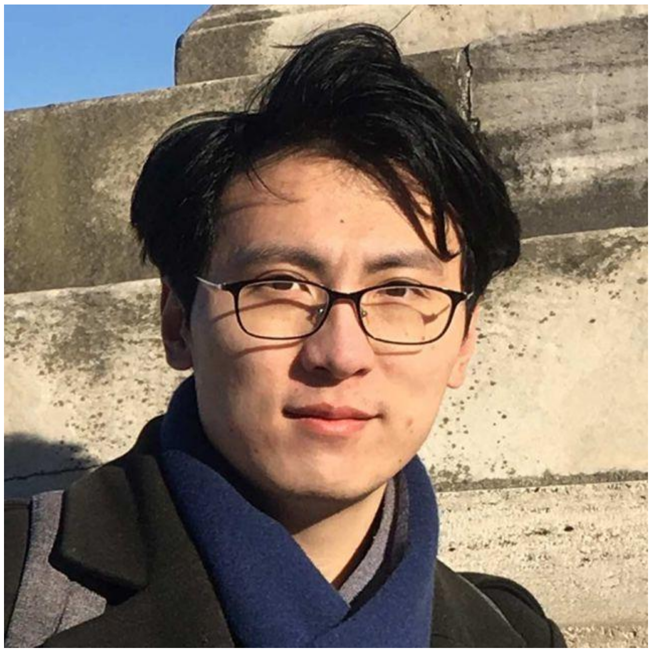
07/24/2024

MC-GTA: Metric-Constrained Model-Based Clustering using Goodness-of-fit Tests with Autocorrelations

Zhangyu Wang, Gengchen Mai, Krzysztof Janowicz, Ni Lao

Presenter Biography



Zhangyu Wang

- 4^{th}-Year PhD Candidate
- Department of Geography, University of California Santa Barbara
- Advisor: Krzysztof Janowicz (UCSB and UWien).
- Research Interest: spatially explicit GeoAI; machine learning/deep learning for geography; spatial information theory.

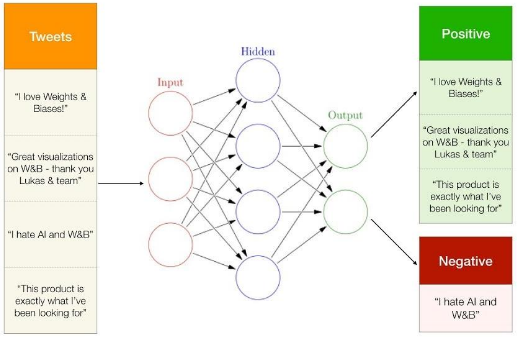
Introduction

- **Q**: Temporal/Spatial is special, but **how** and **why**?

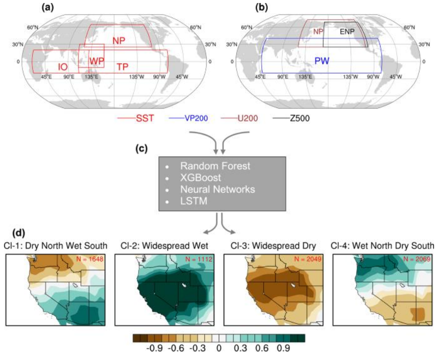
Segmentation

VS

Classification

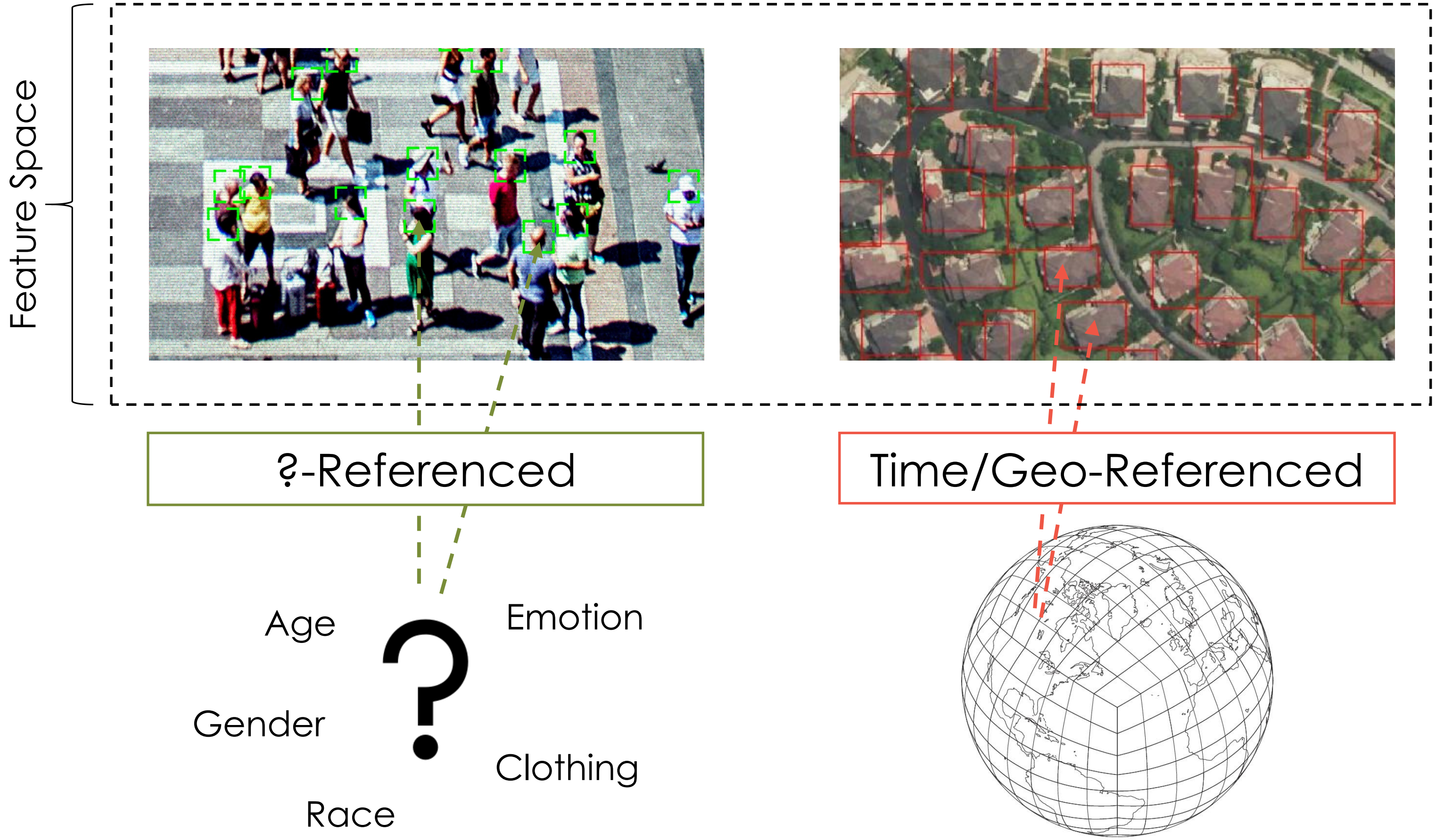


VS



Introduction

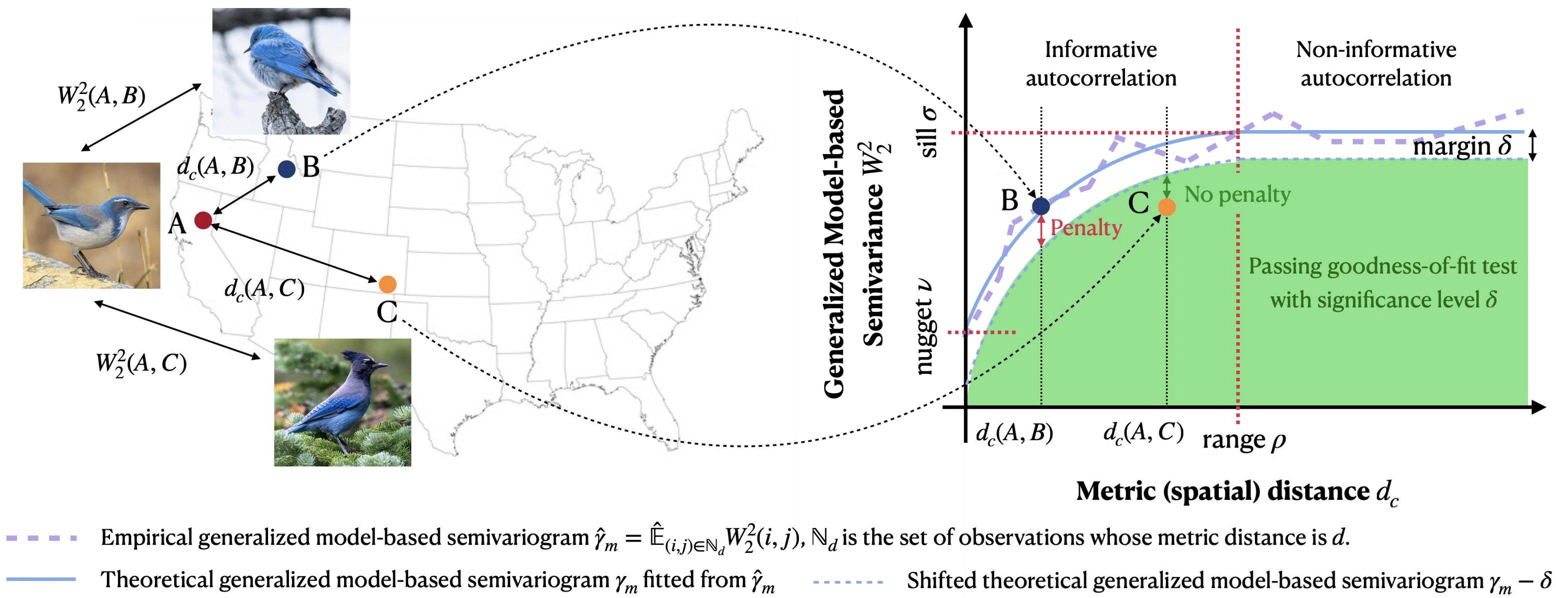
- **A**: The underlying continuous **metric constraints (MC)**



Zhangyu Wang, Department of Geography

MC-Aware Clustering: Necessity

- When we compare spatial samples, we need to notice that part of the difference is the **systematic variance** due to the distance decay of homogeneity, stated in spatial theories as **semivariogram**.



MC-Aware Clustering: Challenges

- We need to carefully distinguish the systematic variance from the true difference between samples, like filtering background noise.

- However, the classic semivariogram has fatal limitations:

$$\hat{\gamma}(h \pm \epsilon) := \frac{1}{2|N(h \pm \epsilon)|} \sum_{\{(\mathbf{p}_i, \mathbf{p}_j) \in N(h \pm \epsilon)\}} |z_i - z_j|^2$$

Univariate

- Modern machine learning/deep learning models are all high-dimensional. The classic univariate definition of semivariogram does not generalize to multivariate cases.

Incompatible with gradient descent algorithms

- Semivariogram is a **function of distance**. It does not fit into real-value based losses.

- Solutions?

  - Multivariate, differentiable generalization of semivariogram – **generalized model-based semivariogram**

Generalized Model-based Semivariogram

- Model-based generalization:

Replace the univariate variance with multivariate statistical distance

$$\hat{\gamma}(h \pm \epsilon) := \frac{1}{2|N(h \pm \epsilon)|} \sum_{\{(\mathbf{p}_i, \mathbf{p}_j) \in N(h \pm \epsilon)\}} \boxed{|z_i - z_j|^2}$$

variance between random variables

$$\hat{\gamma}_m(h \pm \epsilon) := \frac{1}{2|N(h \pm \epsilon)|} \sum_{(\mathbf{p}_i, \mathbf{p}_j) \in N(h \pm \epsilon)} \boxed{W_2^2(i, j)}$$

statistical distance between distributions

Here we use square Wasserstein-2 distance (a.k.a. square Earth Mover's Distance)

$$W_2^2(i, j) = d_2^2(\mu_i, \mu_j) + Tr\left(\Sigma_i + \Sigma_j - 2A\right)$$

Is this a valid generalization?

- If we view random variables as a special case of distributions (single-point distribution), then the variance effectively equals the square Wasserstein-2 distance.

Zhangyu Wang, Department of Geography

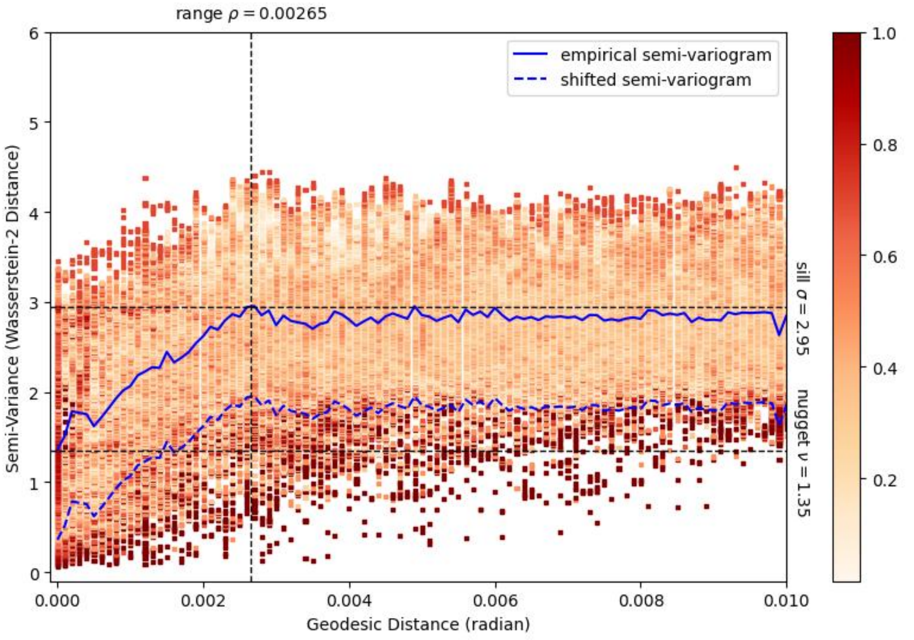
UC SANTA BARBARA

Generalized Model-based Semivariogram

- Theoretical intuition: samples that belong to the same cluster should have lower variance than the average variance of the entire dataset at each distance lag (i.e., the semivariogram).

- Empirical results meet the theoretical intuition extremely well:

  - Samples that show high chance of belonging to the same cluster/class (deep red) form a clear border that resemble the curve of the generalized semivariogram.
  - This border can be used to help cluster spatial data as regularizing information (i.e., soft constraints).



MC-Aware Clustering Objective

- MC-Aware Clustering Objective

Metric-constraint-unaware loss

$$\mathcal{L}(\mathcal{C}) = \sum_{\{C_k \in \mathcal{C}\}} \sum_{\{i,j \in C_k\}} d_m(i,j)$$

Metric-constraint penalized loss

$$\mathcal{L}^{\text{mcm}}(\mathcal{C}) = \sum_{\{C_k \in \mathcal{C}\}} \sum_{\{i,j \in C_k\}} [\boxed{d_m(i,j)} + \beta \boxed{r(i,j)}]$$

sample difference measure

metric-constraint penalty

Criteria for a good choice of $r(i,j)$?

- $r(i,j)$ should be differentiable.
- $r(i,j)$ should be addible to $d_m(i,j)$, i.e., $d_m(i,j) + \beta r(i,j)$ itself should be a valid, well-defined mathematical amount.
- Most desirably, we wish $d_m(i,j) + \beta r(i,j)$ to have some clear theoretical interpretation, i.e., we can intuitively understand what we are minimizing.

MC-Aware Clustering Objective

- MC-GTA Objective

We propose the MC-GTA (Model-based Clustering via Goodness-of-fit Tests with Autocorrelations) objective

$$\mathcal{L}^{\text{MC-GTA}}(\mathcal{C}) = \sum_{\{C_k\in\mathcal{C}\}}\sum_{\{i,j\in C_k\}} \Big[\underbrace{\lfloor W_2^2(i,j) - (\widehat{\mathbb{E}}_{i',j'\in\mathbb{N}} W_2^2(i',j') - \delta^0)\rfloor_+}_{d_m(i,j)} + \beta \underbrace{\lfloor W_2^2(i,j) - (\widehat{\mathbb{E}}_{i',j'\in\mathbb{N}_{i,j}} W_2^2(i',j') - \delta)\rfloor_+}_{r(i,j)} \Big]$$

average sample difference over the entire data

How much more different of the given sample pair is than the data average.

average sample difference over the distance lag

Intuition behind the objective:

- Minimizing the loss primarily encourages sample pairs within each cluster to have lower than data average difference; secondarily, the sample pairs with higher than the average difference over its distance lag, i.e., the semivariance, are punished.

MC-Aware Clustering Objective

- MC-GTA Objective

Generalized semivariogram hinge penalty

$$r(i,j) = \lfloor W_2^2(i,j) - [\gamma_m(d_c(i,j)) - \delta] \rfloor_+$$

Spatially penalized clustering objective

$$\mathcal{L}^{\text{MC-GTA}}(\mathcal{C}) = \sum_{\{C_k \in \mathcal{C}\}} \sum_{\{i,j \in C_k\}} [W_2^2(i,j) + \beta r(i,j)]$$

Merits of the MC-GTA objective:

- This objective is a natural extension of the conventional, non-penalized clustering objective. It is obviously differentiable.
- The property of square Wasserstein-2 distance ensures that minimizing the penalty equals passing a **goodness-of-fit test** with null hypothesis being that "the two distributions are statistically the same".
- Further math proves that square Wasserstein-2 distance is the **tightest possible** penalty.

Experimental Results

- Experimental results: temporal and spatial clustering

		Synthetic Datasets				Real-world Datasets															
						Temporal						Spatial									
		Temporal d=5, c=5 N=1,000		Spatial d=5, c=5 N=10,000		Pavement d=10, c=3 N=1,055		Vehicle d=7, c=5 N=16,641		Gesture d=3, c=8 N=704,970		Climate d=5, c=14 N=4,741		iNat2018 d=16, c=6 N=24,343		POI d=7, c=10 N=23,019		Landuse d=7, c=5 N=8,964	
Model Type	Model	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI
No-Constraint Model-Free	k-Means	1.03	1.69	1.26	1.66	8.02	6.59	8.94	21.54	2.78	5.23	5.47	22.14	6.91	14.71	18.37	43.44	2.39	4.21
	DBSCAN	2.44	2.50	3.69	5.38	15.25	18.75	33.67	41.83	1.18	2.07	3.61	17.89	34.91	34.69	15.03	39.29	11.91	7.19
	HDBSCAN	0.90	0.61	1.00	1.39	7.10	11.66	37.51	41.64	-	-	11.52	28.01	7.65	17.92	20.78	62.55	1.00	7.64
	DTW	2.52	2.13	-	-	17.13	17.55	8.11	23.35	-	-	-	-	-	-	-	-	-	-
Constrained Model-Free	PCK-Means	5.12	5.68	2.30	2.89	7.42	5.13	4.80	14.17	NC	NC	18.50	34.67	25.51	28.96	0.12	0.18	0.11	0.26
	MDST-DBSCAN	-	-	1.12	5.73	-	-	-	-	-	-	11.32	27.89	8.43	18.13	1.33	0.97	1.29	1.01
	SKATER	-	-	23.87	32.29	-	-	-	-	-	-	**23.44**	**44.10**	0.51	0.35	1.52	0.91	1.03	0.74
No-Constraint Model-Based	GMM	7.82	9.54	9.26	10.35	28.05	28.74	57.87	58.78	2.44	4.15	19.06	34.97	21.72	35.91	16.38	42.96	2.86	4.61
	(S)TICC-β=0	80.11	83.95	91.28	89.28	58.54	58.83	40.12	45.86	3.26	6.56	13.30	30.53	NC	NC	13.29	27.08	7.22	12.60
	MC-GTA-wo	86.38	84.56	87.34	84.74	76.10	74.36	63.31	58.60	8.12	33.60	16.63	36.73	21.90	36.47	30.45	66.23	12.91	28.72
Constrained Model-Based	(S)TICC	84.88	86.13	91.84	89.85	62.27	61.89	50.53	53.68	12.20	23.20	17.62	37.29	NC	NC	NC	NC	11.04	15.35
	MC-GTA-w	**90.50**	**87.96**	**94.49**	**91.98**	**77.64**	**77.22**	**65.04**	**59.36**	**26.51**	**55.34**	20.08	40.91	**42.70**	**40.49**	**39.81**	**68.27**	**36.54**	**42.97**

Table 2. Comparing different feature similarity measures

Wasserstein-2		Euclidean		Cosine		Total Var.		KL-D		JS-D	
ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI
77.64	77.22	23.11	22.43	0.34	1.36	3.37	3.61	56.73	66.10	15.55	18.84

Table 4. Comparing different distance-based clustering algorithms

Method	TICC (Baseline)		MC-GTA (DBSCAN)		MC-GTA (HDBSCAN)		MC-GTA (OPTICS)	
Performance	ARI 62.27	NMI 61.89	ARI 77.64	NMI 77.22	ARI 72.35	NMI 69.61	ARI 69.77	NMI 68.58

Thank You!

UC SANTA BARBARA