberts (1991) and Nichols and Williams (2006) deplore too many monitorings are "planned backward on the collect now (data), think-later (of a useful question) principle". A forum (Hayward et al., 2015) wrote

after conflicting results were published in high-quality scientific journal. It emphases robust methods and appropriate experimental designs must be developed and used by practitioners, avoiding controversy in studies results.

The choice of a sample size and a sampling design is a very important step in the establishment of a survey. Representativeness is brought by the random property of the sampling design (Macdonald, 2009; Sica, 2006), precision by collecting enough data through a substantial sample size (Lohr, 2009). A substantial sample size will increase precision on population estimation but may also increase the survey cost. This is particularly true in ecology where sampling on field necessitates human and/or expert resources, sometimes expensive gears (boats, trucks...) differing than, for example, in some cases in sociology where survey can just be a sample on the web. In the great barrier reef monitoring (Kang et al., 2016), divers must have the necessary skills and qualifications to do the monitoring and statisticians must curate the data. The practitioner has to found a trade-off (Stehman and Overton, 1996) between a sufficient amount of sample, to achieve a precise estimation, and a price that will be reasonable for financers. Before constructing any survey procedure, clear objectives about total sample size and estimator quality have to be fixed. Priority can be given to maximize estimator quality (it's accuracy) or minimize total sample size (Guillera-Arroita et al., 2010). No survey designs will be good for all purposes (Kenkel et al., 1990).

In ecology, the studied population is almost always a spatial population because species always display spatial distribution. With new developments in statistics and geostatistics, a large amount of probabilistic sampling designs was developed last decades (McDonald, 2014). Now, a significant number of them are available and the issue is that it may be very tricky to determine which one is better to use for each surveys. This is especially true because environmental scientists are not prepared to this. Some environmental studies use non-probabilistic sampling designs to draw samples from spatially distributed populations. Unfortunately, data gathered with a non-probabilistic design can be biased (Albert et al., 2010; Levy and Lemeshow, 2013) because the random component is not taken into consideration. No design-based inference and statistical studies could be done from data collected with a non-probabilistic sampling design. Probabilistic sampling designs, displaying a random property, must be used for design-based sampling. Kermorvant et al. (2019b) published a review of probabilistic spatially bal-

anced sampling designs and a tutorial to use them on R software. Simple random sampling (SRS) design is one of the most commonly used survey design in ecology, due to its ease of use and its flexibility. Systematic sampling (SS) design sets randomly the first sample on studied population and then distributes the other equidistantly from each other. Spatially balanced sampling designs are probabilistic designs that spread the sampling effort evenly over the resource. The most popular of them is generalized random tessellation stratified (GRTS) (Stevens and Olsen, 2004). It has many desirable features including a spatially balanced sample, design-based estimators and the ability to select spatially balanced oversamples. Spatially balanced sampling designs are particularly useful for environmental samplings because they produce good sample coverage over the resource, have precise design-based estimators and they can potentially reduce the sampling cost. GRTS is not the only spatially balanced sampling. BAS - Balanced acceptance sampling (Robertson et al., 2013, 2017), HIP - Halton iterative partitioning (Robertson et al., 2018), SCPS - spatially correlated Poisson sampling (Grafström et al., 2012) and LPM - Local Pivotal Method (LPM) (Grafström, 2012) are also spatially balanced sampling designs. Financial constraints are the main reason given for using qualitative (poor) methods (Legg and Nagy, 2006), which do not guarantee survey success. Cost-efficiency of surveys is under scrutiny.

We are not the first team that is interested into cost-efficiency optimisation of monitoring programs. Field et al. (2005) also use simulation study to show that monitoring in environment can be optimized by using power analysis. Their method permits to select a sample size and a number of visits that maximize statistical power within fixed budget constraints. Power analysis are, for now, able to be calculated only for simple random sampling design and so are not very relevant when using a more advanced sampling design (i.e spatially balanced sampling design). Spatially balanced sampling designs are proved to be more efficient than simple random sampling (Robertson et al., 2013; Brown et al., 2015; Grafström and Matei, 2018; Kermorvant et al., 2019b) and so need fewer samples to detect same level of change between two survey seasons. We believe that overestimation of sample size is probable in such cases. As long as power analyses are not available for these advanced sampling designs, it is very tricky to use them when the purpose is to compare their efficiency. Another team worked on an optimisation framework for monitoring biological invasions under global change (Vicente

et al., 2016). Liberts (2013) explains how to obtain high quality estimates of population from an artificial population data with low cost. Moore and McCarthy (2016) integrate imperfect detection and different times spent in surveying by selected sites. Another framework, aiming at optimizing monitoring networks of multi species monitoring programs (Carvalho et al., 2016), permits to include a predetermined number of sites while reducing the total survey costs. Rudders (2011) shows that we can evaluate sampling designs performance fixing the level of sampling intensity (three different levels to estimate sea scallops abundance in this case). The methodology is valid when the practitioner wants to know how much accuracy will be achieved with the number of samples he can find.

The major challenge of this paper is to provide a sequential method permitting to choose optimal sampling design and optimal sample size for a survey. First of all, the process takes into account prior data. Prior knowledge of the studied area and the population can dramatically reduce the uncertainty in the sampling estimate (Wang et al., 2012). Without this knowledge of the population distribution, optimization of the survey can be very tricky. This knowledge is very relevant to provide initial idea of quantification and spatial (and/or temporal) limitations of the survey area (and/or duration). Secondly, these data will be the basis to compare several sampling designs to choose the most efficient on this population. Rajabi and Ataie-Ashtiani (2014) define performance (= efficiency) as the capacity of a design strategy to require fewer samples to reach a certain level of accuracy. In the comparison of various sampling designs, efficiency can be viewed as a measure of quality of these sampling designs (Brown, 2003).

Important issues with this challenge are to keep an acceptable accuracy on estimation results and emancipate from sampling fluctuations. Following this, we choose to define the "optimal sampling design" as the more efficient design between those assessed. The "optimal sample size" will be the corresponding sample size. The developed process compares sampling designs efficiency and determines the most optimal of them. For each assessed sampling design, the method simulates a large number of samples (e.g > 1000) of size and calculates reached accuracy on targeted estimate for this. If this accuracy is not smaller, or at least equal, than the accuracy fixed by the user, a greater is assessed. Once the fixed accuracy is reached, the associated sample size is selected as optimal sample size for this sampling design.

As sampling designs perform differently following the statistical population, optimal sample sizes are different following the used sampling design. The sampling design that has the smaller optimal sample size is chosen as the optimal sampling design and must be applied on field with its associated optimal sample size.

2. THEORETICAL FRAMEWORK

Let us denote Ω a finite statistical population composed N of elementary units ω. On a purely spatial research problem, statistical population would be the area of interest and an elementary unit would be a point, a line or a polygon. On a temporal research problem, statistical population would be a time lapse and an elementary unit would be a punctual date or a time interval. Finally, on a spatio-temporal problem, statistical population and elementary unit would be a combination of both spatial and temporal features.

Let us consider Y a numerical statistical variable of interest unknown on all statistical units ω of a spatial statistical population. We will note $Y = \omega_1; \omega_2; ...; \omega_n$ all the possible values of Y on specific statistical units. We want to estimate a particular parameter of this variable. In this paper, we will focus on the total parameter. For example, if we work on total of abundance, Y would be the number of individuals. Let's note $T(Y)$ on Ω the total of variable Y on the statistical population. In sampling theory, we need to collect some information about on a sample of statistical units. We will note $y_{\omega_1}; y_{\omega_2}; ...; y_{\omega_n}$ the values of Y sampled in one sample of size S. To estimate the interest parameter on from a sample, we should then construct an estimator or choose between existing ones.

In this paper, the Horvitz-Thomson's estimator (Horvitz and Thompson, 1952) is chosen for our total parameter estimation example because it is the best linear unbiased estimator (BLUE) (Tillé, 2011). Horvitz-Thomson formula gives a linear estimator of total from samples values, without bias and valid for all probabilistic sampling designs. Horvitz-Thompson estimator of total only depends on number of statistical units in the population, number of samples and values taken by these samples. Hence, the chosen statistical population may have an effect on final estimates and so must be well defined at the beginning of the study. Values taken by samples are non-manageable and so only sample size can affect estimation.

3. SEQUENTIAL PROCESS

We are interested on finding the optimal sampling design and associated sample size for a survey with prior data.

3.1. Step 1 - data available and reconstruct $\hat{Y}$ on Ω

In the following we assume that Ω is a spatial domain. The entire distribution of Y variable on Ω is rarely known. For example, it is difficult to know the abundance of one species within its all living area. But, do we have some previous data on this space or not? And, if yes, what can we do with such data? These questions are the beginning of our process. We will now detail the two cases: data are available or they are not.

If previous data of Y were drawn with a probabilistic sampling design (design-based method), random ensures data independence. In this case estimates can be derivate directly from samples values, without assumption on Y distribution (Petitgas, 2001). When data are not collected with a random process, a model of Y spatial structure need to be inferred. The estimate is so model-based (Cochran, 1977; Petitgas, 2001).

Spatial Interpolation Methods (SIMs) are usually used to reconstruct this spatial field. Commonly, SIMs are divided into three main categories: non-geostatistical interpolation methods, geostatistical interpolation methods and combined methods. Li and Heap (2014) provide a review of 25 methods among SIMs and a decision tree to determine the most appropriate method depending on the data user hold. To demonstrate our sequential process can be used with geostatistical and non-geostatistical methods, we will provide two examples based on each these SIMs categories. For non-geostatistical interpolation methods, we will use a regression model. For geostatistical interpolation model we will use Nearest Neighbor Gaussian Process (Datta et al., 2016).

Non-geostatistical interpolation methods link data to exogenous variables known on though a model (i.e. GLM, GAM...). In environmental studies, theses model are the basis of the family called "species model distribution (SDM)" (for further information see Guillera-Arroita (2017)). For these methods, data does not require to be geo-referenced but, to be able to extrapolate, there must be an establish link between Y values and exogenous

variables X known on each statistical units of Ω. Main difficulties are to construct the statistical model. Statistical units and variables X must be independent. User must choose a distribution law and a link function. Then, goodness of fit to data has to be checked and predictions on space must have a slight confidence interval (Gregoire, 1998). Problems may also rise when wanted to model from a qualitative variable and data are unbalanced on different modalities. As with any modeling approach, the interpretation and the quality of model output depend on the initial data set and whether the model assumptions are met sufficiently (Guillera-Arroita et al., 2015).

A second type of SIMs available to reconstruct Y on Ω is geostatistical interpolation methods. There is a large literature on this type of SIM but here we choose to focus on hierarchical nearest neighbor Gaussian process (NNGP) models (Datta et al., 2016). The main advantage of these methods is an affordable data computation time, even for large geostatistical datasets. This is due to subsuming estimation of the model parameters, prediction outcome and interpolation. NNPG models allow fast-approximated computation of the Gaussian Process (GP) likelihood on large spatial data sets. The approximation starts by writing the GP likelihood under conditional recursive form. The conditioning variables in each conditional density are then replaced by a much smaller subset of the variables. Choosing the nearest neighbors to condition often constitutes a good heuristic, hence giving the Nearest Neighbor Gaussian Process denomination.

When no data are available, a method is to construct a distribution model elsewhere and adjusted onsite. Bayesian statistical models can also be used when no data are available (Choy et al., 2009). In this case, prior information is obtained from expert knowledge. Initial model can be progressively updated once data are available. Third solution would be to conduct a pilot study.

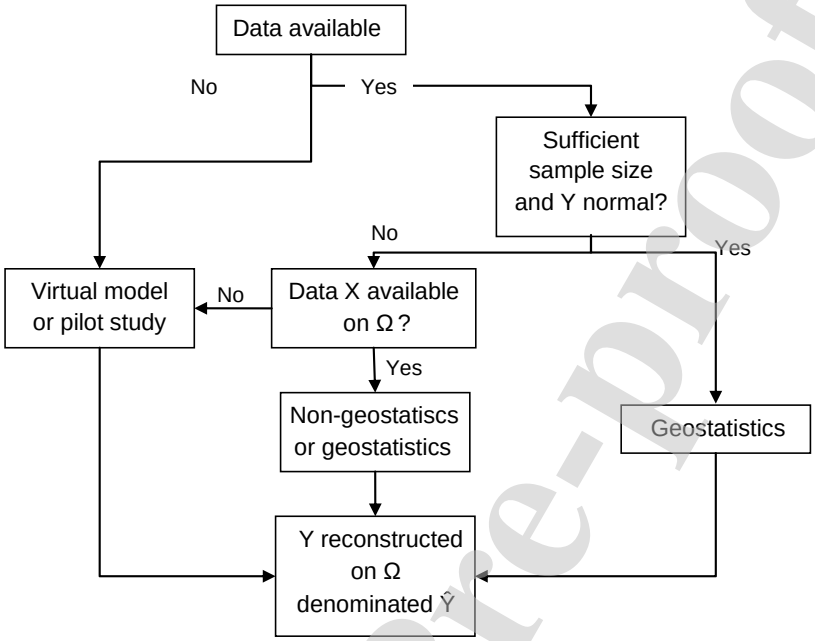


Figure 1: Process to reconstruct $\hat{Y}$ on Ω

We choose "meuse" dataset from R package sp. It comprises four heavy metals measured in the top soil in a flood plain along the river Meuse. You can find further detail here: https://cran.r-project.org/web/packages/gstat/vignettes/gstat.pdf. We will focus on the spatial distribution of zinc concentration in soil (in ppm).

3.1.1. Step 1.1: Available data analysis

So applied to this problem, the statistical population ω is the Meuse river watershed. We are interested on the zinc concentration parameter, previously called Y. 155 geo-referenced samples ω_j points were done and are available (whatever the way of drawing them). Fig.2 a) present their spatial dispersion. Fig.2 b) is zinc concentration density and Fig.2 c) is log-transformation.

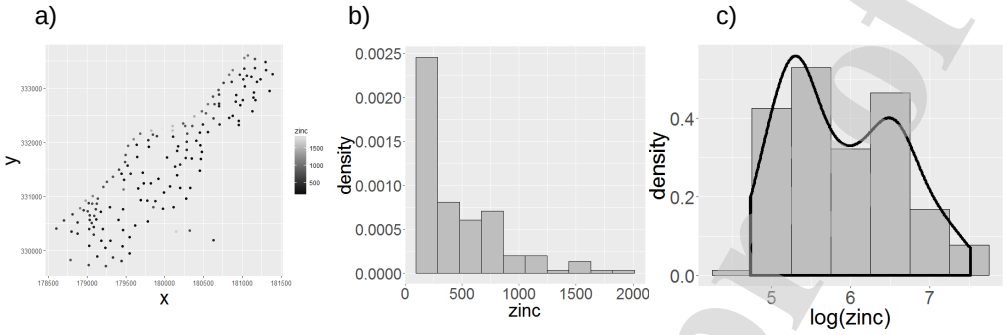


Figure 2: Visualisation of Y

3.1.2. Step 1.2: Reconstruction of $\hat{Y}$ on Ω

To reconstruct $\hat{Y}$ on Ω, exogenous variable *dist* (previously called X) is available for all statistical units ω_j of Ω. The statistical model-based $\hat{Y} \sim X$ can be constructed. Requirement for statistical model method is to have an exogenous variable X that can be linked to explain Y and available on all statistical units of Ω. Modelling is easier when there is a linear effect between X and Y (third plot). So here, data were log- transformed and root-squared transformed. The statistical model used here (Fig.3) is a linear model $log(zinc_i) = \alpha_0 + \sqrt{dist_i} + \epsilon_i$. The tab in Fig.3 represents model summary. Residuals plot shows the model fit well data. Last map displays a prediction of zinc concentration on Meuse watershed with the linear model.

We choose to use linear model for this example but we could have used other methods to reconstruct this random field. Theoretically, both the SIMs method lead to a prediction map, but some may have better predictive power. One way to evaluate models predictive power is by using leave-one-out cross-validation index. Cross-validation is a model validation technique mainly used when model goal is prediction. Leave-One-Out (or LOO) is an exhaustive cross-validation. Each learning set is created by taking all the samples except one, the test set being the sample left out. LOO cross-validation estimate of prediction error is the mean squared of the difference between the observed value and the predicted one:

$$CV_n = \frac{1}{n}\sum_{j=1}^{n}(\frac{y_j - \hat{y_j}}{1 - h_j})^2$$

Where h_j is the diagonal element of the hat matrix. It tells how much

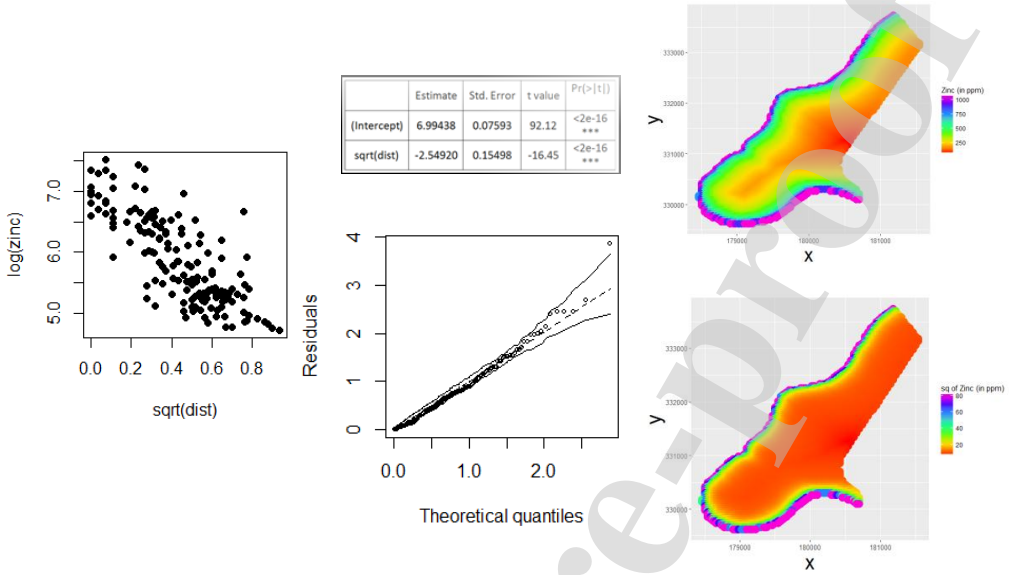


Figure 3: Model fitting and spatial prediction (Leave-One-Out Cross-Validation estimate of prediction error = 0.1914)

influence an observation has on its own fit. The more this index is close to 0, the more the model's ability to predict $\hat{Y}$ on Ω (in our case) is good. The distribution map of $\hat{Y}$ on Ω is the basis of the following step of our sequential process.

3.2. Step 2 – Simulations and sampling designs comparisons

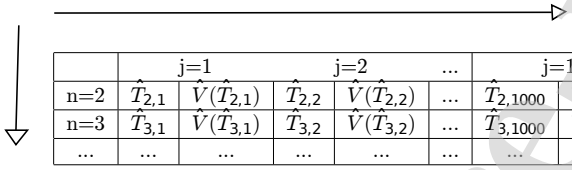
Once the statistical population $\hat{Y}$ on Ω is reconstructed, we can start assessing performances of chosen sampling designs.

3.2.1. Step 2.1: One by one sampling design simulation study of quality

As shown previously, the quality of a given sampling design can be assessed using efficiency of the corresponding estimator that depends on n. For a given sampling design, we will now show how to find the sample size that will permit to reach a wanted efficiency on the total parameter estimation. The same process will be applied to each sampling design selected by the user to be assessed.

Several values of n are tested. For each value n, a large number (>1 000) of simulations of samples arrangements are computed following the idea of

bootstrapping technique (Fontaine et al., 2008). This permits to remove random sampling fluctuations. Then for each combination $n \times j$ (one simulation) we calculate the estimate following the formula $\hat{T_{n,j}}$ and $V(\hat{T_{n,j}})$ (for the case where we want to estimate a total) and use the mean of the 1 000 simulations $\hat{T_n}$ and $V(\hat{T_n})$. The process is described step by step in the following (Figure 4): 1) Simulate 1 000 samples j for both increasing sampling efforts n; 2) Values of samples j are values of corresponding statistical units ω_j on previously reconstructed population $\hat{Y}$; 3) For all simulations j calculate $\hat{T_{n,j}}$ and $V(\hat{T_{n,j}})$; 4) Calculate mean of the 1 000 simulations $\hat{T_n}$ and $V(\hat{T_n})$.



	j=1		j=2		...	j=1000		$\hat{T}_n$	$\hat{V}(\hat{T}_n)$
n=2	$\hat{T}_{2,1}$	$\hat{V}(\hat{T}_{2,1})$	$\hat{T}_{2,2}$	$\hat{V}(\hat{T}_{2,2})$	...	$\hat{T}_{2,1000}$	$\hat{V}(\hat{T}_{2,1000})$	$\hat{T}_2$	$\hat{V}(\hat{T}_2)$
n=3	$\hat{T}_{3,1}$	$\hat{V}(\hat{T}_{3,1})$	$\hat{T}_{3,2}$	$\hat{V}(\hat{T}_{3,2})$	...	$\hat{T}_{3,1000}$	$\hat{V}(\hat{T}_{3,1000})$	$\hat{T}_3$	$\hat{V}(\hat{T}_3)$
...	...	...	...	...	...	...	...	...	...

Figure 4: Simulation process

These steps have to be done for all assessed sampling designs.

3.2.2. Step 2.2: Define optimal design and optimal sample size

For the previously assessed sampling designs, we have now $\hat{T_n}$ and $V(\hat{T_n})$ depending on sample size. The decision process to choose n_{opt} for each sampling design is based on an acceptable level of accuracy on results estimates. Therefore, the user needs to set this level called L and calculate margin of error with this level at all sampling size. To do so, we:

1) Set the level L of accuracy to be reached on estimates; 2) Calculate margin of error ME_n of size L for all sampling size;

$$ME_n = \frac{2t_\alpha \sqrt{\frac{V(\hat{T_n})}{n}}}{\hat{T_n}} \times 100$$

3) When margin of error ME_n is under the level L fixed in 1), optimal sample size is n_{opt} for this sampling design at this level of accuracy. Among the sampling assessed designs, the one that needs fewer samples than other to reach a same margin of error in total estimation is chosen as the optimal sampling design.

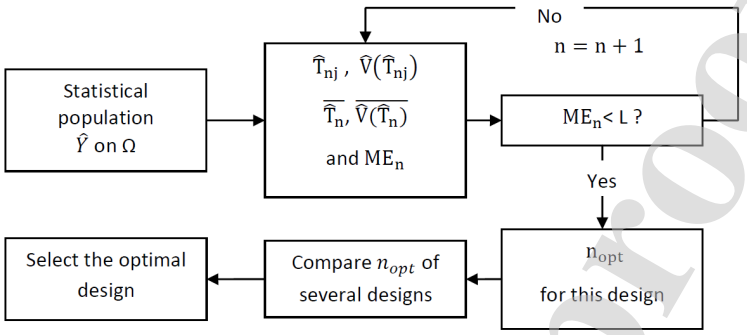


Figure 5: Process to assess optimal sampling design and sample size.

Meuse dataset is still used here. We assessed simple random sampling (SRS), systematic sampling (SSS) and local pivotal method (LPM) on the random field $\hat{Y}$ previously created. We set three different levels of accuracy to reach on total estimator result: 5, 10 and 20 percent, from the more to the less accurate.

	SRS			SSS			LPM		
nopt	5	10	20	5	10	20	5	10	20
%	50	26	40	40	25	15	34	20	11

Table 1: Optimal sampling designs and optimal sample sizes for zinc survey on Meuse watershed

LPM needs fewer samples than the other designs to achieve a same accuracy on total zinc estimation (Tab.2), LPM is so the optimal sampling design for this resource. The optimal sample size depends on the level of accuracy we want to reach on zinc total estimate. Once the number of samples is determined, user can easily define the cost of his survey depending on travel costs, survey times, analysis time (etc. . .).

4. PROCESS ILLUSTRATION

Before publishing this sequential process, we developed and assessed it throughout an example of manila clam stock monitoring program (Kermorvant et al., 2017, 2019a). In Arcachon day (SW France), scientists and commercial fishermen have developed a monitoring survey to estimate clam

stocks to assist in implementing a sustainable management strategy. The survey design used for these surveys was a standard stratified random sampling (StRS). The survey has been undertaken almost every 2 years since 2006 (Caill-Milly et al., 2006, 2008; Sanchez et al., 2010, 2012, 2014). Each survey costs approximately €50 000, with funding provided by ∼20% of the commercial fishermen. Due to shortfall in funds, this monitoring was not done in 2016 and could not be made up in 2017. To avoid this once again, we optimized this monitoring based on previous monitored years. We found that using a Generalized Tessellation Random Sampling (GRTS) instead of a simple random sampling could enhance estimates of the survey or decrease the survey cost. And used GRTS sampling design for 2018 survey.

4.1. Data understanding and spatial field reconstruction

To illustrate our process, we will use Manila clam monitoring in 2018 (Sanchez et al., 2018). During this survey 533 samples were gathered with a Generalized Tessellation Random Sampling (GRTS). The statistical population Ω in this example is defined as the places where samples can effectively be done. The targeted parameters is total of biomass in $g.m$. We use a hierarchical nearest neighbour Gaussian process model to interpolate the spatial field Y at spatial positions ω_j. The particularity of this Gaussian process is it follows a normal distribution with a recursive conditional form NNGP. As it include a random component, any any simulation is just a possible image of $\hat{Y}$ among a lot. We ran the optimisation process part from 3 among these different simulations (Fig.6) to illustrate the random property.

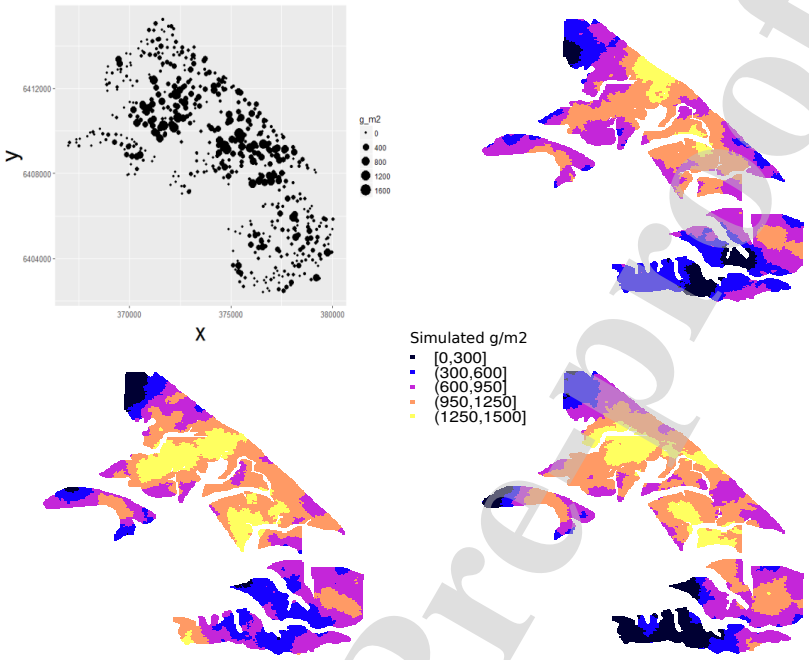


Figure 6: Data gathered during 2018 monitoring and 3 possible interpolations with NNGP.

4.2. Survey optimization

Three sampling designs were assessed: generalized tessellation sampling (GRTS - Stevens and Olsen (2004)), systematic sampling (SS) and simple random sampling (SRS). We fixed the level of accuracy to 10%. The following table summarizes results.

	Sampling Design		
	GRTS	SS	SRS
Sim 1	99	136	138
Sim 2	97	129	133
Sim 3	103	151	151

Table 2: Optimal sampling designs and optimal samples sizes for the three simulated fields (Sim1, Sim 2 and Sim 3) of Manila clam survey in Arcachon Bay

The spatially balanced sampling design (GRTS) achieves desired accuracy with less samples than the two others sampling designs. This means GRTS is the optimal sampling design for this survey. Optimal sample size depends

on the simulated field. Higher number of simulated field must be created to overcome this issue.

5. DISCUSSION

Where to sample and how many samples have to be done is often overlooked in surveys, resulting into non precise and non reproducible data-sets. Results of inadequate surveys can be misleading and hazardous not only because they fail to answer to the study problem but also because they can create the illusion that something useful was done (Peterman, 1990). Following these lacks of robust survey designs and sampling strategies (also highlighted by Hayward et al. (2015)), we construct a robust and reproducible process permitting sampling strategies' results to be non contentious. We assess our sequential framework on the monitoring of Manila clam resource on Arcachon bay. By selecting a more performing sampling design for this survey, we could decrease the number of samples of the overall survey (Kermorvant et al., 2017, 2019a).

Because sampling design efficiency depends on population properties, best sampling design and sample size will vary between studied populations. Hence, a population could be i.i.d., spatial auto-correlated (SAC), spatially stratified heterogeneity (SSH), both SAC and SSH and these characteristics will modify our method results. For example, it is now proved that spatially balanced sampling designs (SBS) are more efficient than simple random sampling when the studied population in SSH (Stevens and Olsen, 2004; Barabesi and Franceschi, 2011; Grafström and Tillé, 2013; Robertson et al., 2013; Kermorvant et al., 2019b). Furthermore, if co-variates are available, some SBS can also balance samples in this co-variates dimensions to gain in efficiency (Brown et al., 2015). The use of a stratified strategy could also enhance the sampling design efficiency. The spatially stratified heterogeneity (SSH) of a population can be tested by the stratified heterogeneity q statistic (Wang et al., 2016). This allow to construct good zoning that is critical to provide good-quality estimates (Li et al., 2008). Then, to estimate mean values for each strata and for the global population, the sandwich estimator would be appropriate (Wang et al., 2013).

We want to warn users of our method that is always risky to build one model on top of the other. The reconstructed distribution of $\hat{Y}$ on Ω needs to be very representative of the targeted real population. Especially because

we know the interpolation model used to reconstruct $\hat{Y}$ on Ω will have uncertainties. If this is not taken into account in studies, method results may be biased or false. For example, the simulations will under or over-estimate the needed sample size and can fail to select the best sampling design. The created random field must be as close as possible of that one encounters when starting to sample. A perspective of the manila clam example is to build a $\hat{Y}$ population even more representative than we currently do. We will enhance the present method (by doing more simulations of the spatial field, calculate confidence intervals and use environmental data) but also compare several SIMs methods to reconstruct this spatial field.

We also want to discuss about the ME_n formula when samples are non-normally distributed. ME_n depends on t-statistics and so fundamental assumptions are normal distribution and independence of data, especially for small sample sizes. This could have an impact on the L parameter of ME_n formula. There are various methods to deal with the normality issue. The most commonly used are the method based on the Central Limit Theorem, the Bootstrap method and the back-transformation method (for further information see Pek et al. (2017)).

The process we presented here differs from already published ones because it allows taking into account prior knowledge of the population. One of its strength is that it is based on a simulation study (Zurell et al., 2010) and so all possible strategies can be assessed, without excessive expenditures. Framework results is an optimal sample size by assessed sampling design for a desired accuracy in the results. However, the practitioner can define more than one *a priori* accuracy to be reached in the estimate and compare sampling designs and sample sizes needed to achieve them. As sample size and total survey cost are closely related, calculating the cost-effectiveness of several combinations is possible and the most appropriate one can be selected, before going on field. Having the possibilities to assess a large amount of sampling designs, choosing the best one and finding the optimal sample size are very relevant for studies where funds are often a limiting factor. In this sense, we choose to use virtual ecology. They reproduce as close as possible the distribution of the variable of interest and so they can be used to compare sampling designs (Albert et al., 2010; Zurell et al., 2010), without being forced to assess all of them on the field.

6. CONCLUSION

We are convinced that our general process will be useful for scientists and managers. We developed it keeping in mind that it must be adaptable to any survey and its special features. It allows choosing the more efficient sampling design, leading in reducing sampling size and/or increasing accuracy of results. This process can also be employed when attempting to develop a new monitoring, thus by selecting the best sampling design and the best sampling size from the beginning.

7. Acknowledgments:

This work was supported by "Communauté d'Agglomération Pays Basque – Euskal Hirigune Elkargoa" through a thesis grant. We particulary want to acknowledge Verena Trenkel (Ifremer), Ismaël Bernard (Eurêka Mer), the AUDAP (Agence D'Urbanisme Atlantique et Pyrénées) and professional fishing organization for valuable comments and help. CK, FD, NB, SC and NCM conceived the ideas and designed methodology; CK and NCM strongly participated in data collection regarding Manila clam.

References

Albert, C.H., Yoccoz, N.G., Edwards, T.C., Graham, C.H., Zimmermann, N.E., Thuiller, W., 2010. Sampling in ecology and evolution–bridging the gap between theory and practice. Ecography 33, 1028–1037.

Barabesi, L., Franceschi, S., 2011. Sampling properties of spatial total estimators under tessellation stratified designs. Environmetrics 22, 271–278.

Bernhardt, E.S., Palmer, M.A., Allan, J.D., Alexander, G., Barnas, K., Brooks, S., Carr, J., Clayton, S., Dahm, C., Follstad-Shah, J., 2005. Synthesizing US river restoration efforts. American Association for the Advancement of Science.

Brown, J., Robertson, B., McDonald, T., 2015. Spatially Balanced Sampling: Application to Environmental Surveys. Spatial Statistics conference 2015 27, 6–9. URL: http://www.sciencedirect.com/science/article/pii/S1878029615003205, doi:10.1016/j.proenv.2015.07.108.

Brown, J.A., 2003. Designing an efficient adaptive cluster sample. Environmental and Ecological Statistics 10, 95–105.

Caill-Milly, N., Bobinet, J., Lissardy, M., Morandeau, G., Sanchez, F., 2008. Campagne d'évaluation du stock de palourdes du bassin d'Arcachon-Année 2008. Rapport de contrat 17800.

Caill-Milly, N., Duclercq, B., Morandeau, G., 2006. Campagne d'évaluation du stock de palourdes du bassin d'Arcachon-Année 2006. Rapport 2218. Ifremer. France. URL: http://archimer.ifremer.fr/doc/00000/2218/.

Carvalho, S.B., Gonçalves, J., Guisan, A., Honrado, J.P., 2016. Systematic site selection for multispecies monitoring networks. Journal of Applied Ecology 53, 1305–1316.

Choy, S.L., O'Leary, R., Mengersen, K., 2009. Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. Ecology 90, 265–277.

Cochran, W.G., 1977. Sampling Techniques: 3d Ed. Wiley.

Datta, A., Banerjee, S., Finley, A.O., Gelfand, A.E., 2016. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. Journal of the American Statistical Association 111, 800–812.

Field, S.A., Tyre, A.J., Possingham, H.P., 2005. Optimizing allocation of monitoring effort under economic and observational constraints. The Journal of Wildlife Management 69, 473–482.

Grafström, A., 2012. Spatially correlated Poisson sampling. Journal of Statistical Planning and Inference 142, 139–147. URL: http://www.sciencedirect.com/science/article/pii/S0378375811002734, doi:10.1016/j.jspi.2011.07.003.

Grafström, A., Lundström, N.L., Schelin, L., 2012. Spatially balanced sampling through the pivotal method. Biometrics 68, 514–520.

Grafström, A., Matei, A., 2018. Spatially Balanced Sampling of Continuous Populations. Scandinavian Journal of Statistics 45, 792–805.

Grafström, A., Tillé, Y., 2013. Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. Environmetrics 24, 120–131.

Gregoire, T.G., 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Canadian Journal of Forest Research 28, 1429–1447.

Guillera-Arroita, G., 2017. Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. Ecography 40, 281–295.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R., Wintle, B.A., 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography 24, 276–292.

Guillera-Arroita, G., Ridout, M.S., Morgan, B.J., 2010. Design of occupancy studies with imperfect detection. Methods in Ecology and Evolution 1, 131–139.

Hayward, M.W., Boitani, L., Burrows, N.D., Funston, P.J., Karanth, K.U., MacKenzie, D.I., Pollock, K.H., Yarnell, R.W., 2015. Ecologists need robust survey designs, sampling and analytical methods. Journal of Applied Ecology 52, 286–290.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association 47, 663–685.

Kang, S.Y., McGree, J.M., Drovandi, C.C., Caley, M.J., Mengersen, K.L., 2016. Bayesian adaptive design: improving the effectiveness of monitoring of the Great Barrier Reef. Ecological applications 26, 2637–2648.

Kenkel, N., Juhász-Nagy, P., Podani, J., 1990. On sampling procedures in population and community ecology. Progress in theoretical vegetation science 83, 195–207. doi:https://doi.org/10.1007/BF00031692.

Kermorvant, C., Caill-Milly, N., Bru, N., D'Amico, F., 2019a. Optimizing cost-efficiency of long term monitoring programs by using spatially balanced sampling designs: The case of manila clams in Arcachon bay. Ecological Informatics 49, 32–39. doi:https://doi.org/10.1016/j.ecoinf.2018.11.005.

Kermorvant, C., Caill-Milly, N., D'Amico, F., Bru, N., Sanchez, F., Lissardy, M., Brown, J., 2017. Optimization of a survey using spatially balanced sampling: a single-year application of clam monitoring in the Arcachon Bay (SW France). Aquatic Living Resources 30, 37.

Kermorvant, C., D'Amico, F., Bru, N., Caill-Milly, N., Robertson, B., 2019b. Spatially balanced sampling designs for environmental surveys. Environmental monitoring and assessment 191, 524.

Legg, C.J., Nagy, L., 2006. Why most conservation monitoring is, but need not be, a waste of time. Journal of environmental management 78, 194–199. doi:https://doi.org/10.1016/j.jenvman.2005.04.016.

Levy, P.S., Lemeshow, S., 2013. Sampling of populations: methods and applications. John Wiley & Sons, New Jersey.

Li, J., Heap, A.D., 2014. Spatial interpolation methods applied in the environmental sciences: A review. Environmental Modelling & Software 53, 173–189.

Li, L., Wang, J., Cao, Z., Zhong, E., 2008. An information-fusion method to identify pattern of spatial heterogeneity for improving the accuracy of estimation. Stochastic Environmental Research and Risk Assessment 22, 689–704.

Liberts, M., 2013. The cost efficiency of sampling designs 14, 7–30.

Likens, G., Lindenmayer, D., 2018. Effective ecological monitoring. CSIRO publishing.

Lohr, S., 2009. Sampling: design and analysis. Nelson Education, Boston.

Macdonald, I.A., 2009. Comparison of sampling techniques on the performance of Monte-Carlo based sensitivity analysis , 992–999.

McDonald, T., 2014. Sampling Designs for Environmental Monitoring, in: Introduction to Ecological Sampling. Chapman and Hall/CRC, pp. 145–166.

Moore, A.L., McCarthy, M.A., 2016. Optimizing ecological survey effort over space and time. Methods in Ecology and Evolution 7, 891–899. doi:https://doi.org/10.1111/2041-210X.12564.

Nichols, J.D., Williams, B.K., 2006. Monitoring for conservation. Trends in ecology & evolution 21, 668–673. doi:https://doi.org/10.1016/j.tree.2006.08.007.

Pek, J., Wong, A.C., Wong, O.C., et al., 2017. Confidence intervals for the mean of non-normal distribution: transform or not to transform. Open Journal of Statistics 7, 405.

Peterman, R.M., 1990. Statistical power analysis can improve fisheries research and management. Canadian Journal of Fisheries and Aquatic Sciences 47, 2–15. doi:https://doi.org/10.1139/f90-001.

Petitgas, P., 2001. Geostatistics in fisheries survey design and stock assessment: models, variances and applications. Fish and Fisheries 2, 231–249. doi:https://doi.org/10.1046/j.1467-2960.2001.00047.x.

Rajabi, M.M., Ataie-Ashtiani, B., 2014. Sampling efficiency in Monte Carlo based uncertainty propagation strategies: application in seawater intrusion simulations. Advances in Water Resources 67, 46–64. doi:https://doi.org/10.1016/j.advwatres.2014.02.004.

Roberts, K.A., 1991. Field monitoring: confessions of an addict, in: Monitoring for conservation and ecology. Springer, Dordrecht, pp. 179–211.

Robertson, B., Brown, J., McDonald, T., Jaksons, P., 2013. BAS: Balanced acceptance sampling of natural resources. Biometrics 69, 776–784. doi:https://doi.org/10.1111/biom.12059.

Robertson, B., McDonald, T., Price, C., Brown, J., 2017. A modification of balanced acceptance sampling. Statistics & Probability Letters 129, 107–112. doi:0167-7152.

Robertson, B., McDonald, T., Price, C., Brown, J., 2018. Halton iterative partitioning: spatially balanced sampling via partitioning. Environmental and Ecological Statistics 25, 1–19. doi:https://doi.org/10.1007/s10651-018-0406-6.

Rudders, D., 2011. A Simulation Study to Evaluate Sampling Designs for Highly Autocorrelated Populations: With an Application to Sea Scallop Closed Areas , 550–550.

Sanchez, F., Caill-Milly, N., De Casamajor Marie-Noelle, L.M., 2012. Campagne d'évaluation du stock de palourdes du bassin d'Arcachon. Rapport de contrat 24114. Ifremer. France.

Sanchez, F., Caill-Milly, N., Lissardy, M., 2018. Campagne d'évaluation du stock de palourdes du bassin d'Arcachon. Année 2018. . Technical Report R.ODE/LITTORAL/LER AR 18.015. URL: https://archimer.ifremer.fr/doc/00477/58897/.

Sanchez, F., Caill-Milly, N., Lissardy, M., Bru, N., 2014. Campagne d'évaluation de stock de palourdes du bassin d'Arcachon. Rapport 34383. Ifremer. France.

Sanchez, F., Caill-Milly, N., Lissardy, M., De Casamajor, M.N., Morandeau, G., 2010. Campagne d'évaluation du stock de palourdes du bassin d'Arcachon. Rapport de contrat 16331. Ifremer. France.

Sica, G.T., 2006. Bias in Research Studies 1. Radiology 238, 780–789.

Stehman, S.V., Overton, W.S., 1996. Spatial sampling. Practical handbook of spatial statistics , 31–63.

Stevens, D.L., Olsen, A.R., 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99, 262–278.

Tillé, Y., 2011. Théorie des sondages: échantillonnage et estimation en population finie: cours et exercices avec solutions. volume 13. Dunod.

Vicente, J.R., Alagador, D., Guerra, C., Alonso, J.M., Kueffer, C., Vaz, A.S., Fernandes, R.F., Cabral, J.A., Araújo, M.B., Honrado, J.P., 2016. Cost-effective monitoring of biological invasions under global change: a model-based framework. Journal of Applied Ecology 53, 1317–1329.

Vos, P., Meelis, E., Ter Keurs, W., 2000. A framework for the design of ecological monitoring programs as a tool for environmental and nature management. Environmental monitoring and assessment 61, 317–344.

Wang, J.F., Haining, R., Liu, T.J., Li, L.F., Jiang, C.S., 2013. Sandwich estimation for multi-unit reporting on a stratified heterogeneous surface. Environment and Planning A 45, 2515–2534.

Wang, J.F., Jiang, C.S., Hu, M.G., Cao, Z.D., Guo, Y.S., Li, L.F., Liu, T.J., Meng, B., 2012. Design-based spatial sampling: Theory and implementation. Environmental modelling & software 40, 280–288.

Wang, J.F., Zhang, T.L., Fu, B.J., 2016. A measure of spatial stratified heterogeneity. Ecological Indicators 67, 250–256.

Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B., 2010. The virtual ecologist approach: simulating data and observers. Oikos 119, 622–635.