Scientific Research Publishing

Open Journal of Statistics, 2018, 8, 706-720
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Estimating the Variance of the Proportion of Contaminated Soil by Petroleum Spills Using Two-Dimensional Systematic Sampling under Different Approaches

Diego Jarquin

Department of Agronomy and Horticulture, University of Nebraska, Lincoln, Nebraska, USA
Email: diego.jarquin@gmail.com

How to cite this paper: Jarquin, D. (2018) Estimating the Variance of the Proportion of Contaminated Soil by Petroleum Spills Using Two-Dimensional Systematic Sampling under Different Approaches. *Open Journal of Statistics*, **8**, 706-720.
https://doi.org/10.4236/ojs.2018.84046

Received: July 6, 2018
Accepted: August 20, 2018
Published: August 23, 2018

Abstract

In leading petroleum-producing countries like Kuwait, Brazil, Iran, Iraq and Mexico oil spills frequently occur on land, causing serious damage to crop fields. Soil remediation requires constant monitoring of the polluted area. One common monitoring method involves two-dimensional systematic sampling, which can be used to estimate the proportion of the contaminated soil and study the oil spills' geographic distribution. A well-known issue using this sampling design involves the analytical derivation of variance of the sample mean (proportion), which requires at least two independent samples. To address the problem, this research proposed a variance estimator based on regression and a corrected estimator using the autocorrelation Geary Index under the model-assisted approach. The construction of the estimators was assisted by geo-statistical models by simulating an auxiliary variable. Similar populations to those in real oil spills were recreated, and the accuracy of proposed estimators was evaluated by comparing their performance with other well-known estimators. The factors considered in this simulation study were: a) the model for simulating the populations (exponential and wave), b) the mean and the variance of the process, c) the level of autocorrelation among units. Given the statistical and computing burdens (bias, ratio between estimated and real variance, convergence and computer time), under the exponential model, the regression estimator showed the best performance; and for the wave model, the corrected version performed even better.

Keywords

Accurate Estimation of Variance, Model-Assisted Approach, Geary Index, Geostatistics, Oil Spills, Simulation

1. Introduction

Frequently, sampling in two dimensions is applied in small areas, resulting in small population in the situation when petroleum spills are studied. In places where these spills are common, oil contamination causes serious damage to soils and water bodies. Remediating the soil is expensive and requires careful monitoring. During this process, a soil sample is taken to estimate the proportion of contaminated area; however this method can be problematic if it does not yield accurate results. Historically, three different approaches have been used to perform and analyze sampling (design-based, model-based and model-assisted). One potentially superior monitoring strategy is the systematic sampling design, which offers a uniform coverage of the study area. Using this strategy, not only punctual estimations of the proportion and the variance are obtained, but also the collected information can be used to perform geo-statistical studies about the distribution of the pollutant area.

Nevertheless, this sampling design has an unresolved difficulty. To obtain an unbiased estimate of the variance of the proportion, at least two independent samples are required; otherwise, by using just one sample, only an approximation of the variance can be computed [1] [2]. This issue is more difficult to resolve in two dimensions than in a linear case because the units are arranged in a plane instead of a line. For example in a bilinear case, under the design-based approach, Marcello [3] proposed estimators that consider the existence of autocorrelation among the units in the sample. He showed that the estimation procedure can be significantly improved by taking the similarity among units in the sample into account. Li [4] in his dissertation and Opsomer [5] proposed a nonparametric version of the variance estimator using a model-based approach. These estimators are robust, but under this approach, a great number of replicates are required for accurate estimations [6], which results in an increased computing time. Recently, a model-assisted approach [7] has been employed to introduce a covariate explaining relatedness, and to improve the accuracy of the estimation. Most recently, Strand [8] compared three estimators for variance in systematic spatial samples. The first of these estimators was based on post-stratification of the data, the second one used a correction factor calculated from the spatial autocorrelation, and the third one was a model-based prediction calculated using values from semivariograms.

This paper introduces two new variance estimators constructed under the model-assisted approach based on geo-statistical models. The estimators are the regression estimator of the variance, $\hat{V}_{REG}(\hat{p})$ and its corrected version, $\hat{V}_{REG\_G}(\hat{p})$, which takes the existing autocorrelation among the units in the sample into account through the Geary Index [9]. This index measures the spatial autocorrelation among the units by determining whether the adjacent observations of the same phenomena are correlated. The spatial autocorrelation is more complex than the linear autocorrelation because the correlation is multi-dimensional and bi-directional. By conducting a simulation study, the accu-

racy was assessed and the performance of the proposed variance estimators was compared to other well-known estimators that exist in different approaches.

The article is organized as follows. Section 2 provides a brief explanation of the two-dimensional systematic sampling. In Section 3, the estimators used to overcome the variance estimator issues are introduced. Section 4 gives a description of the design and the simulation study. The results and discussion are presented in Section 5. Finally, Section 6 formulates the conclusions and offers some recommendations.

2. Systematic Sampling in Two Dimensions

Systematic sampling designs are commonly used in real-life applications due to their straightforward implementation. Moreover, when proportions are estimated in finite populations, their results are frequently more efficient than other sampling design alternatives (*i.e.* simple random sampling, stratified sampling, etc.). These properties made it attractive to consider this over an area where the population units are in a regular spaced array. Thus, this sampling design provides a uniform coverage of the area; which can take advantage of the information of the location of the sample units for accounting for the spatial correlation. For example, when geographically close sampling units show a high positive correlation (geographically closer units tend to be more similar than units more distant from each other, as in the case of an oil spill, it is possible to obtain more accurate estimations.

Two-dimensional systematic sampling consists of the random selection of an initial point, and the remaining points are selected by following a regular pattern (e.g. a rectangular or a square arrangement). In these arrangements, a sample is obtained by randomly selecting a square in the first domain (**Figure 1**), and then the squares that occupy the same position in the remaining domains are selected automatically. More formally, let's consider D as a continuous region that contains a finite number of units N, where these units are represented by some non-overlapping squares in the study area. To obtain these square units, it is necessary to divide the grid of regular squares over D; where this grid is formed by R rows and C columns, such that $N = R \times C$.

Thus, when a two-dimensional systematic sample of size n is needed, the N squares are grouped into $n = nC \times nR$ non-overlapping rectangular sub-regions or domains, and each domain contains $k = kC \times kR$ squares, where $k = N/n$ is the number of all possible systematic samples.

2.1. Estimation

Suppose there is a binary random variable $Z = f(t)$ that may take values 0 or 1 for a fixed value of t. Then, an unbiased estimator of the population proportion p (mean) of Z in D by using the j^{th} two-dimensional systematic sample is given by

$$\hat{p}_{SYS,j} = \frac{\sum_{i=1}^{n} z_{ij}}{n}, \tag{1}$$

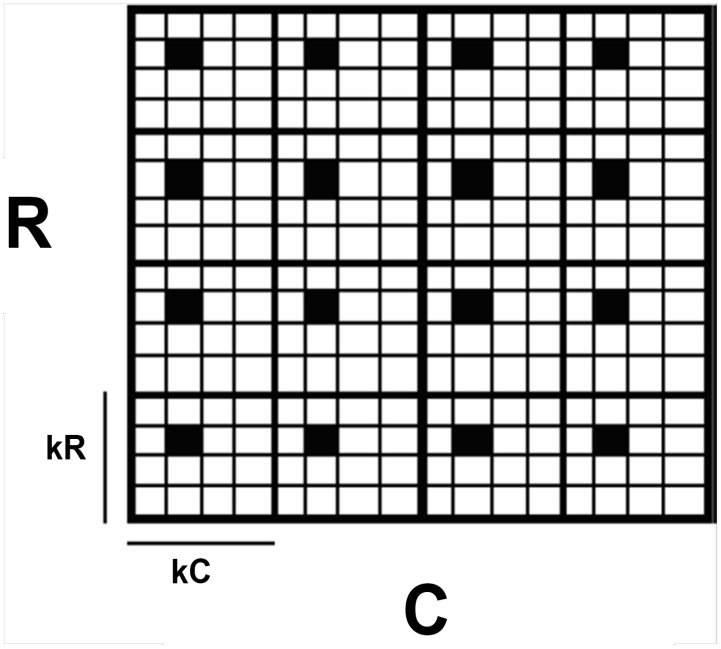

Figure 1. An example of a two-dimensional systematic sample taken from 16 domains and $N = 256$ with $n = 16$, $k = 16$, $kC = 4$ and $kR = 4$.

where $z_{ij} = \begin{cases} 1 & \text{if } T \geq t \\ 0 & \text{if } T < t \end{cases}$

with T as a continuous random variable and t as a fixed predetermined threshold.

The sampling error provides information about the variance of the estimator. An unbiased variance estimator for the population proportion (mean) can be obtained through

$$V_{SYS} = \frac{\sum_{j=1}^{k} \left(\hat{p}_{j.SYS} - p \right)^2}{k}. \quad (2)$$

Under this sampling design, computing an unbiased estimator for the variance requires at least two independent systematic samples, and by using a single sample, only approximations can be derived [3] [10].

2.2. Variance Estimation under Different Approaches

The estimation of the sampling error through a single systematic sample is more difficult in two dimensions than for the linear case, because the units are arranged in a plane instead of a line [3].

For selecting samples and performing correspondent analysis, three different approaches have been considered: design-based, model-based and model-assisted. In the design-based approach [1], the primary source of randomness comes from the sampling design. The model-based approach considers the values in a sample as the realized outcomes of random variables [11] [12]. Finally, the model-assisted approach uses an auxiliary variable, which is related to the

variable of interest [10].

The results (estimations) of these approaches are not directly comparable, because they arise under different assumptions. Nevertheless, a few authors have tried to give explanations to why and how to perform comparisons. For example, Särndal [11] showed that several of the classic results can be obtained and reinterpreted through the model-based theory.

Next, the variance estimators that were considered in this study for comparison purposes are described. Then, the proposed models are introduced.

A) Design-based approaches

1) Simple random sampling estimator

The variance estimator of the proportion (mean) under the simple random sampling estimation scheme can be written as

$$\hat{V}_{SRS}\left(\hat{p}\right)=\left(1-f\right)\frac{\hat{p}\left(1-\hat{p}\right)}{n-1}, \tag{3}$$

where $\hat{p}$ is as defined in (1) and f is the proportion of the population selected for a sample. $(1-f)$ is called the finite population correction or adjustment. In sampling without replacement, the sample variance is reduced by this factor.

The variance estimator (3) of the proportion (mean) under the simple random sampling scheme is unique in a way that it can be used without taking the spatial array of the units into account. Good estimations are expected if the distance between the sampled squares is large enough to have a small spatial correlation, or zero in the best case. In the presence of high homogeneity between sampled units the variance will be underestimated [2].

2) Geary's spatial autocorrelation index

Marcello [3] proposed two estimators that consider the autocorrelation in the sample, and they are obtained by correcting (3).

The first estimator that formally takes the presence of the correlation into account is constructed with the autocorrelation Geary Index, and can be written as

$$\hat{V}_{GI}\left(\hat{p}\right)=\hat{V}_{SRS}\left(\hat{p}\right)c_j, \tag{4}$$

where $c_j=\dfrac{n-1\sum_{i=1}^{n}\sum_{i\neq l}^{n}\left(z_l-z_i\right)^2\delta_{il}}{2\sum_{i=1}^{n}\sum_{i\neq l}^{n}\delta_{il}\sum_{i=1}^{n}\left(\left(z_i-\hat{p}\right)\right)^2}$

is the Geary index computed for the j^{th} sample, $\delta_{il}=1$ if the i^{th} and the l^{th} units are in adjacent domains or 0 otherwise, and z is defined as before (*i.e.*, binary outcome). Here, c_j measures the grade of similarity among sampling units.

3) Moran's spatial autocorrelation statistic

This estimator is constructed with the Moran's spatial autocorrelation statistic.

$$\hat{V}_{MS}\left(\hat{p}\right)=\hat{V}_{SRS}\left(\hat{p}\right)w_j, \tag{5}$$

where

$$w_j = \left[1 + \frac{2}{\log\left(l_j\right)} + \frac{2}{\left(\frac{1}{l_j} - 1\right)}\right],$$

and

$$l_j = \frac{n}{\sum_{i=1}^{n}\sum_{l\neq i}^{n}\delta_{il}} - \frac{\sum_{i=1}^{n}\sum_{l\neq i}^{n}\left(y_i - \hat{p}\right)\left(y_l - \hat{p}\right)\delta_{il}}{\sum_{i=1}^{n}\left(y_i - \hat{p}\right)^2}$$

and δ_{il} is defined as in (4).

Here, I_j is the Moran autocorrelation statistic computed for the j^{th} sample, and it measures the dissimilarity grade between sampling units.

B) Model-based approach

Briefly, this approach considers the N values of a population $\{y_1, y_2, \cdots, y_N\}$ as realized outcomes of N random variables $\{Y_1, Y_2, \cdots, Y_N\}$ resulting in a N-dimensional joint distribution also known as superpopulation.

In the model-based approach under geo-statistical modeling, Aubry & Debouzie [13] proposed

$$\hat{V}_{MB}\left(\hat{p}\right) = \frac{1}{S-1}\sum_{s=1}^{S}\left(E_s - \bar{E}\right)^2, \tag{6}$$

where $\bar{E} = \frac{\sum_{s=1}^{S} E_s}{S}$ and $E_s = \hat{p}_s - P_s$.

In this case, $\hat{p}_s$ is the estimated proportion of P_s obtained by using the j^{th} sample for simulating the population $\{z_1, z_2, \cdots, z_N\}$ in the s^{th} realization (iteration).

3. Proposed Estimators

C) Model-assisted approaches

This research proposes a regression estimator and its' correction by using the Geary Index. These estimators arise from the model-assisted approach, and their construction is assisted by simulating the auxiliary variable concentration of the Total Petroleum Hydrocarbons (TPH).

1) Single regression estimator

The variance estimator of the proportion can be written as

$$\hat{V}_{REG\_single}\left(\hat{p}\right) = \left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)S_{\hat{e}}^2, \tag{7}$$

where $S_{\hat{e}}^2 = \frac{\sum_{i}^{n}\left(\hat{e}_i - \bar{\hat{e}}\right)^2}{n-1}$, $\bar{\hat{e}} = \frac{\sum_{i=1}^{n}\hat{e}_i}{n}$, $\hat{e}_i = y_i - \hat{y}_i$

and $\hat{y}_i = \hat{p}_s + \beta_2\left(x_i - \bar{x}_s\right)$ is the predicted y-value for the i^{th} unit using the j^{th} sample, with $\beta_2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(z_i - \bar{z}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$.

Here $\hat{y}_i$ and β_2 are constructed by simulating the auxiliary covariate x_i (TPH concentration), and z_i denotes the corresponding discrete value (*i.e.*, $z_i = 1$ if $x_i \geq 1000$ and $z_i = 0$ otherwise; a unit will be considered contaminated if the TPH concentration is greater or equal to 1000). Then, the auxiliary variable is simulated for the entire population, and the units that occupy the same position in the original sample across domains are selected for analysis.

2) Correction of the regression estimator using the Geary Index

The variance estimator of the proportion using the Geary Index can be written as

$$\hat{V}_{REG\_single\_c_j}\left(\hat{p}\right) = \hat{V}_{REG\_single}\left(\hat{p}\right)c_j, \tag{8}$$

where the terms are as defined previously.

D) Model-assisted approach (averaged)

The estimators (7) and (8) are obtained by using only one realization of the TPH population. Under the model-based approach, the estimator (6) needs a great number of iterations. To compare the accuracy under the same number of iterations, estimators (9) and (10) are introduced to be able to compare with estimator (6).

$$\bar{\hat{V}}_{REG.AV}\left(\hat{p}\right) = \frac{\sum_{s=1}^{S}\hat{V}_{REG}\left(\hat{p}_s\right)}{S}, \tag{9}$$

where $\hat{V}_{REG}\left(\hat{p}_s\right)$ is obtained for the s^{th} iteration.

Correction to (7) using the Geary Index can be written as

$$\bar{\hat{V}}_{REG.AV\_c_j}\left(\hat{p}\right) = \bar{\hat{V}}_{REG.AV}\left(\hat{p}\right)c_j. \tag{10}$$

4. Simulation Study

The estimation process in the model-based approach requires a great number of iterations; at least 10,000 are recommended by Aubry and Debouzie [6]. In contrast to the other approaches, only a single iteration is necessary. Due to this constraint, the simulation study was divided into two parts.

The first part, called one-step simulation, considered 34 cases that were derived from combinations of the following factors.

Geo-statistical model. Two models were selected for generating data:

The wave model and the exponential model [14] [15].

The mean and the variance of the process. Three values that are common in this type of populations were used for the mean (980, 1000, 1020) and two for the variance (300, 600).

Autocorrelation index. Three different levels of autocorrelation indices were employed: low, medium and high (1.5, 2.8 and 4.8), respectively.

Under the exponential model, estimators (3), (4), (5) and (7) were evaluated, while for the wave model, estimator (8) was also considered. Estimator (8) was employed because it provides better results.

In the second part of the study, called the averaging simulation, 6 cases were evaluated. Three cases corresponded to the exponential model, and the remaining ones to the wave model. In the first group, the mean (980) and the variance (300) were kept constant for all autocorrelation levels (1.5, 2.8 and 4.8); while in the second group, they were held at 1020 and 600, respectively.

The averaging simulation was carried out to observe the performance of the model-based estimator (6) and the model-assisted estimators (9) and (10) by using the same number of iterations (1000). These results were compared against the estimators (7) and (8), which were constructed with a single iteration.

4.1. One-Step Simulation

1) For each one of the 34 cases, the TPH $\left(w_1, w_2, \cdots, w_N\right)$ values were generated, and this population was considered as the real population. Next, the correspondent discrete values were assigned: units with TPH contents higher than 1000 ppm had a value 1 and 0 otherwise $\left(y_1, y_2, \cdots, y_N\right)$. The simulation procedure was as described below.

2) Divide the real population into 9 systematic samples.

a) With a single sample, estimate the parameters (mean, variance and scale) for the geo-statistical model.

b) Several candidate models for semi-variogram were tested; the comparison criteria for selecting the best model were the mean square error and the Akaike Information Criteria [16].

c) Using the estimated model, the population of TPH was simulated once, and the estimators (3)-(5), (7) and (8) were computed.

3) Repeat steps (a)-(c) for the 9 systematic samples.

4) Compute the ratio R between averaged estimate variance of the proportion, using the current estimator, and the average variance of the proportion using all the systematic samples of the real population as

$$R = \frac{E\left(\hat{V}\left(\hat{p}\right)\right)}{V\left(\hat{p}\right)} = \frac{\sum_{j=1}^{9} \hat{V}\left(\hat{p}_j\right)}{\sum_{j=1}^{9}\left(\bar{\hat{p}} - \hat{p}_j\right)^2}, \tag{11}$$

where $\hat{V}\left(\hat{p}_j\right)$ is the estimated variance for the j^{th} simulated sample using the current estimator;

$$\bar{\hat{p}} = \frac{\sum_{j=1}^{9} \hat{p}_j}{9}$$

corresponds to the mean of the estimated proportion by using all samples of the real population. The best estimators are those for which $R = 1$. The higher values of R overestimate, and those that are less than 1 underestimate the parameter of the variance of the proportion, respectively.

5) Steps (1)-(5) were repeated 1000 times, then 1000 initial populations of TPH were considered, and the results of these replications were averaged.

4.2. Average Simulation

1) Equal to step (1) of one-step simulation.

2) Equal to steps (1)-(a, b) of one step simulation, but in (1)-(c) instead of simulating one population of TPH 1000 were generated. For each simulated population estimators (3)-(6), (9) and (10) are obtained, and at the end these values were averaged. By using only one of these populations (7) and (8) were calculated, too.

3) Repeat (a)-(c) for the 9 systematic samples.

4) Compute the ratio as (3) of one-step simulation.

5) Steps (1)-(5) were repeated 200 times, then 200 populations of TPH were considered; results of these iterations were averaged.

For performing the simulations an R (2.3.1) [17] script was developed, which uses the Random Fields (1.3.2.8) package.

5. Results and Discussion

In order to determine the best performance, the following criteria were considered: ratio average closer to 1; minimum risk of sub estimating the parameter; minimum mean square error; stability and accuracy through different levels of autocorrelation for each model. Using these criteria, the systematic selection strategy that provided the best estimator was set up as follows. First, those estimators that incurred in serious sub-estimations through the different factors were discarded. Then the accuracy and the minimum mean square error were calculated, respectively, as an essential criterion for deciding on the best estimator.

5.1. One-Step Simulation Analysis

Under the exponential model, the estimators (3), (4) and (7) showed a similar trend (**Figure 2**); they were biased upward, and the bias increased as the autocorrelation increased. The estimator (5) showed the opposite pattern, because it was biased downward. In this context, estimator (7) was the most accurate; also, it was affected by high and low levels of autocorrelation. For example, when the mean of the population was equal to 1000 ppm (the point for determining a unit as polluted or not), this estimator resulted in a slight sub-estimation of the variance, especially in the absence of autocorrelation. The worst accuracy was given by the estimator (3), which overestimated the real variance by 2 to 3.5 folds. The estimator (4) was more precise than (3); it showed moderate overestimations that varied from 1.2 to 2.1 folds.

For the wave model, by reviewing the behavior through the levels of autocorrelation, two groups of estimators were found. In the first one, the trend of estimates made with (3) and (4) increased as the autocorrelation level increased (**Figure 3**). However, the opposite pattern was shown in estimators (5), (7) and (8). Among all of these estimators, the better behavior was produced through estimates of (8); which never under-estimated the true value and was the most accurate estimator. The best results were reached when the mean of the population was equal to 1000 (threshold to determine if a unit is contaminated or not

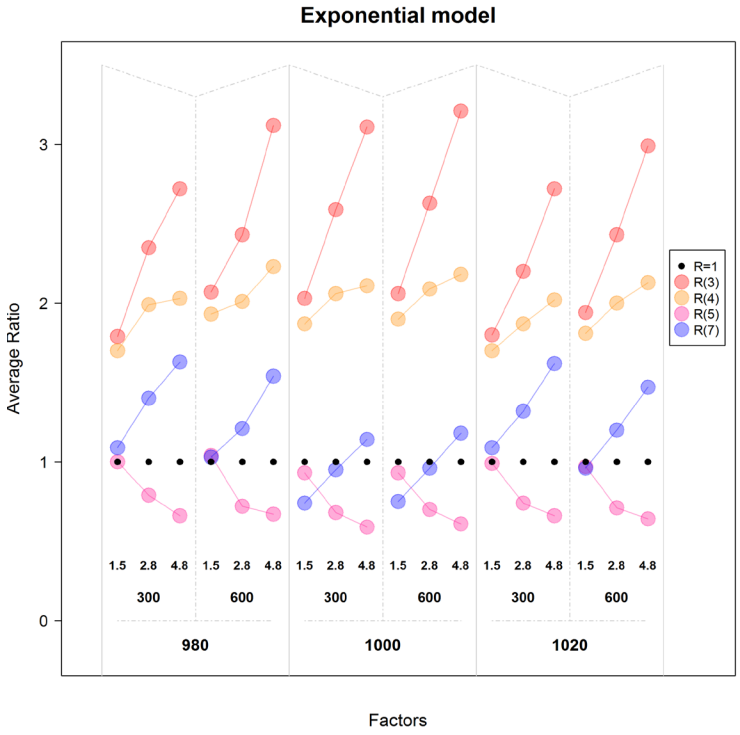

Figure 2. Average ratio (y axis) between estimates performed with estimators (3), (4), (5) and (7) through several factors (x axis) as mean, variance and autocorrelation levels (1.5, 2.8 and 4.8) for the exponential model.

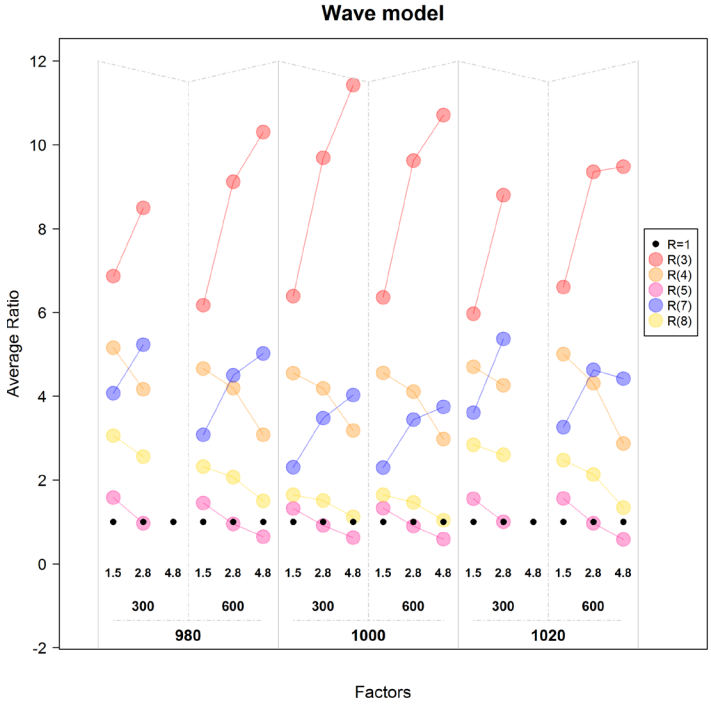

Figure 3. Average ratio (y axis) between estimates for estimators (3), (4), (5), (7) and (8) through several factors (x axis) as mean, variance and autocorrelation levels (1.5, 2.8 and 4.8) for the wave model.

according to [18]). As in the last model, the estimator (3) showed the lowest level of efficiency; its results overestimated the real values by 6.17 to 11.43 folds. The estimates of (5) were the most accurate, but in the presence of highest level of the autocorrelation (4.8) serious sub-estimations were produced. The estimators (4) and (7) presented intermediate results that overestimated the variance by 3.08 to 5.16 folds.

In this simulation study, the estimator (7) showed a periodic and opposite behavior in the accuracy when comparing the exponential and wave models. This behavior is linked to the amount of autocorrelation among sampled units. In the exponential model, the accuracy improved, and the mean square error decreased as the level of autocorrelation increased (**Table 1**).

Under the wave model, different effects occurred: the accuracy decreased and the mean square error remained practically unchanged as the autocorrelation increased (**Table 2**). The inclusion of the estimator (8), which is corrected by the Geary Index, showed an opposite pattern to (7), and a slight reduction of the mean square error in the presence of highly correlated units.

The selection of an estimator must be performed carefully by considering possible implications and risks of each option. For example, in both models the

Table 1. Mean square error of estimates for the exponential model.

			Estimator			
Mean	Variance	Autocorrelation levels	(3)	(4)	(5)	(7)
980	300	1.5	2.32E−08	2.06E−08	2.47E−08	1.56E−08
		2.8	3.25E−08	1.97E−08	1.90E−08	1.09E−08
		4.8	3.86E−08	1.40E−08	1.28E−08	8.48E−09
980	600	1.5	5.83E−08	4.91E−08	5.11E−08	3.84E−08
		2.8	7.41E−08	4.14E−08	3.61E−08	2.08E−08
		4.8	9.31E−08	3.19E−08	2.74−E−8	1.50E−08
1000	300	1.5	1.38E−07	1.11E−07	1.16E−07	1.17E−07
		2.8	1.74E−07	8.50E−08	7.46E−08	5.01E−08
		4.8	2.11E−07	6.28E−08	5.55E−08	2.67E−08
1000	600	1.5	1.39E−07	1.10E−07	1.03E−07	1.04E−07
		2.8	1.80E−07	8.93E−08	7.17E−08	4.88E−08
		4.8	2.20E−07	6.81E−08	5.37E−08	2.77E−08
1020	300	1.5	2.45E−08	2.15E−08	2.59E−08	1.72E−08
		2.8	3.10E−08	1.87E−08	1.81E−08	1.04E−08
		4.8	4.17E−08	1.59E−08	1.35E−08	9.63E−09
1020	600	1.5	5.60E−08	4.72E−08	5.17E−08	3.91E−08
		2.8	7.07E−08	3.91E−08	3.93E−08	2.18E−08
		4.8	9.25E−08	3.28E−08	2.85E−08	1.57E−08

Table 2. Mean square error of estimates for the wave model.

			Estimator				
Mean	Variance	Autocorrelation levels	(3)	(4)	(5)	(7)	(8)
980	300	1.5	7.89E−08	4.02E−08	7.80E−10	2.11E−08	1.31E−08
		2.8	8.48E−08	1.40E−08	1.45E−09	2.31E−08	7.84E−09
		4.8	−−	−−	−−	−−	−−
980	600	1.5	1.57E−07	7.59E−08	5.05E−09	2.26E−08	1.89E−08
		2.8	1.88E−07	2.72E−08	3.10E−09	3.07E−08	1.15E−08
		4.8	1.86E−07	8.76E−09	4.52E−09	2.79E−08	7.29E−09
1000	300	1.5	3.76E−07	1.58E−07	7.64E−09	2.12E−08	2.35E−08
		2.8	4.17E−07	5.23E−08	8.13E−09	3.03E−08	1.70E−08
		4.8	4.00E−07	1.73E−08	1.08E−08	3.04E−08	1.33E−08
1000	600	1.5	3.70E−07	1.56E−07	9.27E−09	2.11E−08	2.43E−08
		2.8	4.25E−07	5.20E−08	6.27E−09	3.09E−08	1.58E−08
		4.8	4.00E−07	1.70E−08	1.15E−08	2.90E−08	1.40E−08
1020	300	1.5	7.05E−07	3.80E−08	2.49E−09	1.75E−08	1.34E−08
		2.8	8.46E−08	1.38E−08	1.46E−09	2.26E−08	7.86E−09
		4.8	−−	−−	−−	−−	−−
1020	600	1.5	1.66E−07	8.17E−08	4.66E−09	2.39E−08	1.90E−08
		2.8	1.82E−07	2.66E−08	2.96E−09	3.02E−08	1.13E−08
		4.8	1.82E−07	8.23E−09	4.28E−09	2.57E−08	7.77E−09

estimator (5) showed a particular behavior: its estimates were the most accurate but in many cases, incurred in serious sub-estimations. The trend of estimator (4) was the opposite. In both models, the bias increased as the autocorrelation increased. The estimator (7) did not show any change in trends through the models. Finally, the estimator (3), which comes from simple random sampling, produced the worse estimations, always over-estimating the variance of the proportion.

5.2. Average Simulation Analysis

In this simulation study, estimator (6) was introduced, which was constructed under the model-based approach. Using this approach, a great number of iterations were necessary to produce reliable results. In this case, 1000 iterations were used. Estimators (9) and (10) were introduced for comparison purposes. First, to compare the estimators' behavior against the model-based estimator under the same number of iterations; and second, to compare the behavior against estimators (7) and (8), which perform the estimation procedure by using only one iteration. Estimators (3), (4) and (5) were also included as references.

Applying the same selection criteria from the one-step simulation to the exponential model, the best performance was shown by estimators (6), (7) and (9). Their estimations, (accuracy and mean square error) were close to each other, respectively. The main difference lies in the fact that the second of them used only 200 iterations in the construction instead of 200 × 1000 = 200,000 iterations for the others (Table 3).

Under the wave model, by adopting a conservative posture, the best estimates were provided through estimators (6), (8) and (10) (see Table 4). Equal to the latter case, the computation of the second of them requires a reduced number of iterations. In general, estimator (4) had the smallest mean squared error values, but as it was pointed out earlier in the case of the one-step simulation, it has serious problems in the presence of high autocorrelation. We must remember that this example is a special case of those showed in Figure 2.

In general, this simulation study shows that for the exponential model estimators (6), (7) and (9) presented similar values for the ratio; however, estimator (7) is preferable because it uses a reduced number of iterations to obtain reliable results. For the wave model, estimator (8) is preferred because it offers the most accurate estimates at the lowest level of computer time.

6. Conclusions

Both simulation studies show promising results that can help improve the accuracy of estimates when performing two-dimensional systematic sampling. The accuracy depends on factors that consider the structure of the population, and takes the relationships among the units in the sample and the use of simulated auxiliary information into account. In general, in the one-step simulation the best results are obtained with estimators constructed under the model-assisted approach and/or taking the presence of autocorrelation into account. When the population presents a structure such as the one produced by the exponential model, estimator (7) is recommended, which shows a periodic behavior for the autocorrelation. In contrast, for the wave model, the best estimator is (8), as the estimates improve as the autocorrelation increases. In the average simulation,

Table 3. Average ratio and mean square error of the estimates for the exponential model.

Estimator	Ratio			MSE		
	1.5	2.8	4.8	1.5	2.8	4.8
(3)	1.73	2.37	2.64	1.68E−08	3.32E−08	3.88E−08
(4)	1.64	2	1.97	1.72E−08	1.97E−08	1.37E−08
(5)	0.98	0.79	0.64	1.24E−08	1.54E−08	1.50E−08
(6)	1.29	1.38	1.28	1.26E−08	9.93E−09	6.75E−09
(7)	1.06	1.43	1.55	1.53E−08	1.08E−08	8.95E−09
(9)	1.06	1.42	1.56	1.93E−08	9.98E−09	7.50E−09

Table 4. Average ratio and mean square error of estimates for the wave model.

Estimator	Ratio			MSE		
	1.5	2.8	4.8	1.5	2.8	4.8
(3)	5.84	9.55	10.94	1.58E−07	1.74E−07	1.78E−07
(4)	4.31	4.43	3.39	7.25E−08	2.48E−08	9.76E−09
(5)	1.3	1	0.69	3.62E−09	2.70E−08	6.08E−09
(6)	3.45	1.85	1.6	6.58E−08	9.53E−09	4.08E−09
(7)	2.9	4.68	5.26	2.20E−08	3.04E−08	2.75E−08
(8)	2.14	2.17	1.72	8.61E−09	5.24E−09	8.86E−09
(9)	2.9	4.67	5.37	2.17E−08	2.79E−08	2.43E−08
(10)	2.14	2.17	1.68	2.38E−08	2.04E−08	1.73E−08

estimators from two different approaches (model-based vs. model-assisted) were compared using the same number of iterations. For the exponential and wave models the best accuracy measures are obtained with estimators (9) and (10), respectively. The estimator (6), which has values close to the estimators mentioned above, seems more robust as a choice of model, but its computation requires a great number of iterations. The estimates of (7) and (8) are as accurate as those obtained with estimators (9) and (10), using only one iteration.

As a result, the regression estimator (7) and its corrected version (8) by the Geary Index are recommended. Within the model-assisted approach, these estimators do not need a great number of iterations in order to achieve estimations as accurate as those obtained with the more complex approaches. Finally, since the systematic sampling method is broadly used in many research areas, this methodology is not limited to this particular problem (petroleum spills). It is easily adaptable to other cases where this sampling design is used, including linear cases.

Acknowledgements

This research was funded by the Mexico's National Council of Science and Technology (Consejo Nacional de Ciencia y Tecnologia, CONACyT).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Cochran, W.G. (1997) Sampling Techniques. 3rd Edition, John Wiley & Sons, New York.

[2] Wolter, K.M. (1985) Introduction to Variance Estimation. Springer-Verlag, New York.

[3] Marcello, D. (2003) Estimating the Variance of the Sample Mean in Two-Dimensional

Systematic Sampling. *Journal of Agriculture*, *Biological and Environmental Statistics*, **8**, 280-295. https://doi.org/10.1198/1085711032174

[4] Li, X. (2006) Application of Nonparametric Regression in Survey Statistics. Ph.D. Thesis, Iowa State University, Ames, Iowa.

[5] Opsomer, J.D., Francisco-Fernández, M. and Li, X. (2012) Model-Based Non-Parametric Variance Estimation for Systematic Sampling. *Scandinavian Journal of Statistics*, **39**, 528-542. https://doi.org/10.1111/j.1467-9469.2011.00773.x

[6] Aubry, P. and Debouzie, D. (2000) Geoestatistical Estimation Variance for the Spatial Mean in Two-Dimensional Systematic Sampling. *Ecology*, **81**, 543-553. https://doi.org/10.1890/0012-9658(2000)081[0543:GEVFTS]2.0.CO;2

[7] Särndal, C.E., Swensson, B. and Wretman, J. (1992) Model Assisted Survey Sampling. Springer, New York.

[8] Strand, G.H. (2017) A Study of Variance Estimation Methods for Systematic Spatial Sampling. *Spatial Statistics*, **21**, 226-240. https://doi.org/10.1016/j.spasta.2017.06.008

[9] Geary, R.C. (1954) The Contiguity Ratio and Statistical Mapping. *The Icorporporated Statistician*, **5**, 115-145. https://doi.org/10.2307/2986645

[10] Lehtonen, R. and Pahkinen, E. (2004) Practical Methods for Design and Analysis of Complex Surveys. 2nd Edition, John Wiley & Sons Ltd, Chichester, 349 p.

[11] Särndal, C.E. (1978) Design-Based and Model Based Inference in Survey Sampling. *Scandinavian Journal of Statistics*, **5**, 27-52.

[12] Thompson, M.E. (1997) Sampling. John Wiley & Sons, New York.

[13] Aubry, P. and Debouzie, D. (2001) Estimation of the Mean from a Two-Dimensional Sample: The Geostatistical Model-Based Approach. *Ecology*, **82**, 1484-1494. https://doi.org/10.1890/0012-9658(2001)082[1484:EOTMFA]2.0.CO;2

[14] Isaaks, E.H. and Srivastava, R.M. (1989) Applied Geostatistics. Oxford University Press.

[15] Chauvet, P. (1993) Processing Data with a Spatial Support: Geostatistics and Its Methods. Cahiers de Géostatistique, Fasc. 4, Centre de Géostatistique, Fontainebleau, France.

[16] Webster, R. and Oliver, M.A. (2001) Geostatistics for Environmental Scientists. John Wiley & Sons Ltd., Chichester.

[17] R Development Core Team (2005) R: A Language and Environment for Statistical Computing, Reference Index Version 2.3.1. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

[18] Diario Oficial de la Federación (2002) Norma Oficial Mexicana de Emergencia NOM-EM-138-ECOL-2002, que establece los límites máximos permisibles de contaminación en suelos afectados por hidrocarburos, la caracterización del sitio y procedimientos para la restauración.