Variograms for spatial functional data with phase variation

SEBASTIAN KURTEK
Department of Statistics
The Ohio State University

Collaborators: Xiaohan Guo (Biostatistics, The Ohio State University), Karthik Bharath (Mathematical Sciences, University of Nottingham)

Research Support: NSF DMS-1613054, NSF DMS-2015226, NSF CCF-1740761, NSF CCF-1839252 and NIH R37-CA214955

DIFFERENT TYPES OF SPATIAL FUNCTIONAL DATA

1. Spatially indexed functional data with phase variation
   - Spatial prediction
   - Clustering
   - Applications: environmetrics, electroencephalogram (EEG) signals

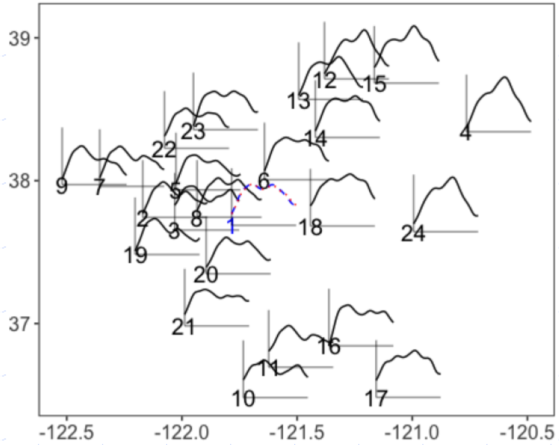
2. Spatially indexed shapes
   - Shapes of curves in two dimensional images with image coordinates serving as spatial locations
   - Shapes of surfaces in three dimensional images
   - Applications: medical imaging, biology, graphics, computer vision

3. Marked point processes
   - Point processes on spatial domain with geometric marks, e.g., functions, shapes, trees, etc.
   - Applications: medical imaging, biology, graphics, computer vision

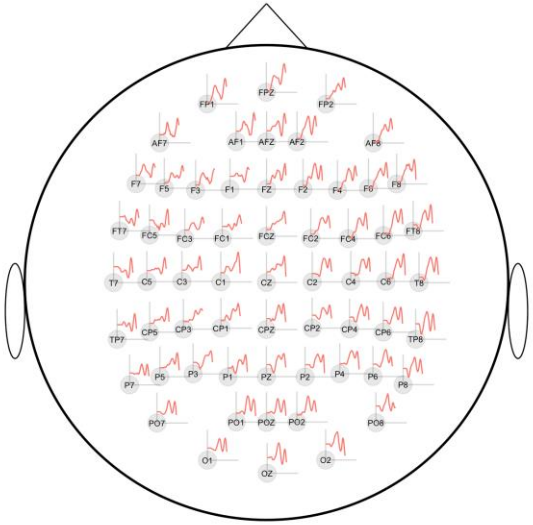
THE OHIO STATE UNIVERSITY

SPATIAL FUNCTIONAL DATA EXAMPLES

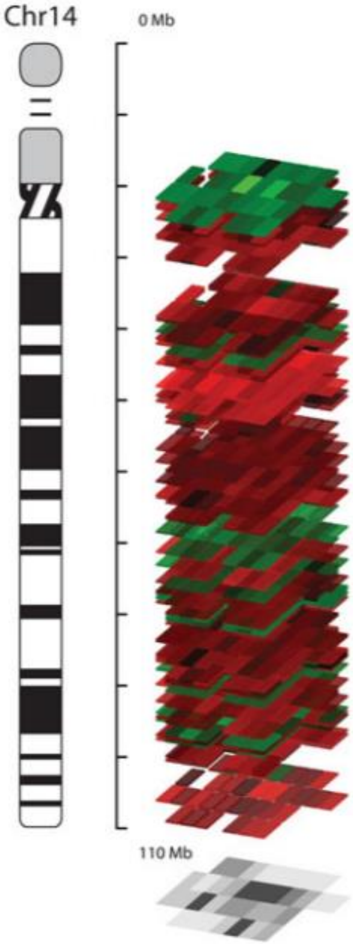
Daily Ozone Data in Northern California



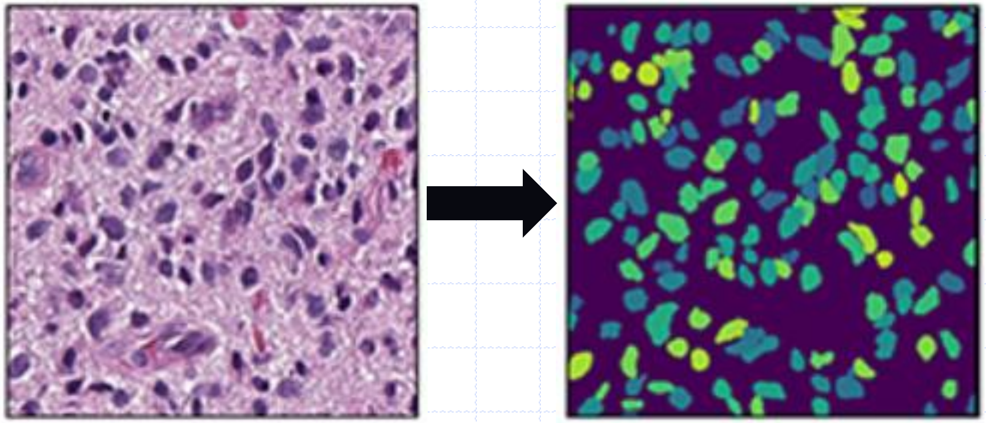
EEG Data



Histologic Image with Genetic Map
(Zhang et al., 2016)



Tumor Cells in Histopathology Images
(Vu et al., 2019)



BRIEF LITERATURE REVIEW

- Brief literature review:
    - Nerini et al., 2010: applied multivariate spatial statistics after dimension reduction
    - Giraldo et al., 2011: defined L^2 metric-based trace-variogram for functional data
    - Mateu and Romano, 2017: applied trace-variogram for spatial prediction and clustering
    - Caballero et al., 2013; Menafoglio et al., 2013; Reyes et al., 2015; Menafoglio & Petris, 2016: extensions that relax assumptions
- Most current trace-variogram based methods assume that given functions are perfectly aligned or treat phase variation as negligible noise.

WHAT IS A VARIOGRAM?

- Second order stationary and isotropic random field: $\{Z_s : \ s \in \mathcal{D}\}$
- Standard definition of variogram:

$$\|s - s'\|_2 = h \mapsto V(h) = \frac{1}{2}E((Z_s - Z_{s'})^2)$$

- Second order stationary and isotropic functional random field: $\{f_s : s \in \mathcal{D}\}$
- L^2 metric-based trace-variogram:

$$\|s - s'\|_2 = h \mapsto V(h) = \frac{1}{2}E(\|f_s - f_{s'}\|^2) = \frac{1}{2}E\left(\int_0^1 (f_s(t) - f_{s'}(t))^2 dt\right)$$

- Captures the amount of spatial dependence in the random field.
- Plays a key role in spatial prediction and clustering.

CONTRIBUTIONS

1. We define an amplitude trace-variogram on the spatial domain.

2. In order to account for the relative nature of phase, we define a conditional phase trace-variogram on an augmented domain comprising shape of the observed functions as a covariate.

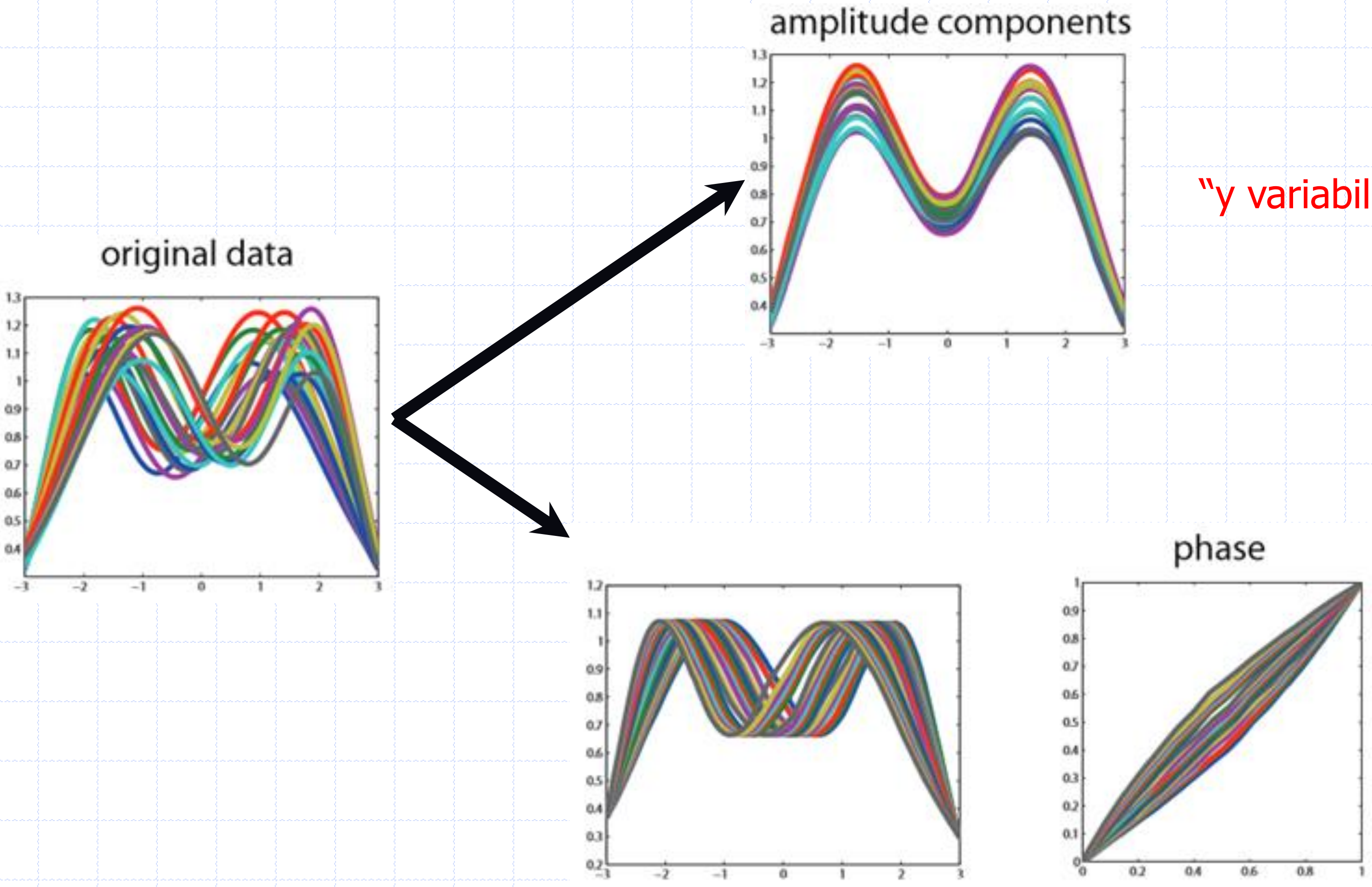
3. We propose an algorithm to compute a spatially-weighted mean, which enables joint alignment of functions and computation of estimators of the amplitude and phase trace-variograms.

4. Based on the variograms, we propose

   a. linear unbiased estimators for spatial prediction of amplitude and phase (and combine them to form the final prediction), and

   b. a method for clustering spatial functional data into amplitude and phase clusters.

AMPLITUDE AND PHASE VARIABILITY



“y variability”

“x variability”

THE OHIO STATE UNIVERSITY

ELASTIC FUNCTIONAL DATA ANALYSIS

- Function space: $\mathcal{F} = \{f : [0,1] \to \mathbb{R} | f \text{ is absolutely continuous}\}$
- Phase component:

  $\Gamma = \{\gamma : [0,1] \to [0,1] | \gamma(0) = 0, \ \gamma(1) = 1, \ \gamma \text{ is a diffeomorphism}\}$
- Function warping: $(f, \gamma) = f \circ \gamma$
- Square-Root Slope Function (SRSF): $Q : \mathcal{F} \to \mathbb{L}^2([0,1], \mathbb{R})$

$$Q(f) = q = \text{sign}(\dot{f})\sqrt{|\dot{f}|}$$

- Simplification: L^2 metric on SRSF space is invariant to warping.
- Function warping through SRSF: $(q, \gamma) = (q \circ \gamma)\sqrt{\dot{\gamma}}$
- Amplitude: $[f] = \{f \circ \gamma | \gamma \in \Gamma\}$ or $[q] = \{(q, \gamma) | \gamma \in \Gamma\}$

THE OHIO STATE UNIVERSITY

ELASTIC FUNCTIONAL DATA ANALYSIS

- Amplitude distance: $d_a(q_1, q_2) = \inf_{\gamma \in \Gamma} \ \|q_1 - (q_2, \gamma)\|$
- Optimal phase: $\gamma^* = \arg\min_{\gamma \in \Gamma} \|q_1 - (q_2, \gamma)\|$
- Shape distance: $d_{sh}(q_1, q_2) = d_a(q_1/\|q_1\|, q_2/\|q_2\|)$
- Representation of phase:

$$\gamma \mapsto Q(\gamma) = \psi = \sqrt{\dot{\gamma}}$$

- $Q(\Gamma) = \Psi$ is the positive orthant of the unit Hilbert sphere.
- Intrinsic relative phase distance:

$$\psi^* = Q(\gamma^*) \longrightarrow d_p^{int}(q_1, q_2) = \cos^{-1}(\int_0^1 \psi^*(t)\psi_{id}(t)dt)$$

- Extrinsic phase distance: $\|\psi_1 - \psi_2\|$

AMPLITUDE/PHASE TRACE-VARIOGRAMS

- Functional random field: $\{f_s : s \in \mathcal{D}\} \longrightarrow \{q_s, s \in \mathcal{D}\}$
- Observed spatial functional data: $q_{s_i},\ s_i \in \mathcal{D}\ (i = 1, \ldots, n)$
  (subscript s_i simply referred to as i)
- Amplitude trace-variogram (under second order stationarity and isotropy):

$$\|s - s'\|_2 = h \mapsto V_a(h) = \frac{1}{2} E\left(\|(q_s, \gamma_s) - (q_{s'}, \gamma_{s'})\|^2\right)$$

- 'Spatial' distance for phase trace-variogram:

$$\|y_1 - y_2\|_\omega^2 = \|s_1 - s_2\|_2^2 + \omega d_{sh}^2(q_1, q_2)$$

($\omega \geq 0$ is a tuning parameter)

- Phase trace-variogram (under second order stationarity and isotropy):

$$\|y - y'\|_\omega = h \mapsto V_p(h) = \frac{1}{2} E\left(\|\psi_y - \psi_{y'}\|^2\right)$$

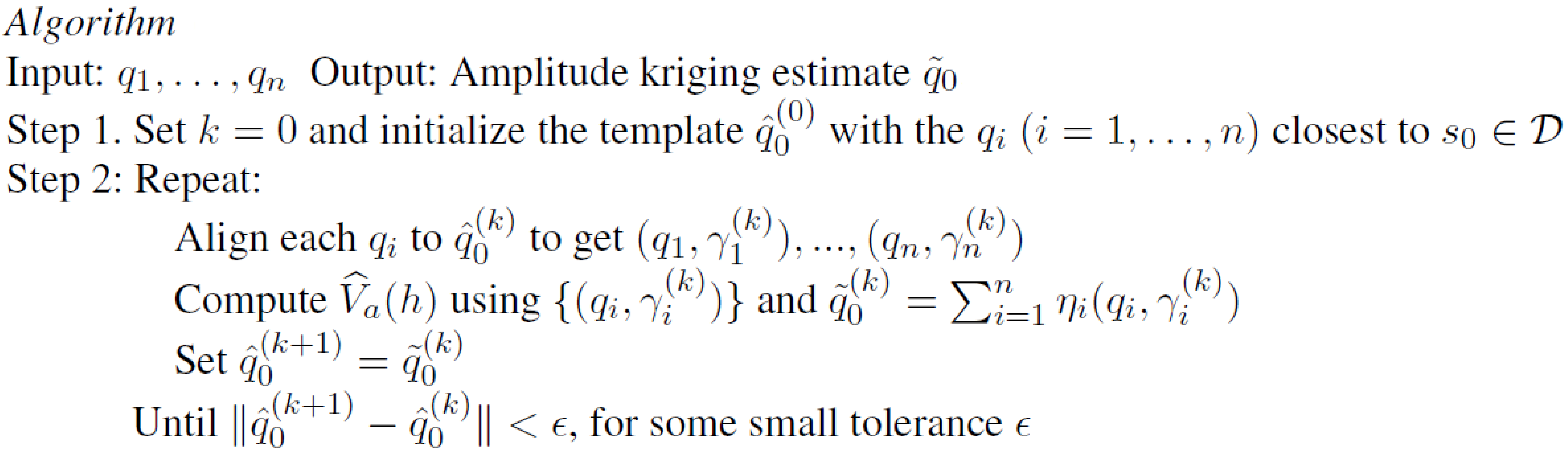
THE OHIO STATE UNIVERSITY

AMPLITUDE SPATIAL PREDICTION

- Goal: given data $\{(s_i, q_i) \mid s_i \in \mathcal{D}\}\ (i = 1, \ldots, n)$, estimate the amplitude component (q_0, γ_0) at a new location $s_0 \in \mathcal{D}$.
- Key idea: define a spatially-weighted amplitude estimator that serves the dual purpose of a local template for alignment and as an estimate of the amplitude component.
- Linear estimator of amplitude: $\tilde{q}_0(t) = \sum_{i=1}^{n} \eta_i (q_i, \hat{\gamma}_i)(t)$
- Minimizes the expected amplitude prediction error functional.
- The optimal weights can be computed solely using the plug-in estimate of the amplitude trace-variogram:

$$\widehat{V}_a(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \|(q_i, \hat{\gamma}_i) - (q_j, \hat{\gamma}_j)\|^2, \quad N(h) = \{(s_i, s_j) \mid \|s_i - s_j\| = h\}$$

AMPLITUDE SPATIAL PREDICTION



```
*Algorithm*
Input: $q_1, \dots, q_n$  Output: Amplitude kriging estimate $\tilde{q}_0$
Step 1. Set $k = 0$ and initialize the template $\hat{q}_0^{(0)}$ with the $q_i$ $(i = 1, \dots, n)$ closest to $s_0 \in \mathcal{D}$
Step 2: Repeat:
        Align each $q_i$ to $\hat{q}_0^{(k)}$ to get $(q_1, \gamma_1^{(k)})$, ..., $(q_n, \gamma_n^{(k)})$
        Compute $\widehat{V}_a(h)$ using $\{(q_i, \gamma_i^{(k)})\}$ and $\tilde{q}_0^{(k)} = \sum_{i=1}^n \eta_i (q_i, \gamma_i^{(k)})$
        Set $\hat{q}_0^{(k+1)} = \tilde{q}_0^{(k)}$
    Until $\|\hat{q}_0^{(k+1)} - \hat{q}_0^{(k)}\| < \epsilon$, for some small tolerance $\epsilon$
```

- At each iteration, we

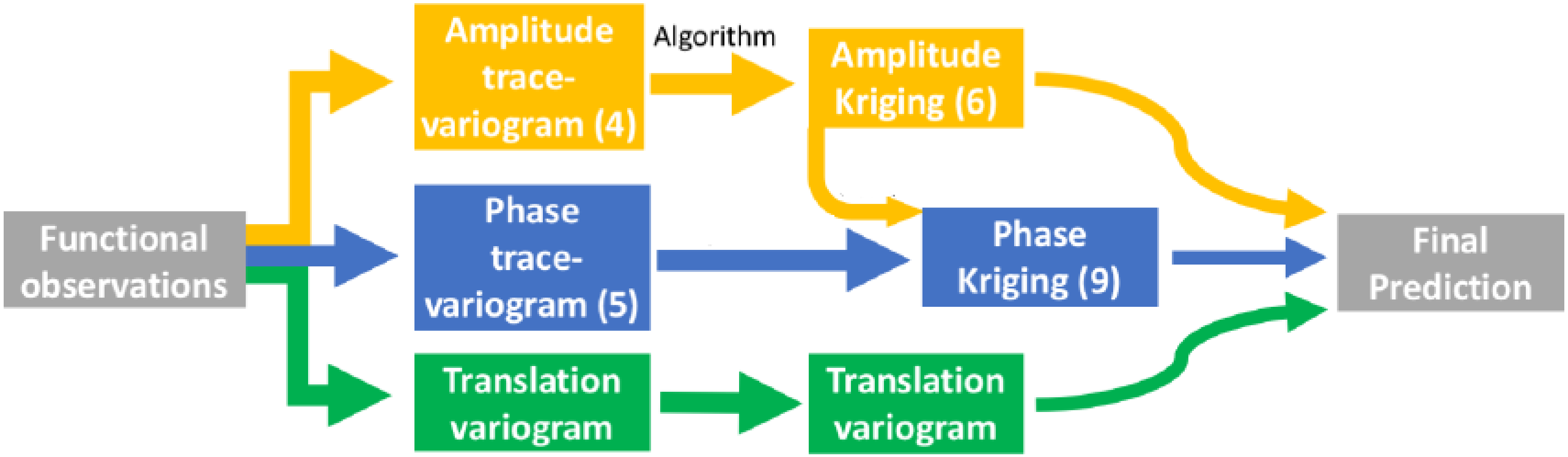
  1. align all observed functions with respect to the current estimate of the spatial amplitude component;
  2. estimate the optimal weights using the trace-variogram;
  3. update the current estimate of the spatial amplitude component.

- The algorithm also results in optimal phase of each function with respect to the spatial amplitude prediction.

SPATIAL PHASE AND FINAL PREDICTION

- The spatial phase prediction is computed in similar fashion (no need for an iterative algorithm), but using the phase trace-variogram based on the modified ‘spatial’ distance.

- The final prediction combines the spatial amplitude and phase predictions, as well as a prediction of function translation (computed using the standard notion of a variogram).



PROPERTIES

1. Under a simple model and some additional conditions, the amplitude estimator computed via the Algorithm converges to an element of the correct orbit as the number of observations in the neighborhood of the new location becomes dense.

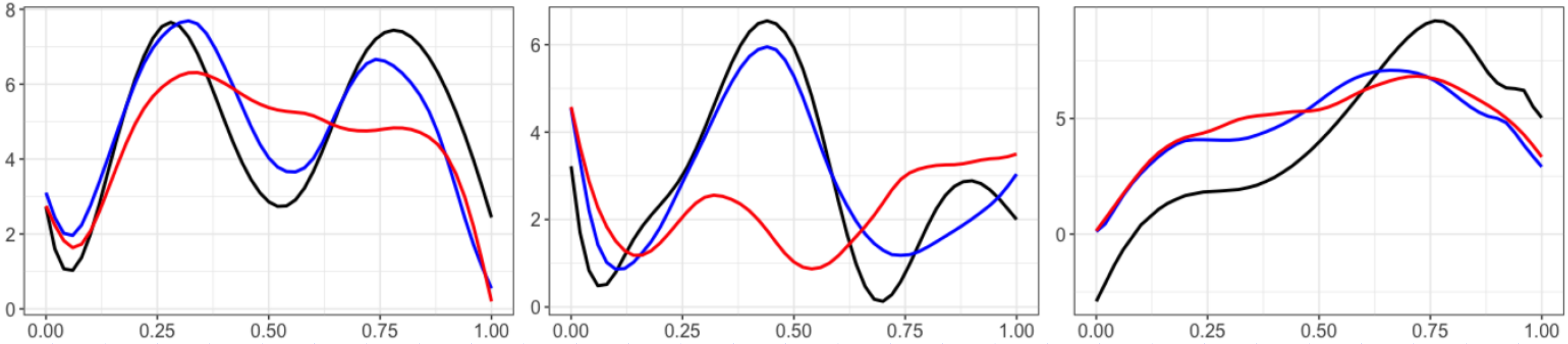
2. Minimizing the amplitude prediction error is equivalent to minimizing a quadratic form that only involves the weights in the linear estimator of amplitude and the amplitude trace-variogram.

3. Property (2) also holds for the phase predictor, but uses the phase trace-variogram.

4. Properties (2) and (3) result in efficient estimation of the amplitude and phase components of the final spatial prediction.

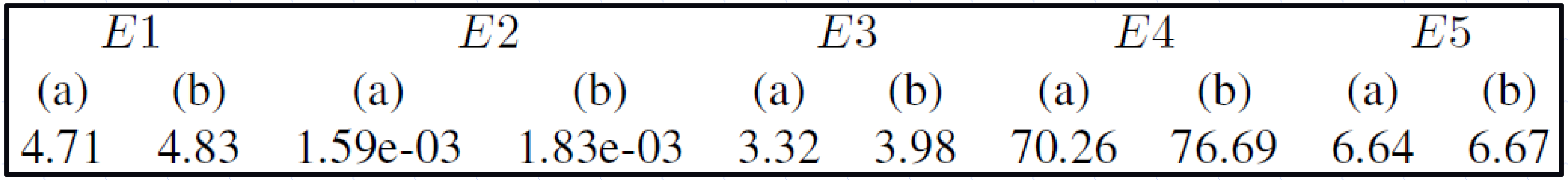
THREE SIMULATED EXAMPLES

- Black: true function
- Red: L^2 metric-based trace-variogram prediction
- Blue: proposed approach



EVALUATION ON REAL DATA

- Leave-one-out cross-validation for the daily ozone data in Northern California.
- Five different error metrics:
  - E1-E3: amplitude errors
  - E4: phase error
  - E5: mean squared error
- (a) proposed approach; (b) L^2 metric-based trace-variogram prediction



$E1$		$E2$		$E3$		$E4$		$E5$	
(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
4.71	4.83	1.59e-03	1.83e-03	3.32	3.98	70.26	76.69	6.64	6.67

SPATIALLY WEIGHTED HIERARCHICAL CLUSTERING

- Spatially-weighted amplitude and phase distance matrices:

$$d_{A,ij} = d_a(q_i, q_j) \times V_a(\|s_i - s_j\|_2)$$

$$d_{P,ij} = d_p^{int}(q_i, q_j) \times V_p(\|y_i - y_j\|_\omega)$$

- Estimate of amplitude trace-variogram:

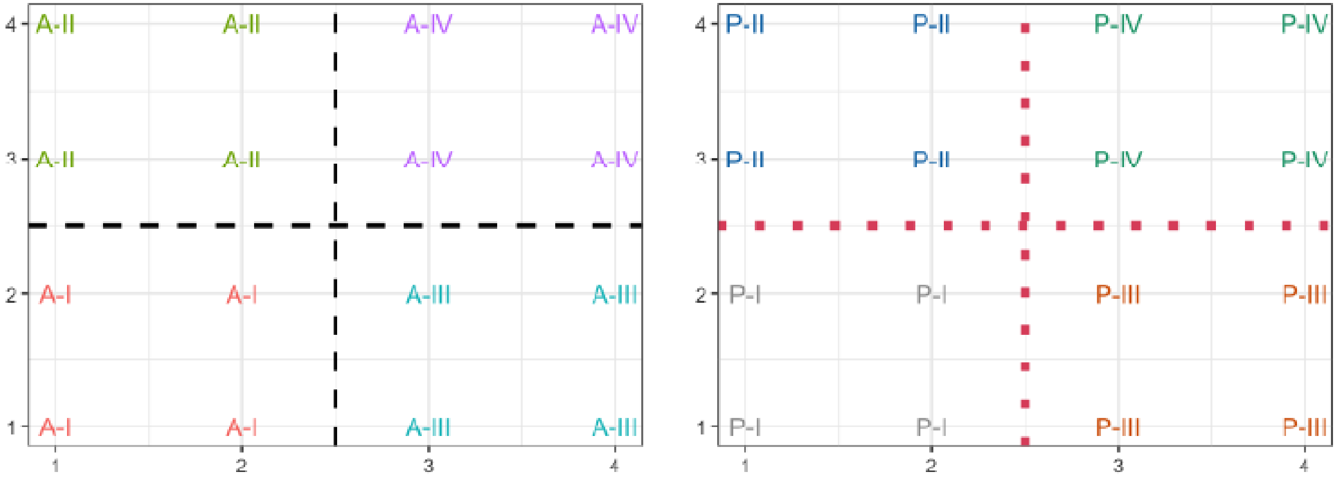
$$\widehat{V}_a(h) = \frac{1}{2|N_a(h)|} \sum_{i,j \in N_a(h)} d_a(q_i, q_j)^2 \qquad N_a(h) = \{(i,j) | h = \|s_i - s_j\|\}$$

- Estimate of phase trace-variogram:

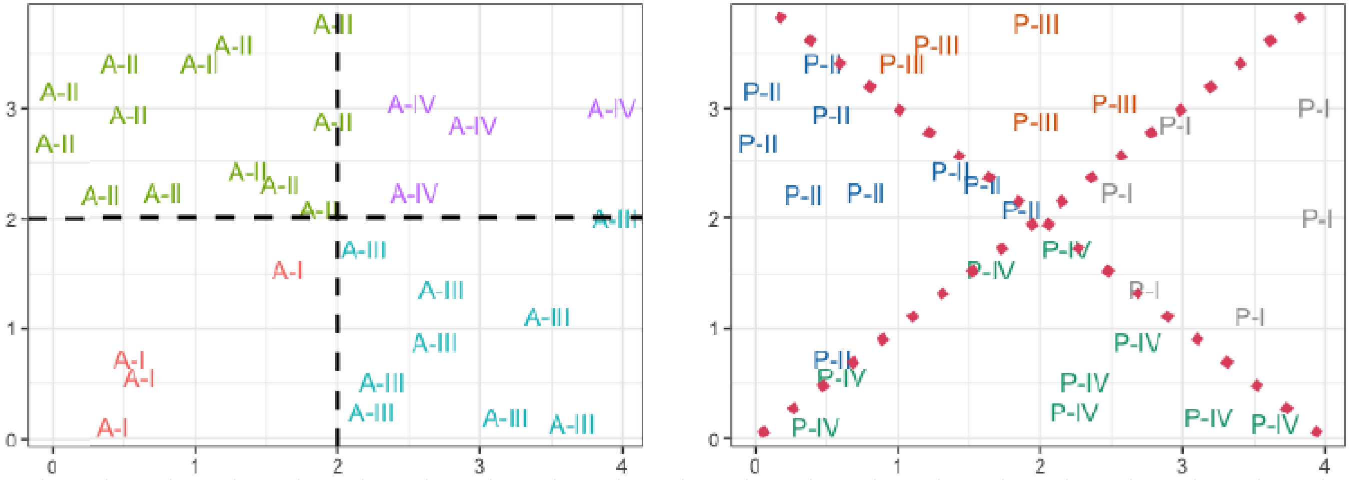
$$\widehat{V}_p(h) = \frac{1}{2|N_p(h)|} \sum_{i,j \in N_p(h)} d_p^{int}(q_i, q_j)^2 \qquad N_p(h) = \{(i,j) | h = \|y_i - y_j\|_\omega\}$$

SIMULATION STUDY

Agree Design

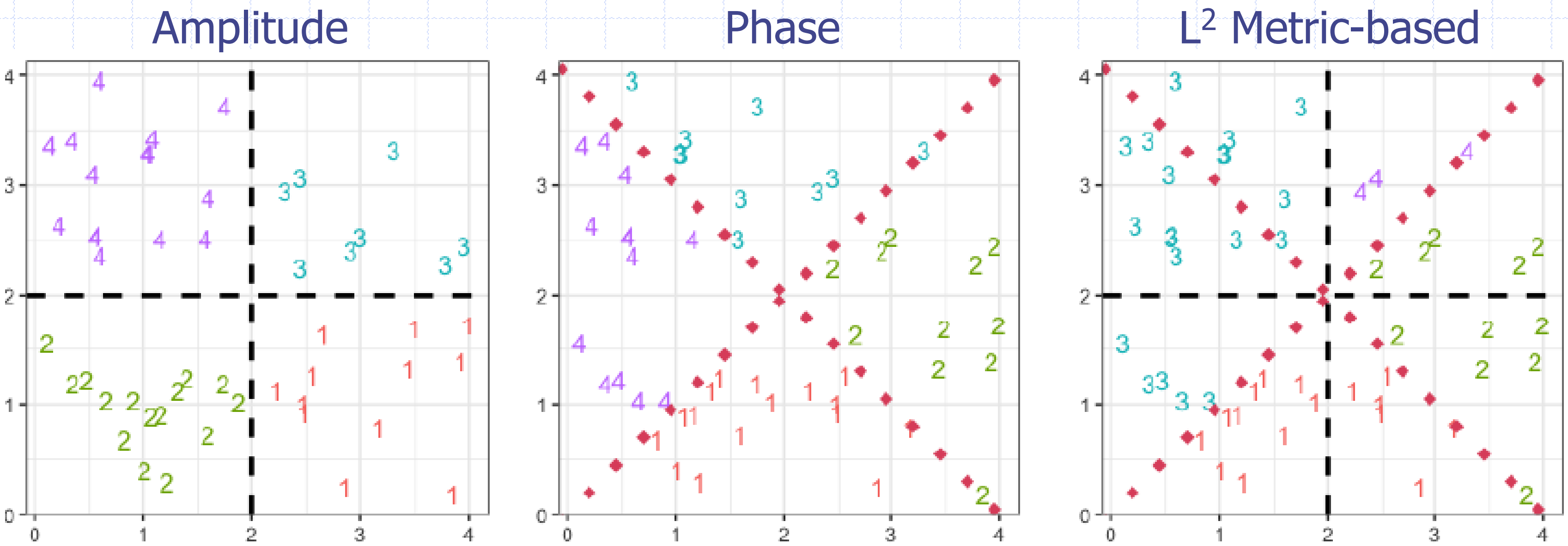


Disagree Design



SIMULATION STUDY

Disagree Design Result



Amplitude

Phase

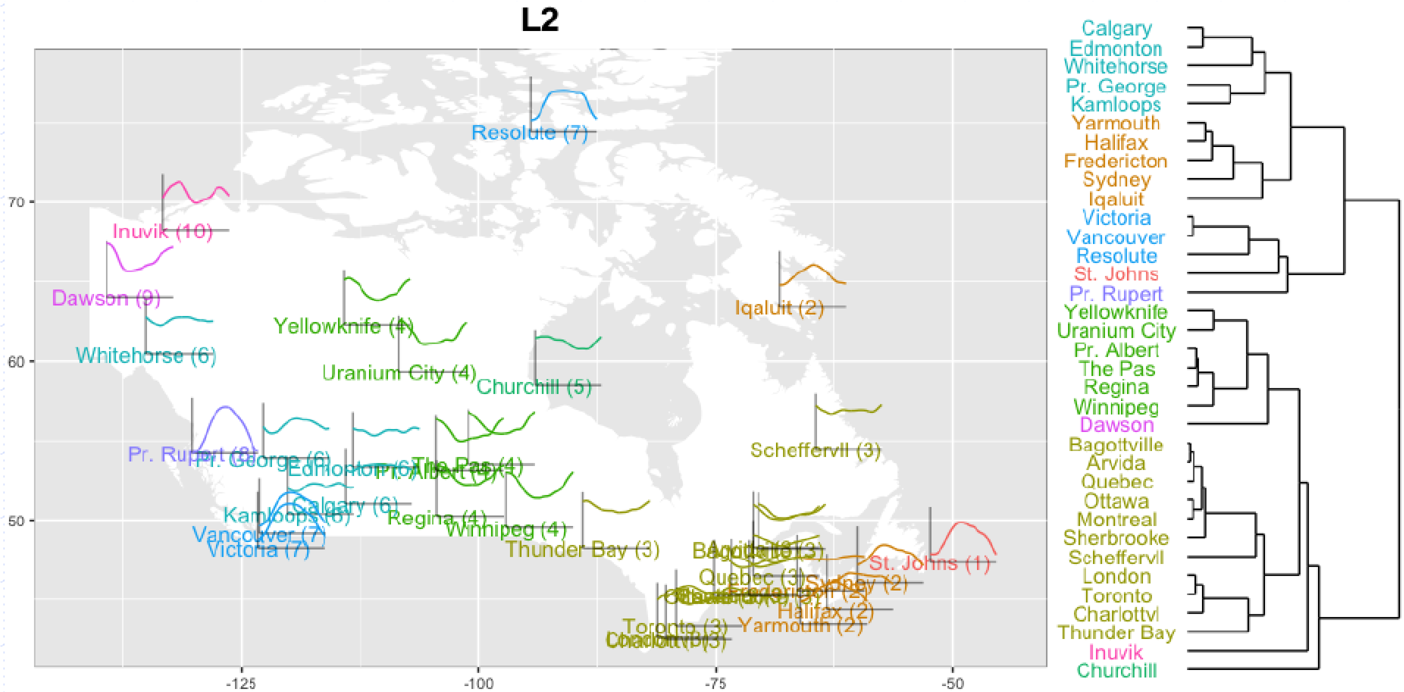
L^2 Metric-based

- L^2 Metric-based clustering must always compromise between the amplitude and phase components.

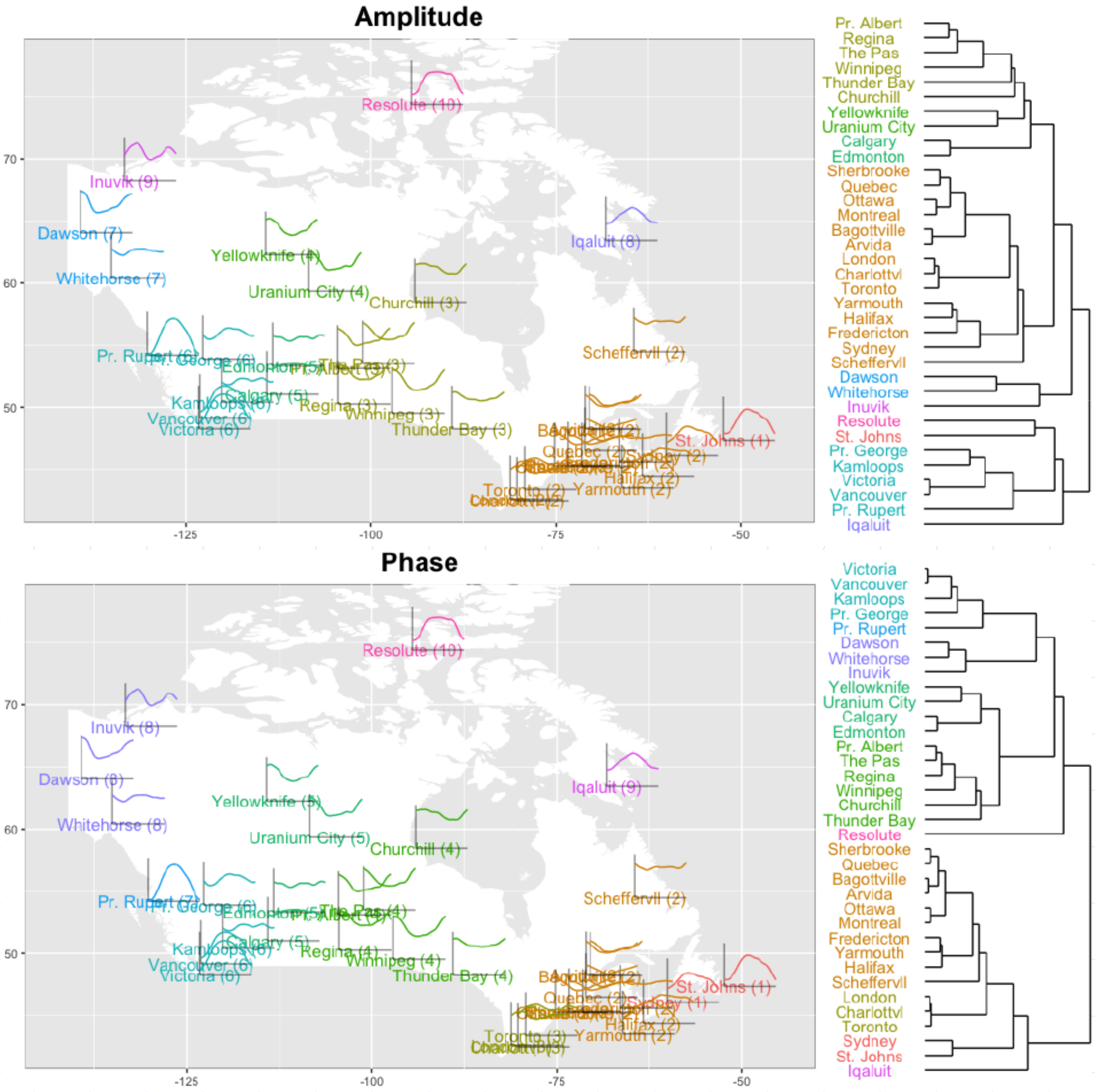
- Using the rand index as an evaluation metric, the proposed method outperforms the L^2 metric-based approach in terms of amplitude and phase clustering.

REAL DATA STUDY

- Clustering of Canadian temperature data: daily temperature data averaged over 1960-1994, recorded at 35 stations.
- Pre-processing: removed latitude and longitude effects using functional linear regression.
- Selected ten clusters for each approach.



REAL DATA STUDY



X. Guo, S. Kurtek, K. Bharath, *Variograms for spatial functional data with phase variation*, arXiv:2010.09578.

THANK YOU!

QUESTIONS?