Spatial Data Compression via Adaptive Spatial Dispersion Clustering

Y. Marchetti[1] H. Nguyen[1] A. Braverman[1] N. Cressie[2]

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

[2]University of Wollongong, Australia

Joint Statistical Meetings, 2016

Outline

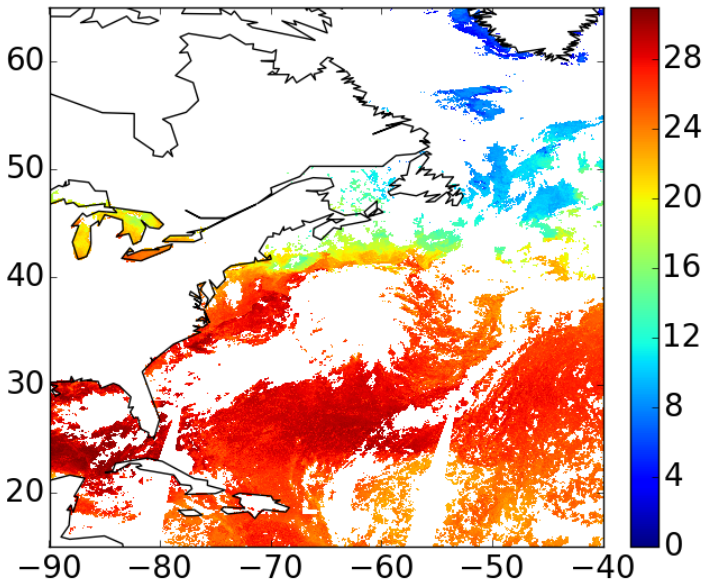
- Motivation
- Spatial data compression
- Adaptive spatial dispersion clustering
- Distance dispersion functions, dispersion matrices, and spectral clustering
- Sea-surface temperature example
- Comparison against other common data reduction methods
- Discussion

Motivation

Remote sensing data are a vast, largely untapped resource.

- Broad range of applications from climate science to fisheries to disaster management
- Massive but sparse
- Challenging for spatial estimation

Current spatial modeling approaches mainly focus on dimension reduction in the model.



Sea-surface temperature from the MODIS-Aqua instrument, daytime on 07/28/2015. Data are 1km resolution. There are about 2.7 million data points in this region.

Motivation

Existing methodologies might not be sufficient (e.g. computational costs and memory limitations).

Idea:

- Make the data smaller by compressing them.
- Perform spatial inference on compressed data.

Adaptive Spatial Dispersion Clustering (ASDC)

Aggregate data into spatial clusters. Requirements:

- produce spatially contiguous clusters
- preserve spatial covariance structure
- minimize information loss for prediction

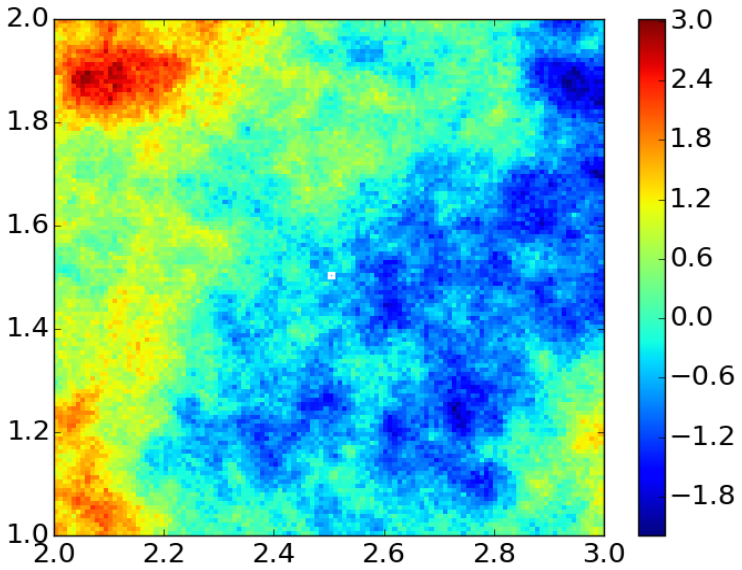
ASDC combines a spatial statistical framework with spectral clustering to meet these requirements.

Distance dispersion functions

Data: $Z(\mathbf{v}) = Y(\mathbf{v}) + \epsilon(\mathbf{v})$,

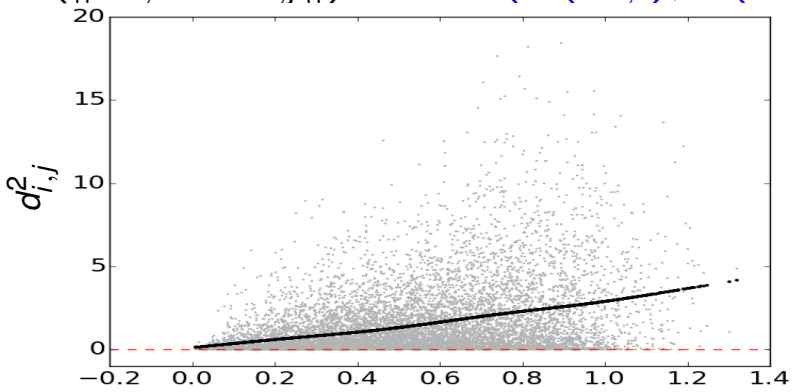
prediction location: $\mathbf{v}_0 \in \mathcal{D}$,

all other locations: $\mathbf{v}_e \in \mathcal{D}_e = \mathcal{D} \setminus \{\mathbf{v}_0\}$.

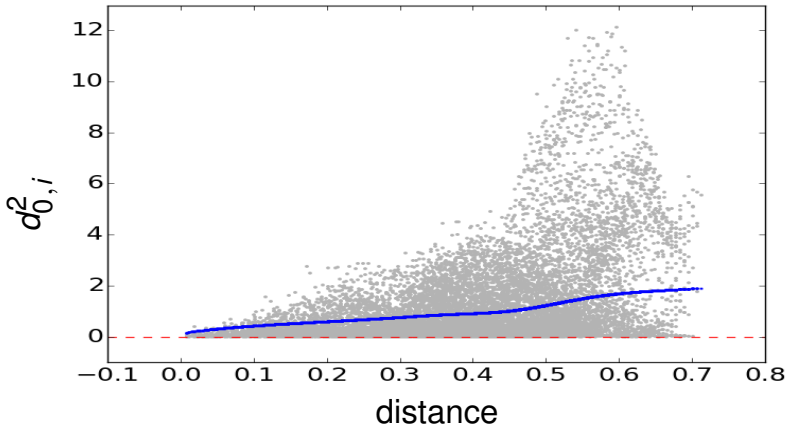


Synthetic spatial domain $\mathcal{D} \subset \mathbb{R}^2$ measurements $Z(\mathbf{v})$, $\mathbf{v} \in \mathcal{D}$.

$f_e(\|\mathbf{v}_{e,i} - \mathbf{v}_{e,j}\|) \Leftrightarrow \mathrm{cov}(Z(\mathbf{v}_{e,i}), Z(\mathbf{v}_{e,j}))$



$f_0(\|\mathbf{v}_0 - \mathbf{v}_{e,i}\|) \Leftrightarrow \mathrm{cov}(Z(\mathbf{v}_0), Z(\mathbf{v}_{e,i}))$



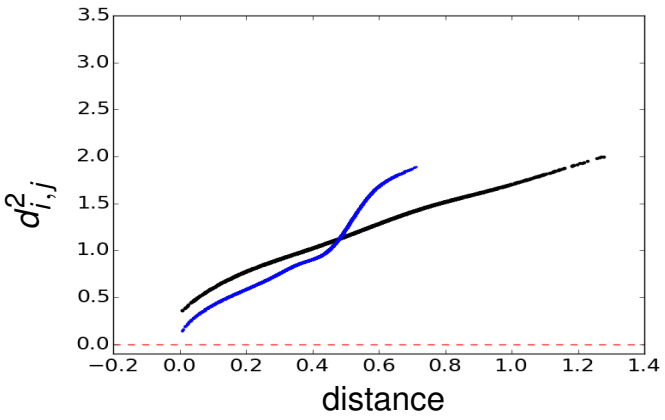
Distance dispersion matrix

Fit f_e and f_0 to $d_{i,j}^2 = |Z(\mathbf{v}_i) - Z(\mathbf{v}_j)|^2$,

$$\gamma_e^{i,j} = f_e(\|\mathbf{v}_{e,i} - \mathbf{v}_{e,j}\|),$$

$$\gamma_0^i = f_0(\|\mathbf{v}_0 - \mathbf{v}_{e,i}\|).$$

f_e and f_0 are fitted by splines



	$\mathbf{v}_0$	$\mathbf{v}_{e,1}$	$\mathbf{v}_{e,2}$	$\cdots$	$\mathbf{v}_{e,N}$	row vector
$\mathbf{v}_0$	0	γ_0^1	γ_0^2	$\cdots$	γ_0^N	$\tilde{\gamma}_0$
$\mathbf{v}_{e,1}$	γ_0^1	0	$\gamma_e^{1,2}$	$\cdots$	$\gamma_e^{1,N}$	$\tilde{\gamma}_e^1$
$\mathbf{v}_{e,2}$	γ_0^2	$\gamma_e^{2,1}$	0	$\cdots$	$\gamma_e^{2,N}$	$\tilde{\gamma}_e^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\mathbf{v}_{e,N}$	γ_0^N	$\gamma_e^{N,1}$	$\gamma_e^{N,2}$	$\cdots$	0	$\tilde{\gamma}_e^N$

H: dissimilarity matrix/table

Spectral clustering

1. Construct a similarity matrix **W** from **H**:

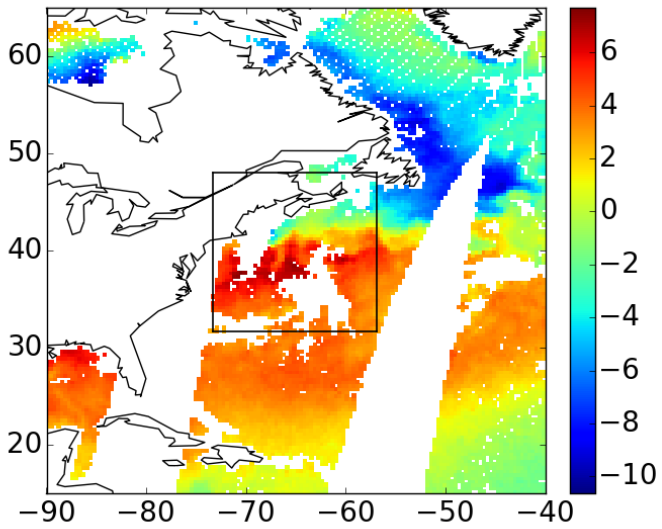
$$w_{i,j} = \exp\left(-\frac{\|\tilde{\gamma}_e^i - \tilde{\gamma}_e^j\|^2}{2\sqrt{\gamma_0^i}\cdot\sqrt{\gamma_0^j}}\right)$$

2. Apply spectral clustering [1] to **W**.

3. Obtain $N \times K$ compression matrix $[\mathbb{Q}]_{ij}$, where ij-th element is 1 if i-th location $\in \Gamma_j$, $j = 1, \ldots, K$.

4. The j-th spatial cluster is $\Gamma_j = \bigcup_{k=1}^{N} \{\mathbf{v}_k : [\mathbb{Q}]_{kj} = 1\}$.

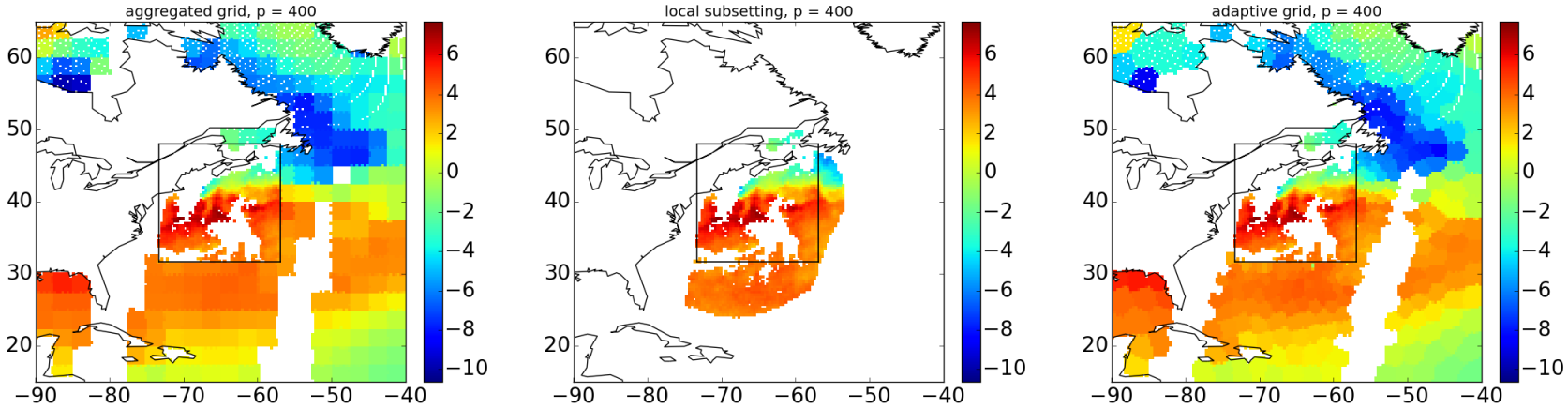
Example: Sea-surface Temperature



Sea-surface temperature from the AMSR-2 instrument on GCOM-W satellite, daytime on 07/28/2015. Data are 25km resolution. There are about 117,000 data points in this region.

- Predict SST in a region of interest (in black square) by kriging compressed data.
- Compare performance to kriging on a proximate subset ("local kriging" [2, 3]) and on binned data.
- Performance benchmark is kriged prediction using the full dataset.

Comparison of methods



binning subsetting ASDC

- A set of M prediction locations $\mathbf{V}_0 = \{\mathbf{v}_{0,1}, \ldots, \mathbf{v}_{0,M}\}$ in $\mathcal{D}_0$ (region of interest).
- Differences in performance are due to differences in compression of $\mathcal{D}_e$ (exterior region).

Step 1: Spatial compression

- Compress the data in $\mathcal{D}_e$ using ASDC, subsetting and binning.
- Obtain compressed datasets $\mathbf{\Psi}^c$, $c = \{\text{Ada}, \text{Sub}, \text{Bin}\}$:

$$\mathbf{\Psi}^c = (\mathbf{Z}_0, \psi^c(\Gamma_1), \ldots, \psi^c(\Gamma_K))^T$$

$$\psi^c(\Gamma_j) = \frac{1}{n_j} \sum_{i=1}^{n} Z(\mathbf{v}_i) 1(\mathbf{v}_i \in \Gamma_j), \quad n_j > 0,$$

is a cluster centroid and n_j is cluster size.

Step 2: Spatial prediction

- Estimate the covariance function parameters from all the available data.
- Perform kriging on full data to predict the region of interest $\mathbf{V}_0$:

$$\hat{Y}(\mathbf{V}_0) = \mathbf{A}^T\mathbf{Z}, \quad \text{where } \mathbf{A} \text{ is a matrix of kriging coefficients.}$$

- Perform kriging on compressed data $\mathbf{\Psi}^c$:

$$\hat{Y}^{\text{Ada}}(\mathbf{V}_0) = \mathbf{B}_{\text{Ada}}^T \mathbf{\Psi}^{\text{Ada}}$$
$$\hat{Y}^{\text{Sub}}(\mathbf{V}_0) = \mathbf{B}_{\text{Sub}}^T \mathbf{\Psi}^{\text{Sub}}$$
$$\hat{Y}^{\text{Bin}}(\mathbf{V}_0) = \mathbf{B}_{\text{Bin}}^T \mathbf{\Psi}^{\text{Bin}}.$$

Performance metrics and results

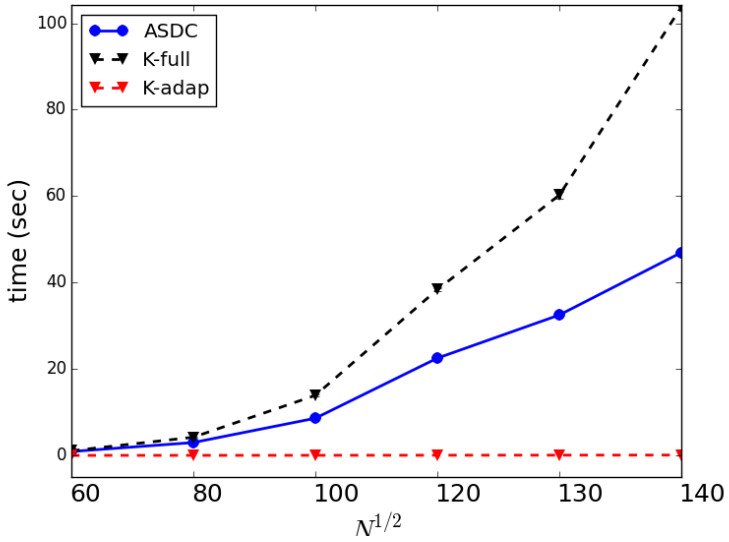
Root mean squared error (RMSE) and mean prediction error ratio (MPER):

$$\mathrm{RMSE}_c = \left[\frac{1}{M} \sum_{i=1}^{M} (\hat{Y}(\mathbf{v}_{0,i}) - \hat{Y}^c(\mathbf{v}_{0,i}))^2 \right]^{1/2}, \qquad \mathrm{MPER}_c = \frac{1}{M} \sum_{i=1}^{M} \frac{\mathrm{var}(\hat{Y}(\mathbf{v}_{0,i}))}{\mathrm{var}(\hat{Y}^c(\mathbf{v}_{0,i}))}.$$

Compressed as	RMSE			MPER		
% of original size	ASDC	Subset	Binned	ASDC	Subset	Binned
1%	0.078	0.208	0.130	1.002	1.029	1.014
2%	0.074	0.177	0.099	0.995	1.015	1.004
4%	0.063	0.106	0.085	0.995	1.005	0.999

Discussion

- Here we demonstrated good performance on a “small" data set against standard competitors.
- Key idea is to preserve both covariance structure in the exterior and between the interior and exterior.
- Simulation studies suggest that computational performance improves with data set size.
- For remote sensing, it will be necessary to combine ASDC with dimension reduction approaches.
- Other potential applications for this methodology include medical imaging, network data, health records, and more.



ASDC computation time vs data size.

References

- Von Luxburg, U., (2007) A tutorial on spectral clustering. *Statistics and Computing* **17**(4) 395–416.
- Haas, T.C. (1990) Kriging and automated variogram modeling within a moving window. *Atmospheric Environment. Part A. General Topics* **24**(7), 1759–1769.
- Hammerling, D.M., Michalak, A.M., Kawa, S.R. (2012) Mapping of CO2 at high spatiotemporal resolution using satellite observations: Global distributions from OCO-2. *Journal of Geophysical Research: Atmospheres,* **117**(D6).

Acknowledgements

This work was performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Government sponsorship acknowledged.

Copyright 2016, California Institute of Technology. All rights reserved.

Contact information: `Yuliya.Marchetti@jpl.nasa.gov`.

We would like to thank JPL's Research and Technology Development program for supporting this work.