Computational Statistics and Data Analysis 117 (2018) 138–153

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Spatial data compression via adaptive dispersion clustering

Yuliya Marchetti [a,*], Hai Nguyen [a], Amy Braverman [a], Noel Cressie [b,a]

[a] *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA*
[b] *University of Wollongong, Wollongong, Australia*

ARTICLE INFO

Article history:
Received 21 September 2016
Received in revised form 8 August 2017
Accepted 8 August 2017
Available online 26 August 2017

Keywords:
Spatial data compression
Spectral clustering
Spatial clusters
Spatial dispersion function

ABSTRACT

Adaptive Spatial Dispersion Clustering (ASDC), a new method of spatial data compression, is specifically designed to reduce the size of a spatial dataset in order to facilitate subsequent spatial prediction. Unlike traditional data and image compression methods, the goal of ASDC is to create a new dataset that will be used as input into spatial-prediction methods, such as traditional kriging or Fixed Rank Kriging, where using the full dataset may be computationally infeasible. ASDC can be classified as a lossy compression method and is based on spectral clustering. It aims to produce contiguous spatial clusters and to preserve the spatial-correlation structure of the data so that the loss of predictive information is minimal. An extensive simulation study demonstrates the predictive performance of these adaptively compressed datasets for several scenarios. ASDC is compared to two other data-reduction schemes, one using local neighborhoods and one using simple binning. An application to remotely sensed sea-surface-temperature data is also presented, and computational costs are discussed.

Published by Elsevier B.V.

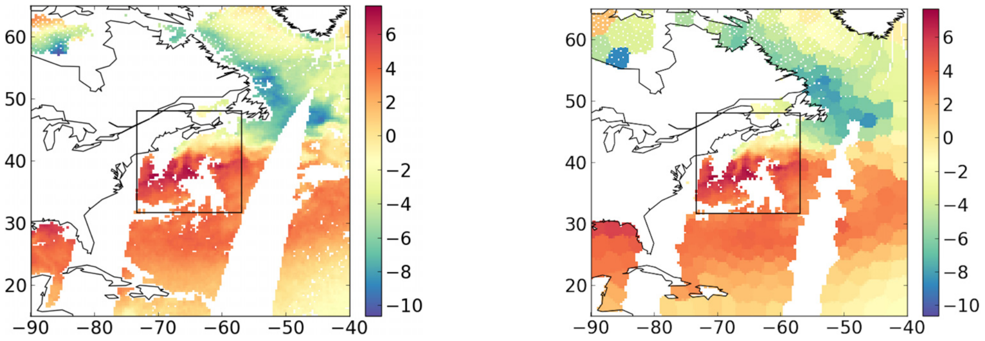
1. Introduction

Very large spatial and spatio-temporal datasets are becoming more commonplace in social, commercial, and scientific research. In the social and commercial realms, this is largely due to the expansion of the internet and the computerization of many aspects of daily life. In science, new technologies for data collection and experimentation have led to the demand for new analysis methods specifically designed for new data types. One area where this is especially true is Earth Science, where satellite remote sensing data play an increasingly important role in understanding the physics of the Earth system and interactions among its components. Remote sensing data can be massive, with hundreds of millions to billions of data points collected per day, but at the same time their spatio-temporal coverage can be sparse, with gaps in coverage due to orbit patterns and observing-technology limitations.

Spatial and spatio-temporal statistical inferences are key to obtaining maximum scientific return from these data, but massiveness poses a serious challenge to conventional spatial-statistical modeling approaches. It is natural to look for ways to make the computations more efficient, and various methods based on simplification of the statistical model have been proposed. Some enforce sparsity on large spatial covariance matrices (Furrer et al., 2006; Kaufman et al., 2008) or precision matrices (Besag and Kooperberg, 1995; Rue and Held, 2005; Lindgren et al., 2011; Eidsvik et al., 2014; Datta et al., 2016; Gramacy and Apley, 2015; Nychka et al., 2015), and others use dimension reduction to reduce the number of parameters required to specify covariance (Banerjee et al., 2008; Cressie and Johannesson, 2008; Finley et al., 2009; Nguyen et al., 2012; Sang and Huang, 2012). However, with ever-increasing data-collection capabilities, the majority of these methods by themselves may not be enough because they still require holding large matrices, such as basis-function matrices, in memory.

[*] Correspondence to: Jet Propulsion Laboratory, MS 158-242, 4800 Oak Grove Drive Pasadena, CA 91109-8099, USA.
E-mail address: yuliya.marchetti@jpl.nasa.gov (Y. Marchetti).

http://dx.doi.org/10.1016/j.csda.2017.08.004
0167-9473/Published by Elsevier B.V.



(a) Sea-surface temperatures.

(b) Compressed data.

Fig. 1. Sea-surface temperature from the Advanced Microwave Scanning Radiometer 2 (AMSR-2) instrument on Global Change Observation Mission — Water (GCOM-W) satellite with about 117,000 data points (a) and the corresponding compressed data to 400 data points (b). The black square highlights the region of interest.

The methodology presented in this article, Adaptive Spatial Dispersion Clustering (ASDC), takes a different approach that is intended to complement dimension-reduction and sparse methods: Our idea is to make the data size smaller in a way that preserves the essential information required for good spatial prediction.

When a spatial dataset is massive, such as is the case for high-resolution global remote sensing data, spatial prediction could be performed by limiting the data to a small region of interest, and ignoring the rest. In fact, local kriging and similar methods rely on such an approach (Haas, 1990; Cressie, 1993; Hammerling et al., 2012). An alternative is to use compressed data instead, where compression here means that data outside the region of interest have been aggregated to coarse resolutions. This approach could be advantageous if aggregation is done in a way that preserves spatial information and produces globally valid spatial covariance structures. "Gridding" or "binning", in which the entire spatial field is aggregated to a coarse spatial resolution, is a form of naive data reduction that does not explicitly address the preservation of spatial covariance.

Clustering is a basic tool of data compression, but spatial dependence is usually not incorporated directly into the "fidelity to the data" part of the clustering criterion. In the case of image compression or segmentation, there are approaches that do account for spatial dependencies and cluster coherence (Hu and Sung, 2006; Craddock et al., 2012), but the goal is to recreate an approximation that is visually indistinguishable from the original image, rather than to preserve spatial contiguity and spatial-dependence structure for purposes of inference *per se*.

In various applications to geospatial data, such as from geological bodies, earthquakes, or climate systems, clustering has been performed to primarily determine spatially contiguous regions, where homogeneous processes are observed. Spatial covariance structure is then incorporated to force spatial contiguity for clusters. The main approaches for clustering geospatial data include using spatial coordinates as additional features, weighting feature dissimilarities by their spatial covariances (Oliver and Webster, 1989; Bourgault et al., 1992), constraining clustering using spatial tessellation (Romary et al., 2015; Heaton et al., 2017); and introducing covariance structure through model-based clustering (Ambroise et al., 1997; Allard and Guillot, 2000; Guillot et al., 2006; François et al., 2006). See Fouedjio (2016b) for a detailed review of such methods.

A very recent approach of Fouedjio (2016b, a) proposes clustering of data locations based on a spatially informed dissimilarity measure, specifically a non-parametric, user-defined kernel estimator of a multivariate cross-variogram function. Fouedjio (2016b) incorporates such a dissimilarity measure into hierarchical clustering to obtain assignment of locations to spatially contiguous groups. Fouedjio (2016a) further defines a similarity measure and uses spectral clustering for grouping of locations, although his work focuses on multivariate data for partitioning data locations into meaningful spatially contiguous clusters and determining an optimal number of clusters. In contrast to ours, his cross-variogram-based similarity measure is of a fixed form and does not allow the size of the clusters to be determined adaptively. See Von Luxburg (2007) for a summary of spectral clustering and a discussion of its relationship to spectral graph theory and graph partitioning problems.

Our method is used explicitly for data compression and incorporates key aspects of spatial covariances through the use of a spatial dispersion function (Sampson and Guttorp, 1992). A demonstration is illustrated in Fig. 1. For spatial predictions in a region of interest, data outside the region of interest are compressed by assigning geographic locations associated with them to spatial clusters. Each spatial cluster is represented by the mean value of the data throughout the cluster. Cluster assignments are obtained by applying spectral clustering (Ng et al., 2002) to a weighted similarity matrix that accounts for covariances among locations outside the region of interest with each other, and covariances between locations inside and outside of the regions of interest. This forces spatial contiguity, and it causes clusters near the region of interest to be smaller than those far away because spatial covariance generally decreases with distance.

Download English Version:

https://daneshyari.com/en/article/4949181

Download Persian Version:

https://daneshyari.com/article/4949181

Daneshyari.com