D. Details on Simulation Studies

All simulation experiments were carried out on an Apple Macbook Pro with M1 Pro processor with 32 GB of memory. The scikit-learn clustering package[24] and scikit-fuzzy[26] package were used for all experiments to perform traditional clustering as well as handling Gaussian process regression for the GPSC algorithm and computing clustering metrics. All code for the simulation studies has also been made available. Note that for all simulations in the main paper, as well as in each of the additional simulations presented in this section, GPSC and all other competitor algorithms were tuned on a single random seed. Experiments were then replicated using these same parameters 49 more times on the next 49 random seeds in order for a total of 50 replicates, and the adjusted Rand index and adjusted mutual information scores are reported as mean $\pm$ standard deviation. Finally, an early stopping condition was employed for all experiments. When both the adjusted Rand index and adjusted mutual information both are above 0.90 (exact value can be set by user) on an iteration by iteration basis, the algorithm is thought to have converged and stopped. The exact values for all experiments are presented with the code.

For parameter tuning of the competitor methods, a grid search over the parameters maximizing the adjusted mutual information score against the true labels was performed as followed: 1) For K-means, the default parameters were used in the scikit-learn package. 2) For spectral clustering, the affinity matrix was determined by nearest neighbors, where the number of neighbors was tuned between 1 and 50 in increments of one. 3) For DBSCAN, the eps parameter (maximum distance between two samples in a single neighborhood) was searched between 1 and 100 in increments of one, and the minimum number of samples in a neighborhood was tuned between 1 and 40 in increments of one. 4) For standard hierarchical clustering, the default parameters were used under the ward linkage. 5) For supervised fuzzy C-means, the default arguments in the scikit-fuzzy package were used, except the algorithm was initialized using the response variable y as the labels. 6) For GDBSCAN, the distance thresholds were tuned individually for each simulation. 7) For spatialized hierarchical clustering, the spatial connectivity matrix was determined by k-nearest neighbors, where the number of neighbors was searched between 1 and at least 75 in increments of one, and where the linkage was also varied between the set {average, complete, ward, single}. 8) Finally for the Gaussian mixture model, the default parameters in the scikit-learn package were used. Any auxiliary parameters unspecified here were left as the default values from the packages.

D.1. *Simulation 1*

The data used in this simulation takes the form $\{(s_i, x_i, y_i)\}_{i=1}^n$, where $s_i \in \mathbb{R}^2$, the spatial domain, $x_i \in \mathbb{R}^2$, the covariate domain, and $y_i \in \mathbb{R}$, the response domain, for visualization purposes.

In this simulation, both $s_i \in \mathbb{R}^2$ and $x_i \in \mathbb{R}^2$ are generated from independent uniform distributions, where $s_i \sim \text{Unif}(-5, 5)$ and $x_i \sim \text{Unif}(-3, 3)$ component-wise.

After generating the data $\{(s_i, x_i)\}_{i=1}^n$, where $n = 1000$ samples, the domain square is subdivided into two clusters, the ball shape cluster and the rest region. This is done by subsetting all points $\{(s_i, x_i)\}$ within 2.8 units of the point (0, 0) solely in the spatial domain into cluster 2 (ball), and the remaining points of the background into cluster 1.

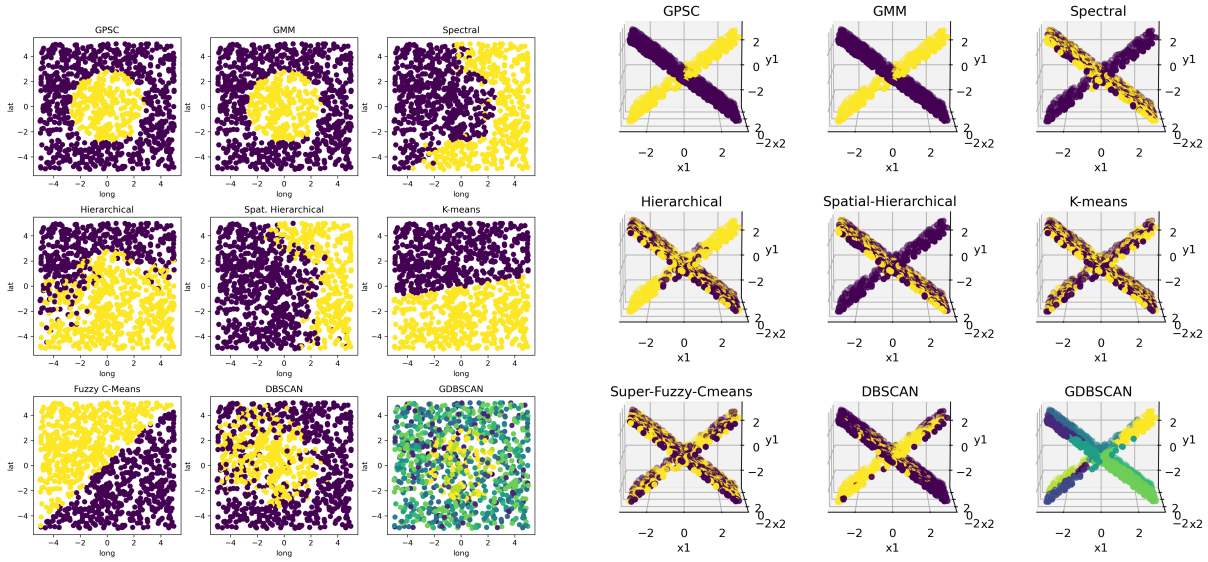
For each cluster, y is generated as a linear function of x. For cluster 1, the true function is:

$$y = -x_1.$$

And for cluster 2, the true function is:

$$y = x_1.$$

After the data was generated, GPSC was applied with the following input: 2 clusters, 50 iterations with early stopping, GP input $\{x_i, s_i\}$, constant bounds $(1e^{-15}, 1e^{6})$, length scale bounds $(1e^{6}, 1e^{15})$, input data. For the GP, the RBF kernel was used. Then K-means clustering, spectral clustering, hierarchical clustering with Ward linkage, and DBSCAN were also applied. For spectral clustering, the affinity matrix was generated with using nearest-neighbors set to 11. For DBSCAN, the maximum distance was set to 77, and minimum number of samples set to 3. The full set of $\{x_i, s_i, y_i\}$ as a vector was input into each algorithm along with $L = 2$ clusters where relevant. For supervised C-means clustering, the supervised labels were set according to the y domain with otherwise default parameters. For GDBSCAN, the covariate distance threshold was set to 3, spatial distance set to 13, and minimum set to 0. For the spatialized hierarchical clustering method, the connectivity matrix was specified using k-nearest neighbors using 1 neighbor and ward linkage. Finally, default parameters were used for Gaussian mixture model. These parameters were found by searching over a wide range of values such that the adjusted mutual information was maximized against the true labels. Parameter tuning for all methods, including GPSC, was only done on the first seed (14), and all replicates used the same set of parameters (for seeds 15-63). Any parameters not mentioned were left as default as per the scikit-learn package. The code is provided for full details and implementation.



(a) Spatial Domain (b) XY Domain

Fig. 9: GPSC and comparisons to spatial clustering and supervised clustering algorithms for Simulation 1.

Table 1: Adjusted Rand index and adjusted mutual information of different methods against the true labels for Simulation 1, replicated over 50 random seeds reported as mean ± standard deviation for.

METHOD	ARI	AMI	METHOD	ARI	AMI
GPSC	$\mathbf{0.91 \pm 0.27}$	$\mathbf{0.90 \pm 0.27}$	GMM	0.82 ± 0.39	0.82 ± 0.39
K-MEANS	0.00 ± 0.00	0.00 ± 0.00	C-MEANS	0.00 ± 0.00	0.00 ± 0.00
HIER.	0.03 ± 0.03	0.11 ± 0.06	SPAT. HIER.	0.03 ± 0.03	0.12 ± 0.06
DBSCAN	0.10 ± 0.08	0.08 ± 0.06	GDBSCAN	0.09 ± 0.03	0.24 ± 0.04
SPECTRAL	0.03 ± 0.14	0.15 ± 0.12			

D.2. *Simulation 2*

The data used in this simulation takes the form $\{(s_i, x_i, y_i)\}_{i=1}^n$, where $s_i \in \mathbb{R}^2$, the spatial domain, $x_i \in \mathbb{R}^2$, the covariate domain, and $y_i \in \mathbb{R}$, the response domain, for visualization purposes.

In this simulation, both $s_i \in \mathbb{R}^2$ and $x_i \in \mathbb{R}^2$ are generated from independent uniform distributions, where $s_i \sim \text{Unif}(-5, 5)$ and $x_i \sim \text{Unif}(-3, 3)$ component-wise.

After generating the data $\{(s_i, x_i)\}_{i=1}^n$, where $n = 1000$ samples, the domain square is subdivided into again two clusters, the ring and background. Cluster 1 (ring) was made by subsetting all points $\{(s_i, x_i)\}$ within 3.5 but greater than 2 units of the point (0,0) solely in the spatial domain, with the rest forming cluster 2.

For each cluster, y is generated as a nonlinear function of just x_i. For cluster 1, the true nonlinear function is:

$$y = -(x_1)^3.$$

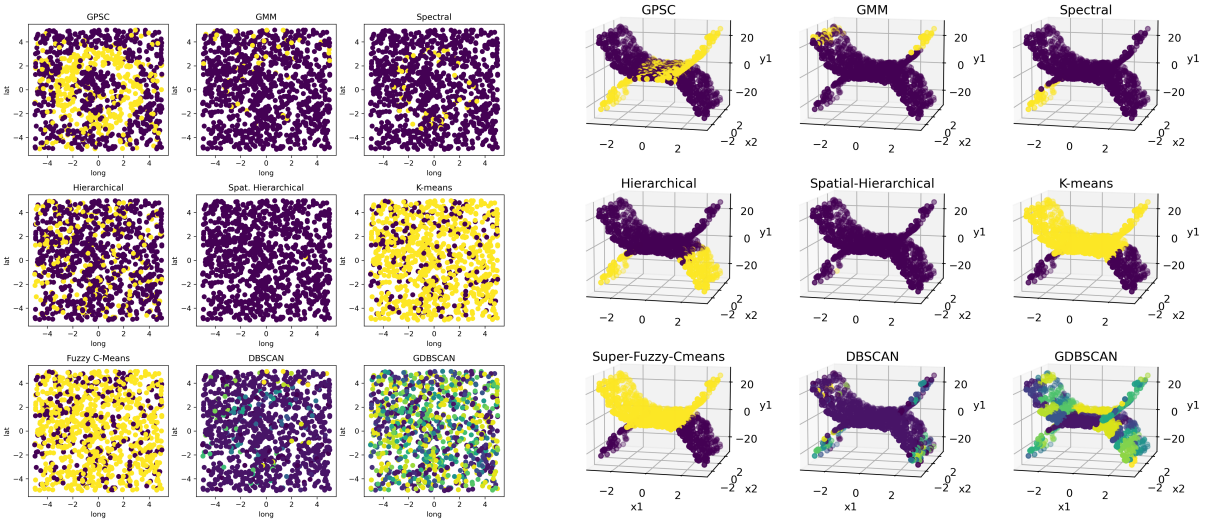
And for cluster 2, the true nonlinear function is:

$$y = (x_1)^3.$$

After the data was generated, the GPSC was applied with the following input: 2 clusters, 20 iterations with early stopping, (note that GP input remains $\{x_i, s_i\}$ even though the true functions generating the clusters are functions only of x), constant bounds $(1e^{-15}, 1e^6)$, length scale bounds $(1e^6, 1e^{15})$, input data. For the GP, again the RBF kernel was used.

Then K-means clustering, spectral clustering, hierarchical clustering with Ward linkage, and DBSCAN was also applied. For spectral clustering, the affinity matrix was generated with using nearest-neighbors set to 5. For DBSCAN, the maximum distance was set to 3, and minimum number of samples set to 3. The full set of $\{x_i, s_i, y_i\}$ as a vector was input into each algorithm along with $L = 2$ clusters where relevant. Any parameters not mentioned were left as default as per the scikit-learn package. For spectral clustering, all neighbors between 1 and 50 were tested by comparing the adjusted mutual information scores. For DBSCAN, the maximum distance distance was tested between 1 and 100, and for each distance, the minimum samples were tested between 1 and 300, again by adjusted mutual information scores.

For supervised C-means clustering, the supervised labels were set according to the y domain with otherwise default parameters. For GDBSCAN, the covariate distance threshold was set to 3, spatial distance threshold set to 13, and minimum set to 0. Finally, for the spatialized hierarchical clustering method, the connectivity matrix was specified using k-nearest neighbors using 11 neighbors and ward linkage. For Gaussian mixture model, the default parameters were used. These parameters were found by searching over a wide range of values such that the adjusted mutual information was maximized against the true labels. Parameter tuning for all methods, including GPSC, was only done on the first seed (14), and all replicates used the same set of parameters (for seeds 15-63). Any parameters not mentioned were left as default as per the scikit-learn package. The code is provided for full details and implementation.



(a) Spatial Domain (b) XY Domain

Fig. 10: GPSC and comparisons to spatial clustering and supervised clustering algorithms for Simulation 2.

Table 2: Adjusted Rand index and adjusted mutual information of different methods against the true labels for Simulation 2, replicated over 50 random seeds reported as mean ± standard deviation.

METHOD	ARI	AMI	METHOD	ARI	AMI
GPSC	$\mathbf{0.38 \pm 0.13}$	$\mathbf{0.28 \pm 0.10}$	GMM	0.11 ± 0.12	0.07 ± 0.11
K-MEANS	0.00 ± 0.01	0.00 ± 0.00	C-MEANS	0.00 ± 0.01	0.00 ± 0.00
HIER.	0.00 ± 0.00	0.00 ± 0.00	SPAT. HIER.	0.00 ± 0.00	0.00 ± 0.00
DBSCAN	0.08 ± 0.05	0.10 ± 0.02	GDBSCAN	0.02 ± 0.01	0.17 ± 0.01
SPECTRAL	0.02 ± 0.09	0.06 ± 0.06			

D.3. *Simulation 3 - Noisy Cluster Results*

The data used in this simulation takes the form $\{(s_i, x_i, y_i)\}_{i=1}^n$, where $s_i \in \mathbb{R}^2$, the spatial domain, $x_i \in \mathbb{R}^2$, the covariate domain, and $y_i \in \mathbb{R}$, the response domain, for visualization purposes. In this simulation, both $s_i \in \mathbb{R}^2$ and $x_i \in \mathbb{R}^2$ are generated from independent uniform distributions, where $s_i \sim \text{Unif}(-5, 5)$, $x_1 \sim \text{Unif}(-6, 6)$, $x_2 \sim \text{Unif}(-2, 4)$ component-wise.

After generating the data $\{(s_i, x_i)\}_{i=1}^n$, where $n = 1000$ samples, the domain square is subdivided into three clusters, the sun shape cluster, moon shape cluster, and the rest region. Cluster 1 was made by subsetting all points $\{(s_i, x_i)\}$ within 2.5 units of the point (-2.2, 2.2) solely in the spatial domain. Cluster 2 was made subsetting all points $\{(s_i, x_i)\}$ within 3 units of (1.8, -1.8) and further than 2 units apart from (1, -1), again solely in the spatial domain, with the remaining points forming cluster 3.

For each cluster, y is generated as a function of just x_i with independent Gaussian distributed noise $\epsilon \sim N(0, \sigma^2)$. For cluster 1, the true nonlinear function is:

$$y = 40x_1^2 - 400 + \epsilon.$$

For cluster 2, the true nonlinear function is:

$$y = -(x_1 - 8)^3 + \epsilon.$$

And for cluster 3, the true nonlinear function is:

$$y = (x_1 + 8)^3 - 20 + \epsilon.$$

After the data was generated, the GPSC was applied with the following input: 3 clusters, 40 iterations with early stopping, (note that GP input remains $\{x_i, s_i\}$ even though the true functions generating the clusters are functions only of x), constant bounds $(1e^{-15}, 1e^4)$, length scale bounds $(1e^6, 1e^{15})$, input data. For the GP, again the RBF kernel was used. Note that here, two forms of GPSC were used. First, standard GPSC was performed with results shown in the table. Then, GPSC with $\lambda = 75$ was used, and it was shown that the GPSC model was better able to find the clusters with this spatial penalty.

Then K-means clustering, spectral clustering, hierarchical clustering with Ward linkage, and DBSCAN was also applied. For spectral clustering, the affinity matrix was generated with using nearest-neighbors set to 12. For DBSCAN, the maximum distance was set to 41, and minimum number of samples set to 27. The full set of $\{x_i, s_i, y_i\}$ as a vector was input into each algorithm along with $k = 3$ clusters where relevant. Any parameters not mentioned were left as default as per the scikit-learn package. For supervised C-means clustering, the supervised labels were set according to the y domain with otherwise default parameters. For GDBSCAN, the covariate distance threshold was set to 675, spatial distance threshold set to 5, and minimum set to 0. For the spatialized hierarchical clustering method, the connectivity matrix was specified using k-nearest neighbors using 9 neighbors and ward linkage. Finally for Gaussian mixture model, the default parameters were used. These parameters were found by searching over a wide range of values such that the adjusted mutual information was maximized against the true labels. Parameter tuning for all methods, including GPSC, was only done on the first seed (14), and all replicates used the same set of parameters (for seeds 15-63). Any parameters not mentioned were left as default as per the scikit-learn package.

D.3.1. $\sigma^2 = 2$

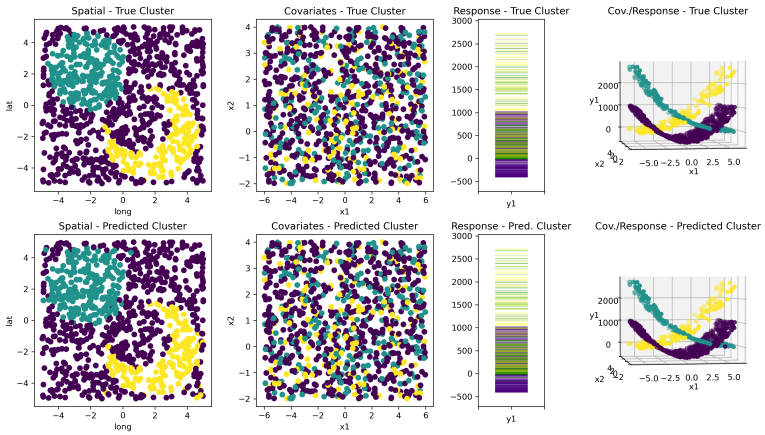
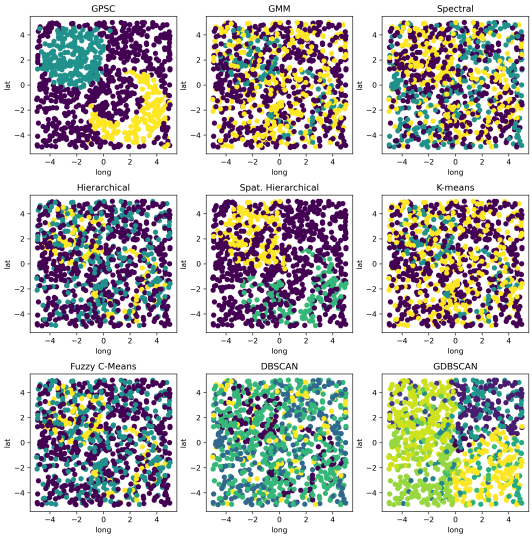
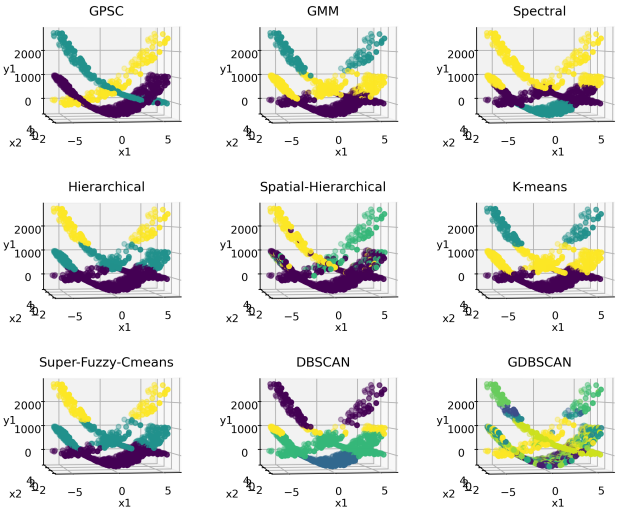


Fig. 11: GPSC results for Simulation 3, $\sigma^2 = 2, L = 3$, colored by cluster and separated by data domain as in previous simulation. The first row indicates ground truth with results from GPSC in the second.



(a) Spatial Domain



(b) XY Domain

Fig. 12: GPSC and comparisons to spatial clustering and supervised clustering algorithms for Simulation 3, $\sigma^2 = 2, L = 3$.

Table 3: Adjusted Rand index and adjusted mutual information of different methods against the true labels for for Simulation 3, $\sigma^2 = 2, L = 3$, replicated over 50 random seeds reported as mean ± standard deviation.

METHOD	ARI	AMI	METHOD	ARI	AMI
GPSC	$\mathbf{0.72 \pm 0.27}$	$\mathbf{0.70 \pm 0.24}$	GMM	0.16 ± 0.02	0.14 ± 0.02
K-MEANS	0.17 ± 0.02	0.13 ± 0.01	C-MEANS	0.16 ± 0.02	0.13 ± 0.01
HIER.	0.17 ± 0.03	0.13 ± 0.03	SPAT. HIER.	0.16 ± 0.11	0.17 ± 0.08
DBSCAN	0.22 ± 0.04	0.15 ± 0.03	GDBSCAN	0.10 ± 0.02	0.24 ± 0.04
SPECTRAL	0.08 ± 0.02	0.16 ± 0.02			

D.3.2. $\sigma^2 = 50$

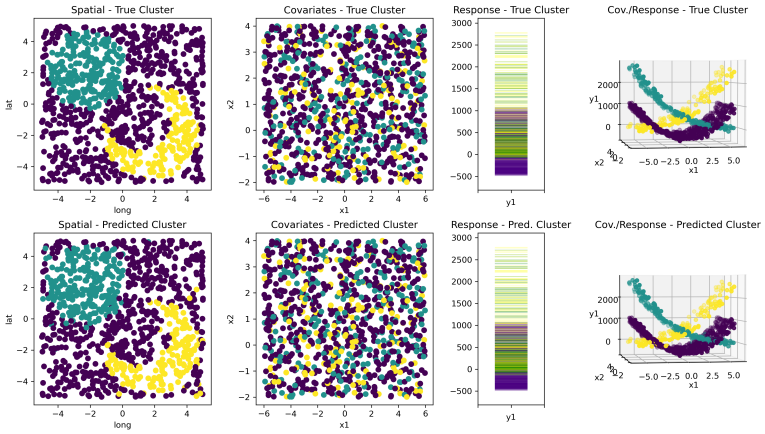
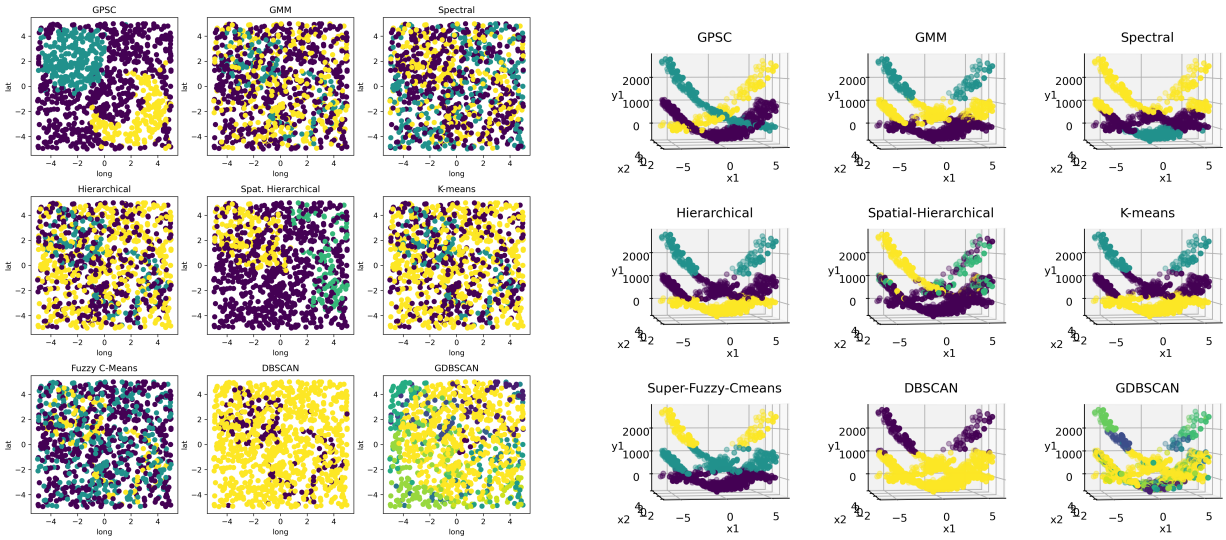


Fig. 13: GPSC results for for Simulation 3, $\sigma^2 = 50, L = 3$, colored by cluster and separated by data domain as in previous simulation. The first row indicates ground truth with results from GPSC in the second.



(a) Spatial Domain

(b) XY Domain

Fig. 14: GPSC and comparisons to spatial clustering and supervised clustering algorithms for Simulation 3, $\sigma^2 = 50, L = 3$.

Table 4: Adjusted Rand index and adjusted mutual information of different methods against the true labels for for Simulation 3, $\sigma^2 = 50, L = 3$, replicated over 50 random seeds reported as mean ± standard deviation.

METHOD	ARI	AMI	METHOD	ARI	AMI
GPSC	$\mathbf{0.73 \pm 0.25}$	$\mathbf{0.71 \pm 0.22}$	GMM	0.16 ± 0.02	0.13 ± 0.03
K-MEANS	0.17 ± 0.02	0.13 ± 0.01	C-MEANS	0.16 ± 0.02	0.13 ± 0.01
HIER.	0.17 ± 0.03	0.13 ± 0.03	SPAT. HIER.	0.17 ± 0.10	0.18 ± 0.08
DBSCAN	0.23 ± 0.03	0.13 ± 0.03	GDBSCAN	0.09 ± 0.03	0.23 ± 0.03
SPECTRAL	0.08 ± 0.02	0.16 ± 0.01			

D.3.3. $\sigma^2 = 100$

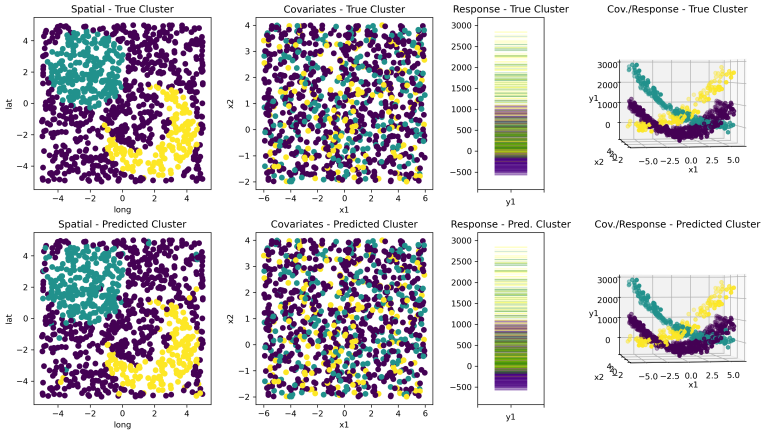
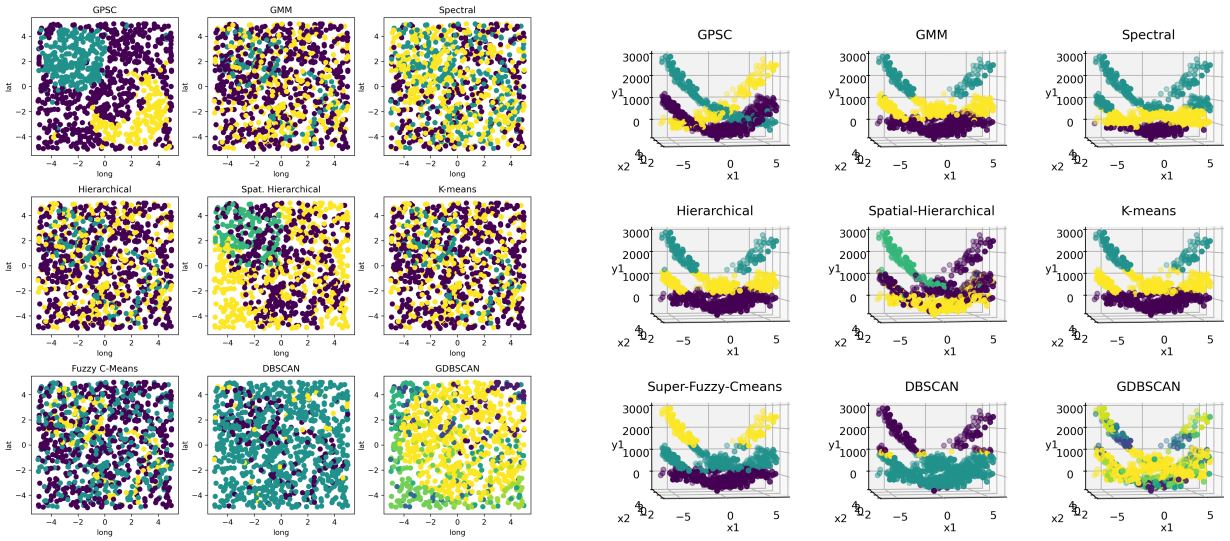


Fig. 15: GPSC results for for Simulation 3, $\sigma^2 = 100, L = 3$, colored by cluster and separated by data domain as in previous simulation. The first row indicates ground truth with results from GPSC in the second.



(a) Spatial Domain (b) XY Domain

Fig. 16: GPSC and comparisons to spatial clustering and supervised clustering algorithms for Simulation 3, $\sigma^2 = 100, L = 3$.

Table 5: Adjusted Rand index and adjusted mutual information of different methods against the true labels for for Simulation 3, $\sigma^2 = 100, L = 3$, replicated over 50 random seeds reported as mean ± standard deviation.

METHOD	ARI	AMI	METHOD	ARI	AMI
GPSC	$\mathbf{0.56 \pm 0.26}$	$\mathbf{0.55 \pm 0.23}$	GMM	0.15 ± 0.02	0.13 ± 0.03
K-MEANS	0.17 ± 0.02	0.13 ± 0.01	C-MEANS	0.16 ± 0.02	0.13 ± 0.01
HIER.	0.16 ± 0.04	0.13 ± 0.03	SPAT. HIER.	0.15 ± 0.09	0.17 ± 0.06
DBSCAN	0.21 ± 0.02	0.10 ± 0.02	GDBSCAN	0.09 ± 0.03	0.23 ± 0.04
SPECTRAL	0.07 ± 0.02	0.14 ± 0.01			

D.3.4. $\sigma^2 = 200$

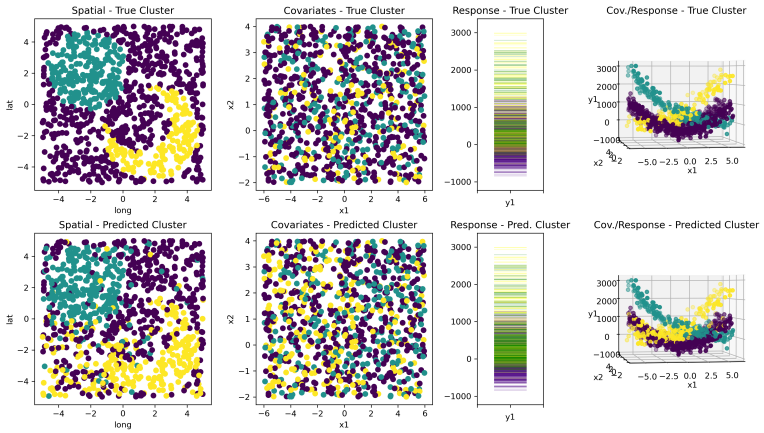
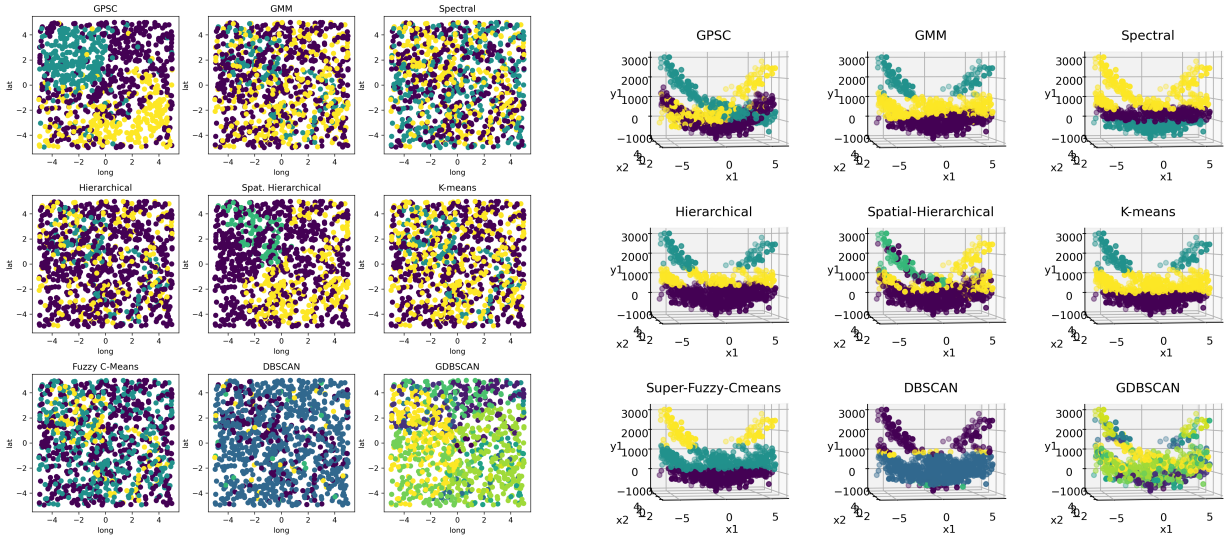


Fig. 17: GPSC results for for Simulation 3, $\sigma^2 = 200, L = 3$, colored by cluster and separated by data domain as in previous simulation. The first row indicates ground truth with results from GPSC in the second.



(a) Spatial Domain (b) XY Domain

Fig. 18: GPSC and comparisons to spatial clustering and supervised clustering algorithms for Simulation 3, $\sigma^2 = 200, L = 3$.

Table 6: Adjusted Rand index and adjusted mutual information of different methods against the true labels for for Simulation 3, $\sigma^2 = 200, L = 3$, replicated over 50 random seeds reported as mean ± standard deviation.

METHOD	ARI	AMI	METHOD	ARI	AMI
GPSC	$\mathbf{0.33 \pm 0.17}$	$\mathbf{0.33 \pm 0.15}$	GMM	0.16 ± 0.02	0.13 ± 0.02
K-MEANS	0.17 ± 0.02	0.13 ± 0.01	C-MEANS	0.16 ± 0.02	0.13 ± 0.01
HIER.	0.16 ± 0.04	0.12 ± 0.02	SPAT. HIER.	0.15 ± 0.08	0.1 ± 0.06
DBSCAN	0.16 ± 0.02	0.06 ± 0.01	GDBSCAN	0.08 ± 0.02	0.22 ± 0.03
SPECTRAL	0.06 ± 0.02	0.10 ± 0.01			