Spatially balanced sampling designs for environmental surveys

Kermorvant Claire [1, *], D'amico Frank [1], Bru Noëlle [1], Caill-Milly Nathalie [2], Robertson Blair [3]

[1] Laboratoire de Mathématiques et de leurs Applications de Pau – MIRACNRS/Univ Pau & Pays Adour/E2S UPPAAnglet, France
[2] Ifremer, Laboratoire Environnement Ressources d'ArcachonAnglet, France
[3] School of Mathematics and StatisticsUniversity of CanterburyChristchurch, New Zealand

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

Abstract :

Some environmental studies use non-probabilistic sampling designs to draw samples from spatially distributed populations. Unfortunately, these samples can be difficult to analyse statistically and can give biased estimates of population characteristics. Spatially balanced sampling designs are probabilistic designs that spread the sampling effort evenly over the resource. These designs are particularly useful for environmental sampling because they produce good-sample coverage over the resource, they have precise design-based estimators and they can potentially reduce the sampling cost. The most popular spatially balanced design is Generalized Random Tessellation Stratified (GRTS), which has many desirable features including a spatially balanced sample, design-based estimators and the ability to select spatially balanced oversamples. This article considers the popularity of spatially balanced sampling, reviews several spatially balanced sampling designs and shows how these designs can be implemented in the statistical programming language R. We hope to increase the visibility of spatially balanced sampling and encourage environmental scientists to use these designs.

Keywords : BAS, GRTS, LPM, Probabilistic sampling, Spatially balanced

28 When sampling an environmental resource, it is important to randomly choose the sampling locations over the

29 study area to provide formal statistical inference from the sample to the population. Smith et al. (2017)

30 established that in ecology, 12% of field studies selected samples using simple random sampling (SRS) and 9%

31 used systematic sampling. These methods are probabilistic sampling designs (meaning there is an element of

32 randomness in selecting their samples) and have well established statistical properties. However, most of the

33 ecological studies in Smith et al.'s (2017) review were not probabilistic sampling designs. Some studies used

34 haphazard or subjective judgement sampling methods and some studies did not specify how their samples were

35 drawn (Smith et al. 2017). This is troubling because data gathered in a haphazard or subjective way can produce

36 unrepresentative samples and biased estimates of population characteristics (Albert et al. 2010; Levy and

37 Lemeshow 2013).

38 Choosing an appropriate sampling design for a particular study can be difficult and there is no *best* design for all

39 research questions (Kenkel et al. 1990; Stehman and Overton 1994). This choice depends on many things

40 including the study objectives, available sampling frames and known auxiliary variables. This paper focuses on

41 making an inference from a sample to the entire population using a specific class of probabilistic sampling

42 designs called spatially balanced sampling designs. These designs were chosen because they are particularly

43 useful for sampling natural resources (Stevens and Olsen 2004). For a full treatment on the subject, the reader is

44 referred to Benedetti et al. (2015).

45

46 **What is spatially balanced sampling?**

47 To achieve good estimates of population characteristics, the spatial pattern of the sample should be similar to the

48 spatial pattern of the population. However, the spatial pattern of the response variable may not be known before

49 the sample is drawn. Fortunately, a common spatial feature in environmental sampling is that nearby locations

50 tend to be more similar because they interact with one another and are influenced by the same set of factors

51 (Stevens and Olsen 2004). Therefore, an effective strategy is to spatially spread the sample evenly over the

52 resource. A sample that is evenly spread over the resource is called a spatially balanced sample. Stevens and

53 Olsen (2004) introduced the phrase *spatially balanced sampling* and proposed a statistic that measures the spatial

54 balance or regularity of a sample using Voronoi polygons.

55

3

Why should environmental scientists use spatially balanced designs?

The potential advantages of using spatially balanced sampling have been demonstrated in the field of environmental science (Stevens and Olsen 2004; Christianson and Kaufman 2016; McGarvey et al. 2016). The first advantage is that spatially balanced samples are evenly spread over the resource. Covering the resource avoids under-coverage and over-coverage, which can happen with probabilistic sampling designs with poor spatial balance (Stevens and Olsen 2004; Christianson and Kaufman 2016). If a researcher's analysis requires clusters of nearby observations, spatially balanced cluster sampling could be useful (Robertson et al. 2017).

Spatially balanced samples can be very efficient when the response variable has a strong spatial trend (Stevens and Olsen 2004; Barabesi and Franceschi 2011; Grafström and Lundström 2013; Robertson et al. 2013; Benedetti et al. 2017), because their design-based estimators take into account spatial heterogeneity (Wang et al., 2012) and spatial auto-correlation (Haining 2003). When estimating a population total or mean using the Horvitz-Thompson estimator, the local mean variance estimator (Stevens and Olsen 2003) is a popular variance estimator. There have been many studies showing the effectiveness of spatially balanced sampling with this estimator on a variety of populations with different spatial structures (c.f. Stevens and Olsen 2004; Grafström et el. 2012; Grafström and Lundström 2013; Robertson et al. 2013; Robertson et al. 2018). If the spatial trend is weak or if there is no trend at all, there is no statistical advantage in choosing spatially balanced designs over other probabilistic designs (Robertson et al. 2013).

Another potential advantage of spatially balanced sampling is reduced sample cost. As mentioned above, spatially balanced designs can produce precise design-based estimators when there is a spatial trend in the response variable. Hence, to achieve a desired level of precision, fewer observations may be required when spatially balanced sampling is used. This was illustrated by Kermorvant et al. (2017, 2019) for a clam monitoring program in Arcachon Bay, France. They showed that when spatially balanced designs were used, the total survey cost was reduced by 30% when compared with simple random sampling.

Another useful feature of some spatially balanced designs is that they can draw spatially balanced oversamples (replacement units). This is particularly useful for environmental sampling because when sample units cannot be observed (private property, inaccessible, too dangerous, etc.), replacement units are often required to achieve the desired sample size (Stevens and Olsen 2004; Robertson et al. 2018; Theobald et al. 2007). For example, in the Oklahoma state-wide stream and river monitoring program, only 130 of the 177 randomly chosen sites could be

84 observed (Oklahoma Water Resources Board 2013). Of the unobservable sites, eight were on private land and

85 39 had dry channels or were not accessible. To achieve the desired sample size, the researchers selected

86 replacement sites from an oversample drawn using simple random sampling. The potential advantage of using

87 spatially balanced oversamples is that the observed sample maintains some degree of spatial balance over the

88 observable resource (Stevens and Olsen 2014; Robertson et al. 2018). Although oversampling is of practical

89 importance, it does not eliminate the non-response from unobservable units or the bias of an inference

90 (Robertson et al. 2018).

91

92 **Generalized Random Tessellation Stratified**

93 Generalized Random Tessellation Stratified (GRTS) (Stevens and Olsen 2004) is the most popular spatially

94 balanced sampling design for environmental studies (Grafström and Tillé 2013; Foster 2016). It was developed

95 by the U. S. Environmental Protection Agency (EPA) for the National Environmental Monitoring and

96 Assessment Program (Messer et al. 1991; Stevens & Olsen 2004). GRTS uses a complex algorithm to draw its

97 sample and we briefly discuss its main steps here. The reader is referred to Olsen et al. (2012) for a full, non-

98 technical description of GRTS. Initially, a grid is superimposed over the study area and each grid cell is

99 hierarchically numbered using a base four numbering system. The numbered grid cells are then randomly

100 permuted using reverse hierarchical ordering and mapped (in order) to the real line. A systematic sample from

101 the ordered grid cells is then drawn and one sampling unit is randomly selected from each of these grid cells. The

102 selected units are then mapped back to their respective locations in the study area, to yield the sample locations.

103 The systematic sampling and hierarchical ordering that GRTS uses ensures that the sample is spatially balanced

104 (Stevens and Olsen, 2004).

105 To investigate GRTS's popularity, the Google Scholar search engine was used because it provided access to a

106 wide range of publication types. It indexes journals papers (published online or in paper format), conference

107 proceedings, posters, and technical reports from research organizations in both the public and private sector. The

108 keyword "GRTS" was used for the search and we did not include citations from 2018. All the documents found

109 were categorised as either "publications" or "reports". Publications included peer-reviewed journal articles and

110 refereed conference proceedings, and reports included all other publication types. Results are displayed in Figure

111 1.

112

Fig. 1 Flow representation of use and/or citations of GRTS in the literature, publication date (stars) of several spatially balanced designs and R packages (arrows) are shown.

115

Our analysis found 600 documents citing GRTS throughout the world. The citation and/or use of GRTS showed a steady increase until 2013, after which it flattened out. At the beginning, there were more reports than publications, but this trend appears to have reversed since 2014. Most of the documents found were in the fields of environmental science (mostly ecology but also environmental chemistry) and statistics (new designs, tests and comparisons). There were only two publications from other fields. The first was in economics, where GRTS was compared to existing sampling designs for business surveys (Dickson et al. 2014) and the second was a thesis on sampling standards for maintenance management quality assurance (Chen 2018).

123

Other spatially balanced sampling designs

Many spatially balanced designs have been proposed in the literature. In this section we mention several approaches.

The Local Pivotal Method (LPM) (Grafström et al. 2012) is a flexible spatially balanced design that can draw equal and unequal probability samples in multiple dimensions. Unequal probability sampling can be more efficient than equal probability sampling if there is a positive correlation between the inclusion probabilities and the response values (Robertson et al. 2013). Additional dimensions could include auxiliary information such as ecological threats, time intervals, species population structure or environmental data (Brown et al. 2015). The Swedish national forest inventory, for example, has implemented LPM with five auxiliary variables (Grafström et al. 2017). LPM is a popular method with 100 citations (using Google Scholar), where most of its applications were related to forestry.

Grafström (2012) also presented spatially correlated Poisson sampling (SCPS). This design is a modification of correlated Poisson sampling (Bondesson and Thorburn 2008) that draws spatially balanced samples. LPM is algorithmically easier than SCPS, but SCPS may produce better results for some populations (Grafström and Schelin 2014).

139   Balanced acceptance sampling (BAS) (Robertson et al. 2013, 2017) is another spatially balanced design. BAS

140   uses the Halton sequence (Halton 1960) to spread its sample across multiple dimensions. BAS is conceptually

141   simple, computationally efficient and is particularly useful for drawing spatially balanced oversamples

142   (Robertson et al. 2018). We found 34 publications citing BAS, where most of the papers were methodological

143   rather than applied. BAS has been used to survey bats in Bighorn Canyon National Recreation Area (Keinath and

144   NRA 2016) and is being used for New Zealand's national monitoring program (van Dam-Bates et al. 2018). BAS

145   is well suited for areal resources (geographic areas), but it can be inefficient on some point resources (Robertson

146   et al. 2018). To improve the performance of BAS on point resources, Robertson et al. (2018) presented Halton

147   iterative partitioning (HIP). This spatially balanced design uses properties of the Halton sequence to partition a

148   point resource into nested boxes to draw its sample, rather than using the sequence itself.

149   Benedetti and Piersimoni (2017) presented a flexible class of spatially balanced designs that draw their samples

150   based on a within sample distance (Benedetti and Piersimoni 2017). The algorithm is simple to implement in

151   multiple dimensions and any distance or similarity measure can be used to define the within sample distance.

152   **Spatially balanced sampling packages in R**

153   Several R software (R Core Team 2014) packages are freely available to draw spatially balanced samples. To

154   draw GRTS samples, spsurvey (Kincaid and Olsen 2015) or SDraw (McDonald 2016) can be used. These

155   packages can draw samples from point resources (geographic locations), linear resources (rivers) and areal

156   resources (geographic areas), and can also draw spatially balanced oversamples. The spsurvey package can also

157   draw stratified spatially balanced GRTS samples with user defined strata.

158   The other spatially balanced designs mentioned in this article can be selected using the following packages.

159   BalancedSampling (Grafström and Lisic 2016) draws equal and unequal probability LPM and SCPS samples

160   from point resources. BAS and HIP samples/oversamples can be selected from point, linear and areal resources

161   using SDraw (McDonald 2016). Historical or legacy sites can also be incorporated into a BAS design (Foster et

162   al. 2014) using the MBHdesign package (Foster 2016). Finally, the R package Spbsampling (Pantalone et al.

163   2019) can be used to draw the within sample distance-based methods of Benedetti and Piersimoni (2017).

164   Figure 2 shows examples of equal probability spatially balanced samples of 150 points drawn from a point

165   resource using SDraw and BalancedSampling. An oversample of 20 points is also illustrated for BAS and GRTS.

7

166 Note how the oversample points are spatially balanced with respect to the primary sample. To illustrate the R

167 syntax for these packages, an annotated R script that creates Figure 2 is given in the supplementary material

168 section.

169

170 **Fig. 2. Several spatially balanced samples drawn using different designs, where open**
171 **symbols denote oversamples sites.**

172

173 **Conclusion**

174 Environmental scientists are beginning to use more advanced sampling designs to achieve robust statistical

175 results. Spatially balanced designs are particularly useful for environmental science because they can produce

176 good sample coverage over a resource, precise design-based estimators and potentially reduce sampling cost.

177 GRTS is the most popular spatially balanced sampling design and it is easy to implement using freely available

178 R packages like spsurvey and SDraw. Another useful feature of GRTS is that spatially balanced oversamples can

179 be drawn. Although oversampling is of practical importance, it does not eliminate the non-response from

180 unobservable units or the bias of an inference. Several other spatially balanced designs are also available, each

181 with an accompanying R package. LPM and SCPS samples can be drawn using BalancedSampling and SDraw

182 selects BAS and HIP samples/oversamples. Although spatially balanced sampling has mostly been used in

183 ecology, we encourage all environmental scientists to these designs when the research objective is to make an

184 inference from a sample to the entire population.

185

190 **References:**

191 Albert CH, Yoccoz NG, Edwards TC, et al (2010) Sampling in ecology and evolution—bridging the gap between
192   theory and practice. Ecography 33:1028–1037

8

193   Barabesi L, Franceschi S (2011) Sampling properties of spatial total estimators under tessellation stratified

194   designs. Environmetrics 22:271–278

195   Benedetti R, Piersimoni F (2017) A spatially balanced design with probability function proportional to the within
196   sample distance. Biom J 59:1067–1084

197   Benedetti R, Piersimoni F, Postiglione P (2015) Sampling spatial units for agricultural surveys. Springer

198   Benedetti R, Piersimoni F, Postiglione P (2017) Spatially balanced sampling: a review and a reappraisal. Int Stat
199   Rev

200   Bondesson L, Thorburn D (2008) A list sequential sampling method suitable for real-time sampling. Scand J Stat
201   35:466–483

202   Brown JA, Robertson BL, McDonald T (2015) Spatially Balanced Sampling: Application to Environmental
203   Surveys. Spat Stat Conf 2015 27:6–9. doi: 10.1016/j.proenv.2015.07.108

204   Chen Z (2018) Statistical analysis and sampling standards for maintenance management quality assurance
205   (MMQA)

206   Christianson DS, Kaufman CG (2016) Effects of sample design and landscape features on a measure of
207   environmental heterogeneity. Methods Ecol Evol 7:770–782. doi: 10.1111/2041-210X.12539

208   Conn PB, Moreland EE, Regehr EV, et al (2016) Using simulation to evaluate wildlife survey designs: polar bears
209   and seals in the Chukchi Sea. R Soc Open Sci 3:150561. doi: 10.1098/rsos.150561

210   Dickson MM, Benedetti R, Giuliani D, Espa G (2014) The use of spatial sampling designs in business surveys.
211   Open J Stat 4:345–354. doi: http://dx.doi.org/10.4236/ojs.2014.45034

212   Foster SD (2016) MBHdesign : Spatial Designs for Ecological and Environmental Surveys

213   Foster SD, Hosack GR, Hill NA, et al (2014) Choosing between strategies for designing surveys: autonomous
214   underwater vehicles. Methods Ecol Evol 5:287–297

215   Grafström A (2012) Spatially correlated Poisson sampling. J Stat Plan Inference 142:139–147. doi:
216   10.1016/j.jspi.2011.07.003

217   Grafström A, Lisic J (2016) BalancedSampling: balanced and spatially balanced sampling [online]. R package
218   version 1.5. 2

219   Grafström A, Lundström NL (2013) Why well spread probability samples are balanced. Open J Stat 3:36–41

220   Grafström A, Lundström NL, Schelin L (2012) Spatially balanced sampling through the pivotal method.
221   Biometrics 68:514–520

222   Grafström A, Matei A (2018) Spatially Balanced Sampling of Continuous Populations. Scand J Stat

223   Grafström A, Schelin L (2014) How to select representative samples. Scand J Stat 41:277–290

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

226   Grafström A, Zhao X, Nylander M, Petersson H (2017) A new sampling strategy for forest inventories applied to
227   the temporary clusters of the Swedish national forest inventory. Can J For Res 47:1161–1167

228   Haining RP (2003) Spatial data analysis: theory and practice. Cambridge University Press

9

229   Halton JH (1960) On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional
230   integrals. Numer Math 2:84–90

231   Keinath DA, NRA BC (2016) Bat population monitoring of Bighorn Canyon National Recreation Area: 2015
232   progress report. 3. pp

233   Kenkel N, Juhász-Nagy P, Podani J (1990) On sampling procedures in population and community ecology. In:
234   Progress in theoretical vegetation science. Springer, pp 195–207

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

238   Kermorvant C, Caill-Milly N, D'Amico F, et al (2017) Optimization of a survey using spatially balanced sampling:
239   a single-year application of clam monitoring in the Arcachon Bay (SW France). Aquat Living Resour
240   30:37

241   Levy PS, Lemeshow S (2013) Sampling of populations: methods and applications. John Wiley & Sons

242   MacKenzie DI (2006) Occupancy estimation and modeling: inferring patterns and dynamics of species
243   occurrence. Academic Press

244   McDonald T (2016) SDraw: Spatially Balanced Sample Draws for Spatial Objects

245   McGarvey R, Burch P, Matthews JM (2016) Precision of systematic and random sampling in clustered
246   populations: habitat patches and aggregating organisms. Ecol Appl 26:233–248

247   Messer JJ, Linthurst RA, Overton WS (1991) An EPA program for monitoring ecological status and trends.
248   Environ Monit Assess 17:67–78

249   Oklahoma Water Resources Board (2013) The Statewide Stream/River Probabilistic Monitoring Network -Final
250   Report. 3800 N. Classen, Oklahoma City, Oklahoma 73118

251   Olsen A, Kincaid T, Payton Q (2012) Spatially balanced survey designs for natural resources. Des Anal Long-
252   Term Ecol Monit Stud 126–150

253   Pantalone, Benedetti R, Piersimoni F (2019) Spbsampling: Spatially Balanced Sampling. [online]. R package
254   version 1.2.0

255   R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical
256   Computing, Vienna, Austria

257   Robertson B, Brown J, McDonald T, Jaksons P (2013) BAS: Balanced acceptance sampling of natural resources.
258   Biometrics 69:776–784

259   Robertson B, McDonald T, Price C, Brown J (2017) A modification of balanced acceptance sampling. Stat Probab
260   Lett

261   Robertson B, McDonald T, Price C, Brown J (2018) Halton iterative partitioning: spatially balanced sampling via
262   partitioning. Environ Ecol Stat 1–19

263   Sica GT (2006) Bias in Research Studies 1. Radiology 238:780–789

264   Smith AN, Anderson MJ, Pawley MD (2017) Could ecologists be more random? Straightforward alternatives to
265   haphazard spatial sampling. Ecography. doi: 10.1111/ecog.02821

266   Stehman SV (2009) Sampling designs for accuracy assessment of land cover. Int J Remote Sens 30:5243–5272

267   Stehman SV, Overton WS (1994) 9 Environmental sampling and monitoring. In: Handbook of Statistics. Elsevier,
268   pp 263–306

269   Stevens DL, Olsen AR (2004) Spatially balanced sampling of natural resources. J Am Stat Assoc 99:262–278

270   Stevens DL, Olsen AR (2003) Variance estimation for spatially balanced samples of environmental resources.
271   Environmetrics 14:593–610

272   Theobald DM, Stevens Jr DL, White D, et al (2007) Using GIS to generate spatially balanced random survey
273   designs for natural resource applications. Environ Manage 40:134–146

274   van Dam-Bates P, Gansell O, Robertson B (2018) Using balanced acceptance sampling as a master sample for
275   environmental surveys. Methods Ecol. Evol.

276

Figure2

GRTS

SRS

LPM

BAS

Figure1