Optimal land cover sampling (draft)

Lauri Häme, March 9, 2023
lauri.hame@terramonitor.com

Abstract

A new method for optimizing the land cover sampling radius is presented in order to maximize the information obtained by land cover sampling. Initial numerical experiments performed on CORINE land cover data suggest that the sampling radius has a significant effect on the land cover characterization. A small sampling radius will result in equal land cover characterizations for all points with the same land cover class and a large sampling radius will render the characterizations indifferent due to the overlap of the sampling areas. The optimal sampling radius is found by studying the spread of land cover distribution vectors in sampling areas of different sizes. In the initial experiments, the optimal sampling radius for CORINE land cover data is found at around 1.3 kilometers for 16085 random points in built-up areas. The proposed method is easily implemented for any data sets and enables the sampling radius to be determined in a well-justified manner in all cases.

Introduction

Land cover is defined as the physical material at the surface of Earth. Different types of land cover *classes* include for example grass, asphalt, trees, bare ground and water. Information on land cover is usually captured using field surveys and analysis of remotely sensed imagery (Cracknell & Reading, 2014).

Land cover sampling is a process where land cover information is extracted from a land cover dataset around predefined geographical sample points. The size of the sampling neighborhood is determined by the *sampling radius*, which has an effect on the resulting land cover characterization of the points. In this article we describe a method for optimally sampling land cover data around the sample points. The general idea is to *maximize the information obtained by land cover sampling*, by adjusting the size of the sampling radius around the sample points. The method is primarily motivated by characterizing the human habitats in medical cohort studies. However, the proposed sampling method can be used in any sampling problem with n-dimensional spatial data and a countable number of classes.

Problem definition

Let us consider a geographical *land cover data set*, in which each point on the ground is associated with exactly one land cover class. More formally, a land cover data set L with m land cover classes is defined as a mapping $L : \mathbb{R}^2 \to \{1, \ldots, m\}$, that is, each point in the 2-dimensional space is associated with a land cover class between 1 and m. For each class $c \in \{1, \ldots, m\}$, the binary *indicator function* is defined as a mapping $I_c : \mathbb{R}^2 \to \{1, 0\}$, where $I_c(x) = 1$ if $L(x) = c$ and $I_c(x) = 0$ otherwise for all $x \in \mathbb{R}^2$. The value of the indicator function $I_c(x)$ for class c at a given point $x \in \mathbb{R}^2$ indicates if the land cover class at point x is c or not.

Next, we want to characterize n geographically defined points based on the land cover around the points. This is performed by calculating the land cover distribution around each point using a *sampling radius* r, which is equal for all points. In other words, for a point $p \in \mathbb{R}^2$ and sampling radius $r > 0$, the *land cover distribution vector* is defined as an m-dimensional vector $v(p, r) = (v_1, \ldots, v_m)$, where v_i is given by the formula

$$v_i = \frac{1}{\pi r^2} \int_{\{x \in \mathbb{R}^2 \ | \ |p-x| \leq r\}} I_i(x). \tag{1}$$

The element v_i of the land cover distribution vector describes the share of land cover class i of the area within radius r from the point p.

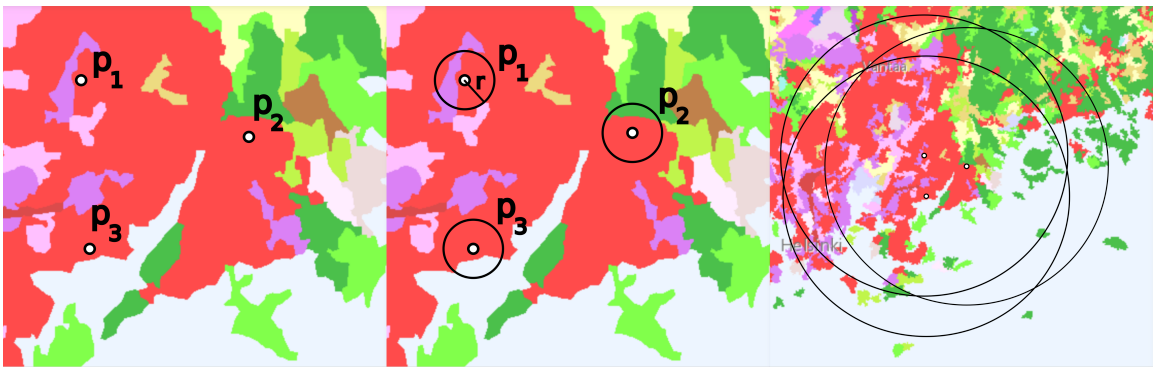

Figure 1: Land cover sampling with three different sampling radii. Figure (a) shows three geographical points p_1, p_2 and p_3 located in the Helsinki area and the CORINE land cover map, where each color represents a specific land cover class. All points are located in the same land cover class (red="Discontinuous urban fabric"). By using a small sampling radius (a), the land cover characterization of all three points is equal. By increasing the sampling radius as depicted in Figure (b), different land cover class distributions are obtained for the areas around the three points, resulting in larger spread in the land cover distribution vectors. Using a very large sampling radius for land cover sampling as in Figure (c), the spread of the land cover distribution vectors will diminish due to the overlap of the sampling areas.

Finally, we define the *spread* of land cover distribution vectors. The *distance* between two land cover distribution vectors v and w is given by the Euclidean norm, $|v - w| = \sqrt{\sum_{i=1}^{m}(v_i - w_i)^2}$. Furthermore, the spread $S(V)$ of a set $V = \{v_1, \ldots, v_n\}$ of vectors is defined as the average of all distances between pairs of vectors in V, that is,

$$S(V) = \frac{1}{n^2} \sum_{v,w \in V} |v - w|. \tag{2}$$

Example 1. Let us consider the spread of the land cover distribution vectors for the three sample points p_1, p_2 and p_3 depicted in Figure 1b. All points are located in the "Discontinuous urban fabric" land cover class (red). Using the sampling radius r shown in the figure, in addition to the red class, the sampling areas contain parts of the "Industrial or commercial units" (purple), "Coniferous forest" (green) and "Sea and ocean" (blue) land cover classes for points p_1, p_2 and p_3, respectively. The land cover distribution vectors obtained using the sampling radius $r = 950$ meters are shown in Table 1. The spread of the vectors is given by calculating the average Euclidean distance between all vector pairs, that is,

$$\begin{aligned} S(V) = \frac{1}{3}\Big(& \sqrt{(0.81-0.86)^2 + (0.19-0)^2 + (0-0.14)^2} \\ & + \sqrt{(0.81-0.77)^2 + (0.19-0)^2 + (0-0.23)^2} \\ & + \sqrt{(0.86-0.77)^2 + (0.14-0)^2 + (0-0.23)^2} \Big) \approx 0.28. \end{aligned}$$

Problem statement: Given a land cover data set $L : \mathbb{R}^2 \to \{1, \ldots, m\}$ and a set of points $P \subset \mathbb{R}^2$, find the optimal sampling radius r which maximizes the spread $S(V_{P,r})$, where $V_{P,r}$ denotes the set of land cover distribution vectors of the points $p \in P$ with sampling radius r.

Sample point	Land cover distribution vector			
	Urban	Industrial	Forest	Sea
p_1	0.81	0.19	0	0
p_2	0.86	0	0.14	0
p_3	0.77	0	0	0.23

Table 1: Land cover distribution vectors for the three sample points presented in Figure 1b.

Solution

The optimal sampling radius for land cover sampling can be easily determined by means of direct search. Given a land cover data set L and a set of points P, the spread of land cover distribution vectors is calculated using different values of the sampling radius r and the optimal value for r is chosen in a way that the spread $S(V_{P,r})$ of land cover distribution vectors is as high as possible. To speed up the search in large data sets, different direct search methods can be applied, such as the Golden section or Fibonacci search methods. Furthermore, to accelerate the computation of the spread function for large values of n, instead of considering all pairs of vectors, the distance between vectors can be calculated for a subset, for example a limited number of randomly chosen pairs of vectors.

Computational experiments

The computational experiments were performed using CORINE 2018 land cover data (Aune-Lundberg & Strand, 2021). The sampling was executed using the QGIS zonal histogram algorithm (QGIS, 2022).

Three points experiment

First, the example case with three points presented in Figure 1 was studied. The optimal sampling radius was determined by calculating the land cover spread for different values of r with 50 meter intervals, see Figure 2. The spread of land cover distribution vectors is maximized with $r = 2050m$. With small values of r, the spread is zero since all three points are covered by the same land cover class. With high values of r the spread is decreased due to overlapping sampling areas.

Large scale experiment

The optimization process was repeated for a data set consisting of randomly chosen points located in Southern Finland between latitudes $[60.62, 62.35]$ and longitudes $[22.11, 27.43]$, see Figure 3. First, 1000000 points were chosen randomly from the area according to a uniform probability distribution. Then, all points in non-built-up areas were removed, resulting in a data set of 16085 points representing hypothetical locations of people's addresses. A 50 meter interval was used for searching the optimal sampling radius.

Computational performance

In the final experiment we studied the effect of limiting the number of vector pairs that are used for calculating the spread in the large scale experiment. The number of randomly chosen vector pairs was constrained by different values of the *depth limit* and standard deviation was calculated over 100000 runs. The sampling radius $r = 1300$ meters was used in the experiment.

The results presented in Figure 5 indicate that the spread of land cover distribution vectors can be computed relatively accurately by considering a relatively small amount of randomly chosen vector pairs. For example, calculating the spread using 3000 randomly chosen vector pairs, the standard deviation is

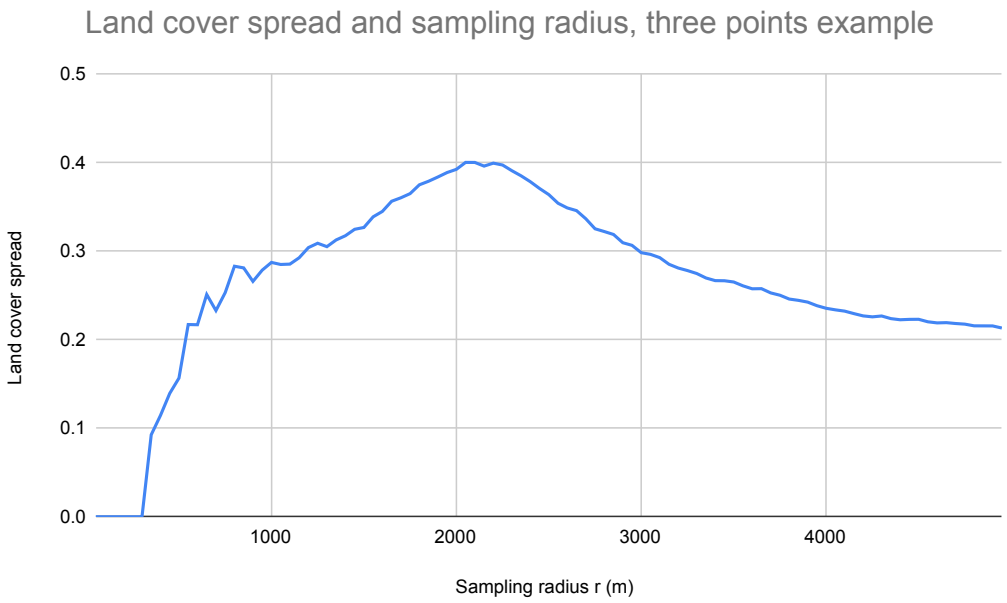

Figure 2: Land cover spread as a function of sampling radius for the three points example in Figure 1. The spread of land cover distribution vectors is maximized with $r = 2050$ meters. With small values of r, the spread is zero since all points are covered by the same land cover class. With high values of r the spread is decreased due to overlapping sampling areas (see Figure 1).

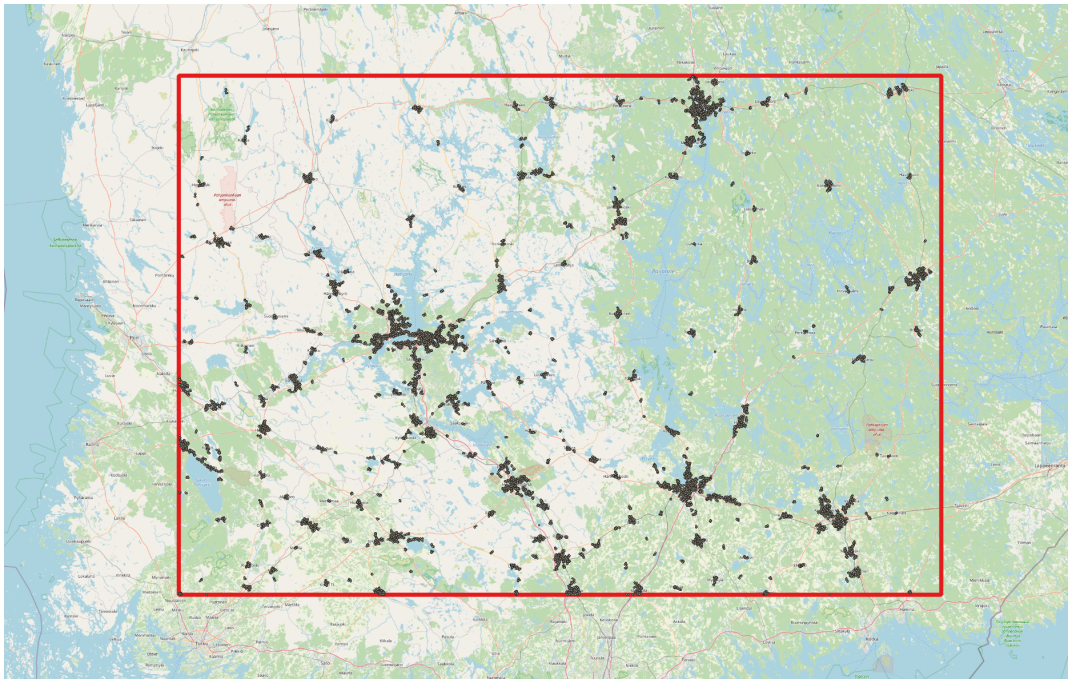

Figure 3: Large scale experiment: 16085 randomly chosen points located in built-up areas in Southern Finland between latitudes $[60.62, 62.35]$ and longitudes $[22.11, 27.43]$.

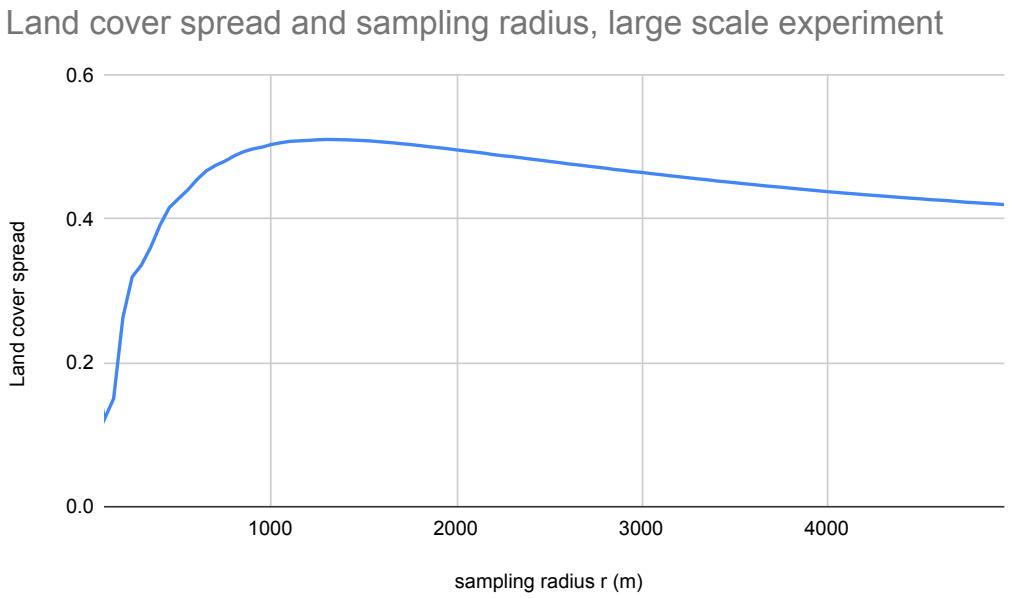

Figure 4: Land cover spread as a function of sampling radius for the large scale experiment. The land cover spread behaves similarly as in the three point experiment. The spread of land cover distribution vectors is maximized with $r = 1300$ meters.

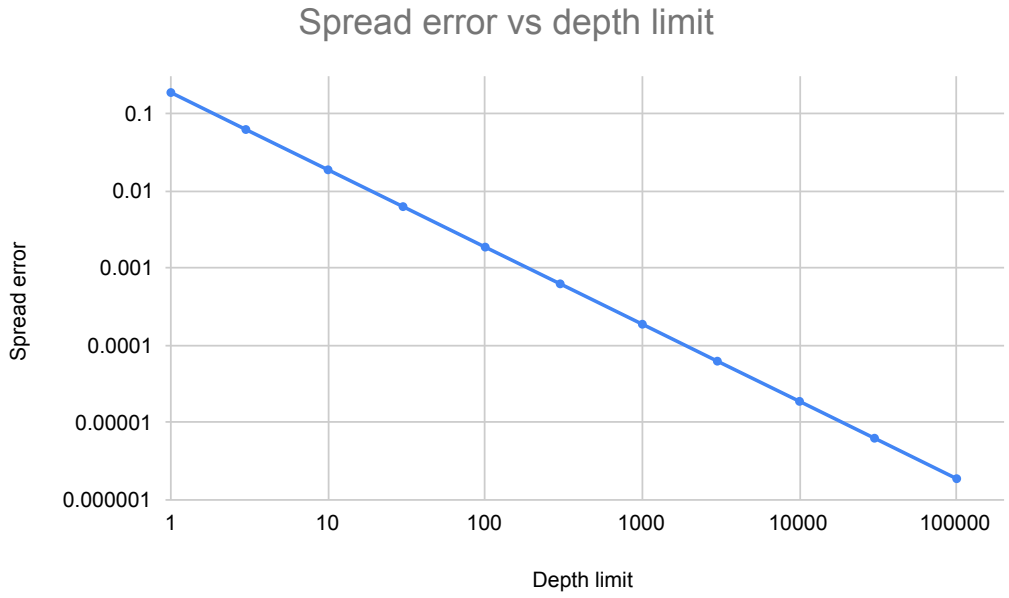

Figure 5: The spread of land cover distribution vectors was calculated using subsets of vector pairs, constrained by different values of the *depth limit*. The spread of land cover distribution vectors can be computed relatively accurately by considering a small amount of randomly chosen vector pairs. For example, calculating the spread using 3000 randomly chosen vector pairs, the standard deviation is less than 0.0001 in 100000 runs.

less than 0.0001. This observation suggests that the calculation of spread can be performed accurately for large data sets with reasonable computational effort.

Conclusions

A new method for optimizing the land cover sampling radius is presented in order to maximize the information obtained by land cover sampling. Initial numerical experiments performed on CORINE land cover data and random locations in built-up areas suggest that the sampling radius has a significant effect on the land cover characterization of the locations. As expected, a small sampling radius will result in equal land cover characterizations for all points with the same land cover class and a large sampling radius will render the land cover characterizations indifferent due to the overlap of the sampling areas. The optimal sampling radius is found by studying the spread of land cover distribution vectors in sampling areas of different sizes.

In the initial experiments, the optimal sampling radius for CORINE land cover data is found at around 1.3 kilometers for random points in built-up areas. However, both the land cover data as well as the sampling point data set have an effect on the size of the optimal sampling area. The method proposed in this paper is easily implemented for any data sets and enables the sampling radius to be determined in a well-justified manner in all cases.

Directions of future work include analytical and computational studies of the optimal sampling radius in different theoretical and real-life scenarios, as well as computational performance optimization of the search algorithm.

Bibliography

Matthew J. Cracknell, Anya M. Reading, *Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information*, Computers & Geosciences, Volume 63, 2014, Pages 22-33, ISSN 0098-3004, https://doi.org/10.1016/j.cageo.2013.10.008

Linda Aune-Lundberg, Geir-Harald Strand, *The content and accuracy of the CORINE Land Cover dataset for Norway*, International Journal of Applied Earth Observation and Geoinformation, Volume 96, 2021, 102266, ISSN 1569-8432, https://doi.org/10.1016/j.jag.2020.102266

QGIS, *QGIS zonal histogram algorithm*, QGIS documentation, 2022, https://docs.qgis.org/3.22/en/docs/user_manual/processing_algs/qgis/rasteranalysis.html