Topological Generalization of Continuous Valued Raster Data

Padraig Corcoran
Cardiff University, Wales, UK
corcoranp@cardiff.ac.uk

ABSTRACT

We propose a novel method for generalizing continuous valued raster data with respect to topological constraints whereby smaller scale connected components and holes in the data sublevel sets are removed. The proposed method formulates the problem of generalization as an optimization problem with respect to persistent homology. We prove the objective function to be locally continuous with analytical gradients which can be used to perform optimization using gradient descent. Furthermore, we prove the convergence of gradient descent to a global optimal solution. The proposed method is general in nature and can be applied to raster data of any dimension. The utility of the method is demonstrated with respect to generalizing two- and three-dimensional raster data corresponding to digital elevation models (DEM) and subsurface mineral interpolation respectively.

CCS CONCEPTS

• **Mathematics of computing → Algebraic topology**.

KEYWORDS

Generalization; Topology; Raster

ACM Reference Format:
Padraig Corcoran. 2019. Topological Generalization of Continuous Valued Raster Data. In *27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '19), November 5–8, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3347146.3359071

1 INTRODUCTION

The raster data model is fundamental approach to modelling and storing geographical data. In this model the data in question is modelled using an array or grid of data points where these data points can be categorical or real/continuous values. For example, digital elevation models (DEM) are commonly modelling using a raster model where the data points are continuous values equalling height. On the other hand, land-use/land-cover classifications are commonly modelling using a raster model where the data points are categorical values equalling land-use/land-cover class.

Given a raster data in many cases it is useful to perform some form of generalization whereby the information or detail of the data is reduced [3]. For example, generalization is commonly performed prior to rendering a visualization in order to reduce computational complexity. Also, generalization can be used to produce a representation which is tailored to a user's requirements whereby irrelevant information is removed and important information is enhanced [4].

Generalization methods may be categorized in terms of the constraints they perform generalization with respect to [3]. Generalizing categorical valued raster data with respect to topological constraints is a well studied problem for which there exists many solutions. On the other hand, generalizing continuous valued raster data with respect to such constraints is less well studied and represents an open research problem. In this article we propose a novel solution to this problem which employs the theory of *persistent homology*. Persistent homology measures the topological features of a given raster data in terms of the persistence of connected components and holes of various dimensions in the data sublevel sets. In the context of a two-dimensional raster data, connected components and holes in the sublevel sets correspond to valleys and peaks respectively in the data. For example, consider the two-dimensional raster data displayed in Figure 1(a) where a corresponding one-dimensional cross section is displayed in Figure 1(b). This data contains a smaller and a larger scale valley which correspond to connected components in the sublevel sets of smaller and larger persistence respectively. This data also contains a smaller scale peak which corresponds to a hole in the sublevel sets of smaller persistence.

The proposed generalization method poses the generalization problem as an optimization problem with respect to persistent homology. The solution to this problem equals a generalization where connected components and holes in the data sublevel sets of lesser persistence have been removed. To illustrate the proposed method consider again the two-dimensional raster data displayed in Figure 1(a). A generalizing of this data is displayed in Figure 1(c) were a corresponding one-dimensional cross section is displayed in Figure 1(d). Here the connected component and hole in the data sublevel sets of lesser persistence have been removed. Although the above example considers two-dimensional raster data, the proposed method is applicable to raster data of any dimension.

The layout of this paper is as follows. Section 2 describes the proposed generalization method. Section 3 presents an evaluation of the method. Finally, section 4 draws some conclusions.

2 TOPOLOGICAL GENERALIZATION

The proposed generalization method contains the following four steps. In the first step the raster data to be generalized is modelled using a combinatorial data structure known as a *filtration*. In the second step this filtration is used to compute the persistent homology of the raster data. In the third step the problem of generalization is posed as a problem of optimizing an objective function defined in terms of persistent homology. In the final step, the raster data is optimized with respect to this objective function using gradient descent. These steps are now described in the following subsections.

SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6909-1/19/11.
https://doi.org/10.1145/3347146.3359071

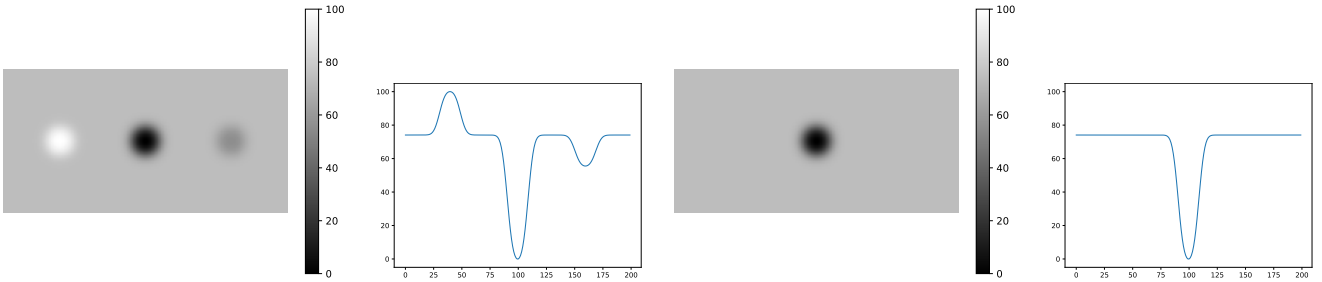

Figure 1: A two-dimensional raster data is displayed in (a) where a corresponding one-dimensional cross section is displayed in (b). This data contains a smaller and a larger scale valley which correspond to connected components of smaller and larger persistence respectively. This data also contains a smaller scale peak which corresponds to a hole of smaller persistence. A generalization of this data is displayed in (c) were a corresponding one-dimensional cross section is displayed in (d).

2.1 Filtration

An (abstract) simplicial complex $\mathcal{K}$ is a finite collection of sets such that for each $\sigma \in \mathcal{K}$ all subsets of σ are also contained in $\mathcal{K}$. Each element $\sigma \in \mathcal{K}$ is called a p-simplex where $p = |\sigma| - 1$ is the corresponding dimension of the simplex. The faces of a simplex σ correspond to all simplices τ where $\tau \subset \sigma$. The dimension of a simplicial complex $\mathcal{K}$ is the largest dimension of any simplex $\sigma \in \mathcal{K}$. A triangulation of a topological space corresponds to a simplicial complex representation of that space. In this work we triangulate the raster data to be generalized using a *Freudenthal triangulation* where each raster cell is represented using a 0-simplex.

Given a simplicial complex $\mathcal{K}$ containing m simplices, consider a function $f : \mathcal{K} \rightarrow \mathbb{R}$ such that $f(\tau) \leq f(\sigma)$ whenever τ is a face of σ. For all $a \in \mathbb{R}$, the sublevel set $\mathcal{K}(a) = f^{-1}(-\infty, a]$ is a subcomplex of $\mathcal{K}$. The ordering of the simplices of $\mathcal{K}$ with respect to the corresponding values of f induces an ordering of the sublevel sets defined in Equation 1 known as a *filtration*. Note that, any two successive sublevel sets $\mathcal{K}_i$ and $\mathcal{K}_{i+1}$ in this sequence differ by only a single simplex. If the function f maps multiple simplices to the same real value, we order them by dimension and if they have equal dimension we order them arbitrary.

$$\emptyset = \mathcal{K}_0 \subset \cdots \subset \mathcal{K}_{m-1} \subset \mathcal{K}_m = \mathcal{K} \tag{1}$$

In this work we represent the data to be generalized as a filtration using the following approach. Let $\mathcal{K}$ be the simplicial complex corresponding to the Freudenthal triangulation of the raster data in question. Let f be the function which maps each 0-simplex in $\mathcal{K}$ to the value of corresponding raster cell. We extend the domain of f to all simplices in $\mathcal{K}$ using Equation 2 which in turn induces a filtration on $\mathcal{K}$ known as a *lower star filtration*.

$$f(\sigma) = \max(f(\gamma) : \gamma \in \sigma, |\gamma| = 0) \tag{2}$$

2.2 Persistent Homology

Let $H_p(\mathcal{K})$ denote the p-homology group of $\mathcal{K}$ [2]. Intuitively an element of the p-homology group corresponds to a p-dimensional hole in $\mathcal{K}$. That is, an element of the 0-homology group corresponds to a path-connected component in $\mathcal{K}$ while an element of the 1-homology group corresponds to a one-dimensional hole in $\mathcal{K}$. Given a filtration of a simplicial complex $\mathcal{K}$, for every $i \leq j$ there exists an inclusion map from $\mathcal{K}_i$ to $\mathcal{K}_j$ and in turn an induced homomorphism from $H_p(\mathcal{K}_i)$ to $H_p(\mathcal{K}_j)$ for each dimension p. An element of the p-homology group is *born* at $\mathcal{K}_{i+1}$ if it exists in $H_p(\mathcal{K}_{i+1})$ but does not exist in $H_p(\mathcal{K}_i)$. An element of the p-homology group *dies* at $\mathcal{K}_{i+1}$ if it exists in $H_p(\mathcal{K}_i)$ but does not exist in $H_p(\mathcal{K}_{i+1})$. If an element of a p-homology group never dies, its death is determined to be at a hypothetical simplicial complex $\mathcal{K}_\infty$.

Let P denote the space $\{(b, d) \in \mathbb{R}^2, b \leq d\}$. An element of the p-homology group which is born at $\mathcal{K}_i$ and dies at $\mathcal{K}_j$ can be represented as a point $(f(\alpha), f(\beta)) \in P$ where α and β are the single simplices added to the filtration at $\mathcal{K}_i$ and $\mathcal{K}_j$ respectively. The value $f(\beta) - f(\alpha)$ is known as the *persistence* of the element in question. The multiset of k points $\{(f(\alpha_i), f(\beta_i)) \in P, i = 1 \ldots k\}$ corresponding to the p-homology group is called the p-dimensional *persistence diagram*. For a given simplicial complex $\mathcal{K}$, let Pers_p be the map defined in Equation 3 from the space of functions defined on $\mathcal{K}$ to the space of p-dimensional persistence diagrams. The persistence diagrams $\text{Pers}_0(.)$ and $\text{Pers}_1(.)$ corresponding to the raster data in Figure 1(a) are displayed in Figures 2(a) and 2(b) respectively.

$$\text{Pers}_p(f) = \{(f(\alpha_i), f(\beta_i)) \in P, i = 1 \ldots k\} \tag{3}$$

2.3 Objective Function

In this section we pose the problem of generalization as an optimization problem where the objective function is defined in terms of persistent homology. Let $\text{Pers}_p(f) = \{(f(\alpha_1), f(\beta_1)), \ldots, (f(\alpha_k), f(\beta_k))\}$ be the p-dimensional persistence diagram corresponding to the raster data to be generalized. The real valued objective function T to be minimized is defined in Equation 4 where τ is a real valued function defined in Equation 5. The function τ has a single hyper-parameter a.

$$T(\text{Pers}_p(f)) = \sum_{i=1}^{k} \tau(f(\beta_i) - f(\alpha_i)) \tag{4}$$

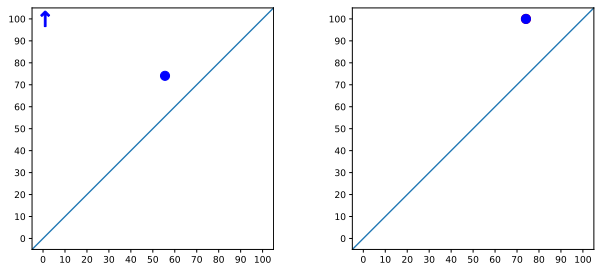

Figure 2: Persistence diagrams $\text{Pers}_0(.)$ and $\text{Pers}_1(.)$ corresponding to Figure 1(a) are displayed in (a) and (b) respectively. Elements of $\text{Pers}_0(.)$ and $\text{Pers}_1(.)$ are represented using blue circles. Elements of $\text{Pers}_0(.)$ that do not die are represented by arrows.

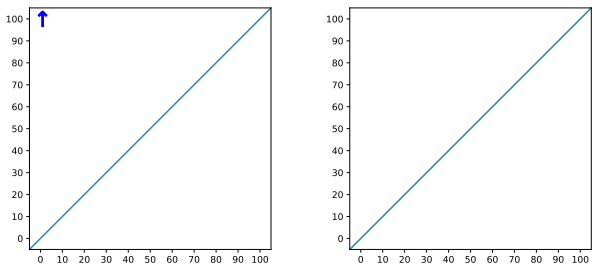

Figure 3: Persistence diagrams $\text{Pers}_0(.)$ and $\text{Pers}_1(.)$ corresponding to Figure 1(c) are displayed in (a) and (b) respectively. Elements of $\text{Pers}_0(.)$ and $\text{Pers}_1(.)$ are represented using blue circles. Elements of $\text{Pers}_0(.)$ that do not die are represented by arrows.

$$\tau(t) = \begin{cases} t & t \leq a \\ 0 & t > a \end{cases} \tag{5}$$

The objective function T can be interpreted as follows. For each point $(f(\alpha_i), f(\beta_i))$ if the corresponding persistence $f(\beta_i) - f(\alpha_i)$ is less than or equal to the hyper-parameter a, a value of $f(\beta_i) - f(\alpha_i)$ will be added to the objective function. Otherwise, a value of 0 will be added to the objective function. This has the effect of penalizing topological features with persistence less than or equal to a while not penalizing those topological features with persistence greater than a.

A global optimal solution to the objective function will correspond to a function f such that $\text{Pers}_p(f)$ does not contain any points with persistence less than or equal to a. Such a solution will have an objective function value of 0. To illustrate this consider again the two-dimensional raster data displayed in Figure 1(a) where the corresponding persistence diagrams are displayed in Figure 2. A generalizing of this data corresponding to a global optimal solution with a hyper-parameter a of value 30 is displayed in Figure 1(c). The corresponding persistence diagrams are displayed in Figure 3. The smaller scale peak and the smaller scale valley in the original data have persistent less than or equal to 30 and therefore do not exist in the generalized result.

2.4 Optimization

In this section we present an optimization algorithm for minimizing the objective function T defined in Equation 4. The algorithm in question is *gradient descent* which uses gradient information to iteratively minimize the objective function. Toward this goal we derive the analytic gradients. Analytic gradients are necessary for performing gradient descent which attempts to minimize the objective function by iteratively taking a step in the direction of negative gradient.

It can be proven that the objective function T is *locally continuous* (this proof is omitted due to page constraints). Given this we derive the analytic gradient of T with respect to the function f evaluated at the 0-simplices of $\mathcal{K}$. Note that, the function f evaluated at the 0-simplices of $\mathcal{K}$ corresponds to the raster cell values. For a given point $(f(\alpha_i), f(\beta_i))$ in $\text{Pers}_p(f)$, application of the sum and chain rules gives the corresponding partial derivatives of T which are defined in Equation 6. Note that, a subgradient is selected at the point a.

$$\begin{aligned} \frac{\partial T}{\partial f(\alpha_i)} &= \begin{cases} -1 & f(\beta_i) - f(\alpha_i) \leq a \\ 0 & f(\beta_i) - f(\alpha_i) > a \end{cases} \\ \frac{\partial T}{\partial f(\beta_i)} &= \begin{cases} 1 & f(\beta_i) - f(\alpha_i) \leq a \\ 0 & f(\beta_i) - f(\alpha_i) > a \end{cases} \end{aligned} \tag{6}$$

Let $\{\gamma_1, \ldots, \gamma_m\}$ be the set of 0-simplices in the Freudenthal triangulation $\mathcal{K}$ of the raster data to be generalized. By applying the chain we extend the domain of the partial derivatives defined in Equation 6 from $\{f(\alpha_1), f(\beta_1), \ldots, f(\alpha_k), f(\beta_k)\}$ to $\{f(\gamma_1), \ldots, f(\gamma_m)\}$ using Equation 7. We subsequently use these partial derivatives to form the gradient vector defined in Equation 8. Given this gradient vector we minimize the objective function $T(\text{Pers}_p(f))$ using the gradient descent algorithm defined in Algorithm 1. In each iteration of this algorithm we compute $\text{Pers}_p(f)$ using the method described in Section 2.2. In line 6, ϵ is an algorithm hyper-parameter corresponding to step size. In all experiments presented in the results section of this paper we used an ϵ value of 0.001.

$$\frac{\partial T}{\partial f(\gamma_i)} = \begin{cases} \frac{\partial T}{\partial f(\alpha_j)} & \gamma_i = \underset{\rho \in \alpha_j}{\arg\max} f(\rho) \\ \frac{\partial T}{\partial f(\beta_j)} & \gamma_i = \underset{\rho \in \beta_j}{\arg\max} f(\rho) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

$$\nabla T = \left(\frac{\partial T}{\partial f(\gamma_1)}, \ldots, \frac{\partial T}{\partial f(\gamma_m)} \right) \tag{8}$$

It can be proven that the gradient descent algorithm defined in Algorithm 1 converges to a global optimal solution where the objective function evaluates to 0 (this proof is omitted due to page constraints). This solution corresponds to a generalization where all topological features with persistence less than or equal to the hyper-parameter a have been removed.

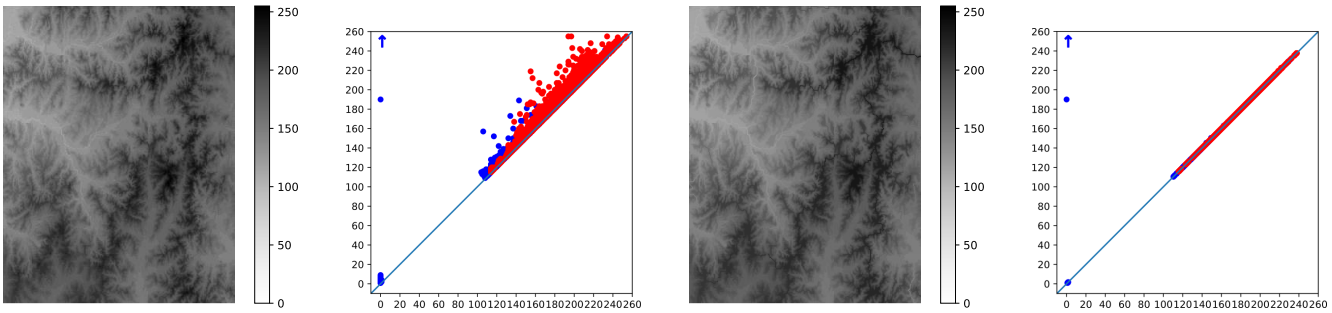

Figure 4: Persistence diagrams corresponding to the DEM in (a) are displayed in (b). The result of generalizing this DEM using a value of 100 for the hyper-parameter a is displayed in (c). The persistence diagrams corresponding to this generalization are displayed in (d). Elements of $\text{Pers}_0(.)$ and $\text{Pers}_1(.)$ are represented by blue and red dots respectively. Elements of $\text{Pers}_0(.)$ that do not die are represented by arrows.

Algorithm 1: Gradient descent optimization

Input: A Freudenthal triangulation $\mathcal{K}$ of the raster data to be generalized. A function $f : \mathcal{K}^0 \rightarrow \mathbb{R}$ where $\mathcal{K}^0$ is the set of 0-simplices in $\mathcal{K}$.

Output: A function $f : \mathcal{K}^0 \rightarrow \mathbb{R}$ such that $T(\text{Pers}_p(f)) = 0$.

```
1  begin
2      prev_objective = ∞
3      current_objective = T(Pers_p(f))
4      while prev_objective - current_objective > 0 do
5          compute ∇T
6          f = f − ε∇T
7          prev_objective = current_objective
8          current_objective = T(Pers_p(f))
9      end
10     return f
11 end
```

3 RESULTS

This section presents an evaluation of the proposed generalization method and is structured as follows. Section 3.1 describes the raster used within the evaluation. Section 3.2 presents an analysis of the convergence properties of the generalization method.

3.1 Raster Data

The raster data used in this evaluation consists of two- and three-dimensional raster data which we now describe. The two-dimensional raster data corresponds to digital elevation models (DEM). We obtained ten DEMs from the National Elevation Dataset provided by the U.S. Geological Survey Science Data Catalog. One of these DEMs corresponding to a region in the Trace State Park Mississippi is displayed in Figure 4(a). The three-dimensional raster data corresponds to a subsurface interpolation of copper mineral percentage. We constructed this raster data using the Brenda Mine dataset obtained from [1].

3.2 Convergence Analysis

To demonstrate convergence with respect to a two-dimensional DEM consider the DEM displayed in Figure 4(a). The corresponding zero-dimensional persistence diagram $\text{Pers}_0(.)$ and one-dimensional persistence diagram $\text{Pers}_1(.)$ are displayed in Figure 4(b). The generalization of this DEM achieved using a value of 100 for the hyper-parameter a is displayed in Figure 4(c) where the corresponding persistence diagrams are displayed in Figure 4(d). The proposed generalization method uses a step size of ϵ (line 6 of Algorithm 1). A consequence of this is that the method cannot reduce the persistence of a point to exactly 0 but instead can only reduce it to a value within ϵ of 0. As mentioned in section 2.4, in our analysis we used an ϵ value of 0.001. It is evident from the persistence diagrams in Figure 4(d) that the generalization method has reduced the persistence of all points with persistence less than or equal to 100 to a value within ϵ of 0. That is, all the points in question now lie along the diagonal of the figure. When applied to the three-dimensional subsurface interpolation, the proposed generalization method was found to also converge correctly.

4 CONCLUSIONS

Generalization of continuous valued raster data with respect to topological constraints represents an open research problem. We proposed a novel solution which poses the problem as an optimization problem with respect to persistent homology. We proved that the optimal solution to this problem can be computed efficiently and corresponds to a generalization where connected components and holes in the sublevel sets of lesser persistence have been removed.

REFERENCES

[1] Isobel Clark. 1979. *Practical geostatistics*. Vol. 3.
[2] Padraig Corcoran and Christopher B Jones. 2017. Modelling topological features of swarm behaviour in space and time with persistence landscapes. *IEEE Access* 5 (2017), 18534–18544.
[3] Padraig Corcoran, Peter Mooney, and Adam Winstanley. 2011. Planar and non-planar topologically consistent vector map simplification. *International Journal of Geographical Information Science* 25, 10 (2011), 1659–1680.
[4] Peter Mooney and Padraig Corcoran. 2012. Using OSM for LBS–an analysis of changes to attributes of spatial objects. In *Advances in Location-Based Services*. Springer, 165–179.