Supervised Stratified Subsampling for Regression Problems

Ming-Chung Chang

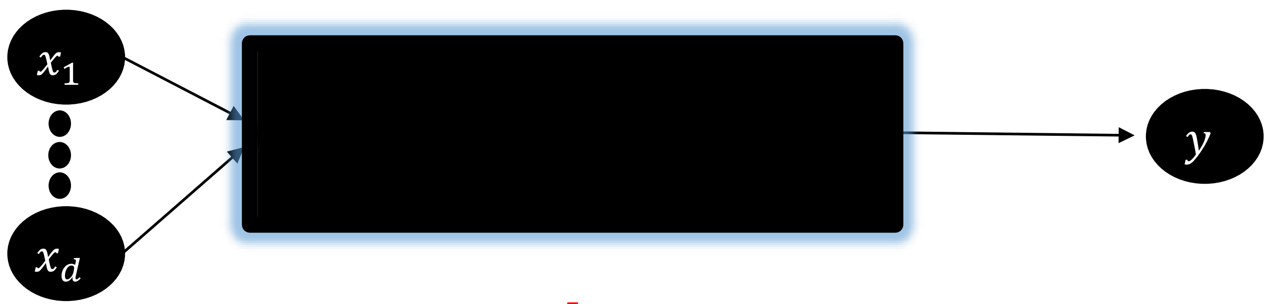
Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

June 18, 2024

JOINT RESEARCH CONFERENCE 2024

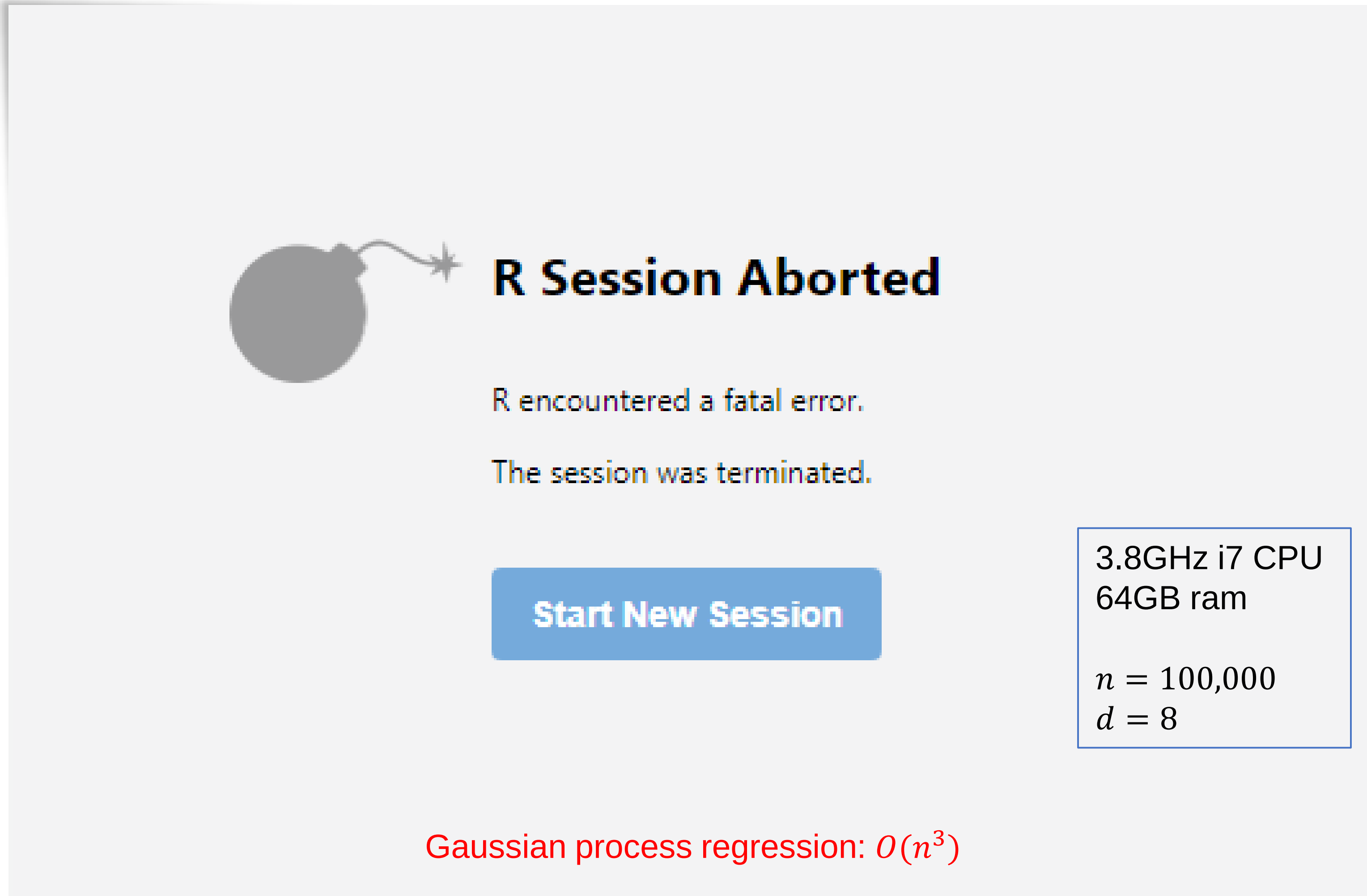
Regression problem / Supervised learning

Unveil the **blackbox** among features/variables



Statistical model fitting
(*Linear model*, *Gaussian process regression*, etc.)

$$f(\mathbf{x}) = -\sum_{i=1}^{d} \sin(x_i) \sin^{2m}\left(\frac{i x_i^2}{\pi}\right)$$



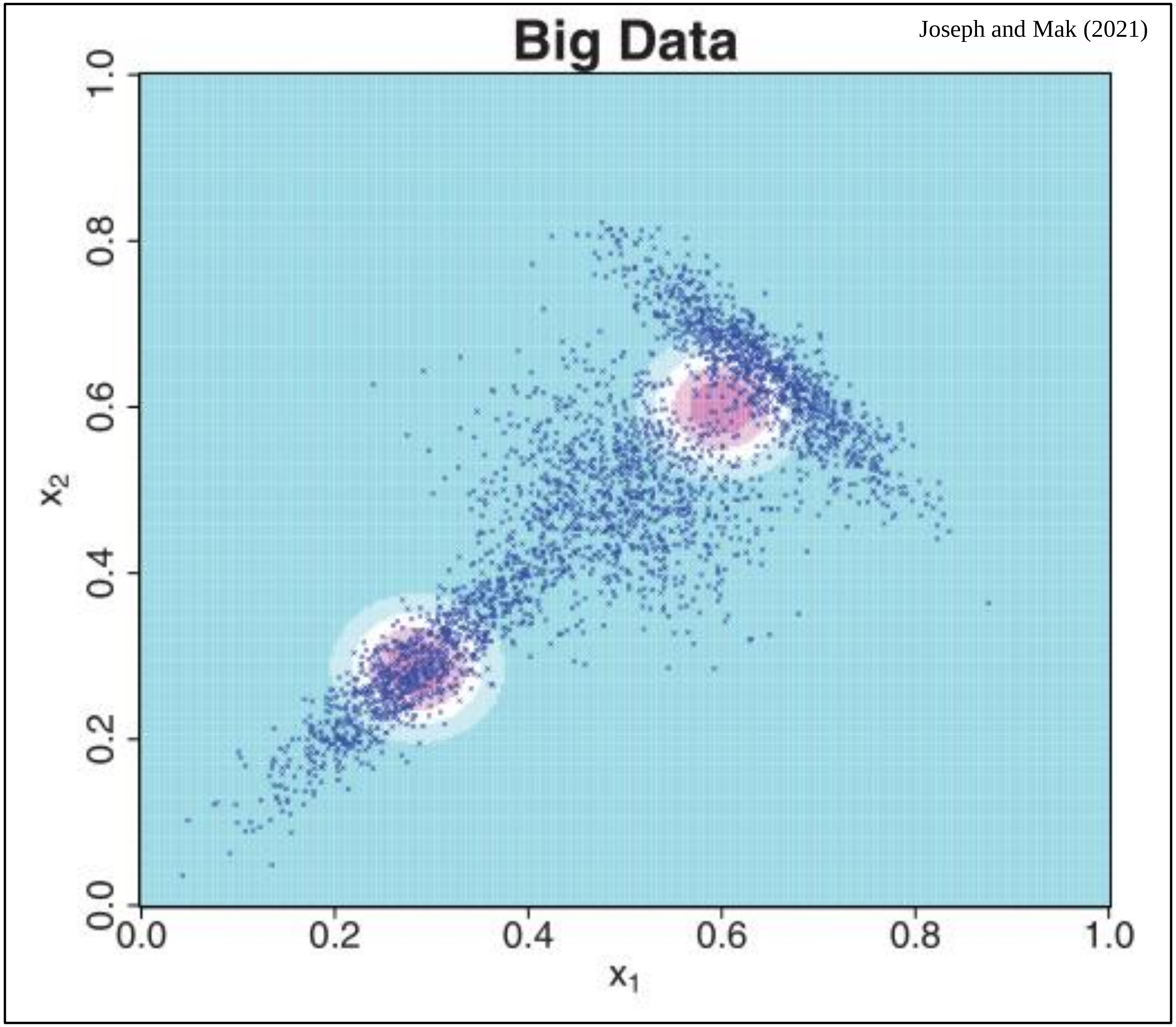
3.8GHz i7 CPU
64GB ram

$n = 100{,}000$
$d = 8$

Gaussian process regression: $O(n^3)$

Outline

- Idea
- The Proposed Method
- Numerical Examples
- Conclusion


Big Data
Joseph and Mak (2021)
1.0
0.8
0.6
0.4
0.2
0.0
x2
0.0
0.2
0.4
0.6
0.8
1.0
x1

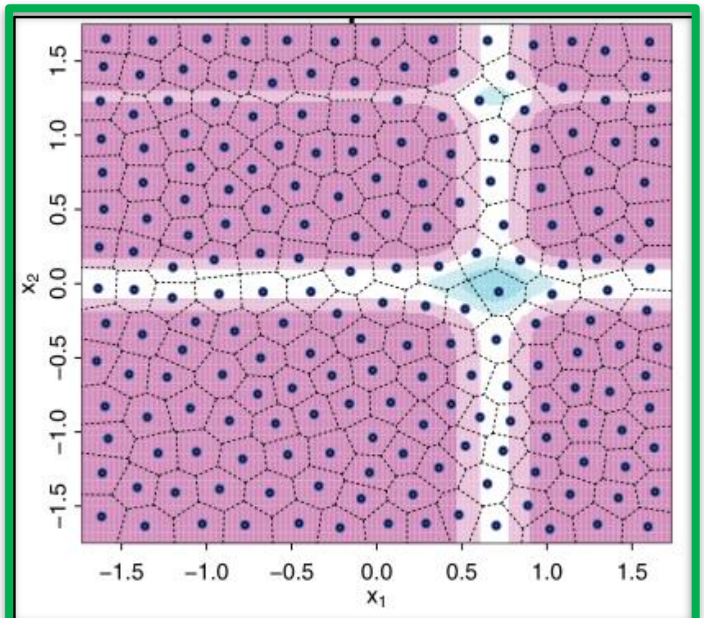
Literature Review

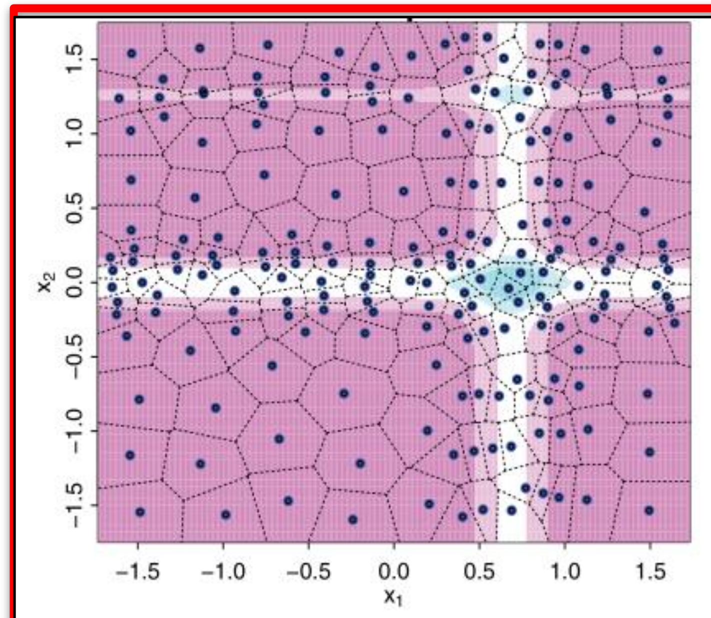
How to cite this article: Joseph VR, Mak S. Supervised compression of big data. *Stat Anal Data Min: The ASA Data Sci Journal*. 2021;14:217–229. https://doi.org/10.1002/sam.11508

- Contour plot: Michaelwicz function in 2 dimensions
- Choose 200 points from 20,000 points (~ $U(0,1)^2$)
- 200 points → 200 nearest regions (Voronoi regions)

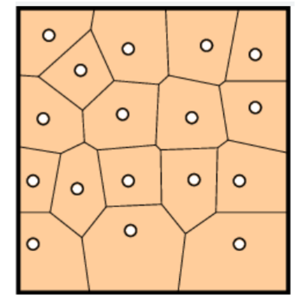
K-means clustering on X

Supercompress (Joseph and Mak, 2021)





Methodology



- Idea:
  - Input space ➔ **response-homogeneous** regions (strata)
  - Sampling from every region (stratum)

- Partitioning estimate is relevant
  - Nonparametric regression estimate
  - Aka *Regressogram*, *Regression histogram*

$$\hat{f}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} y_i \mathrm{I}(\boldsymbol{x}_i \in A(\boldsymbol{x}))}{\sum_{i=1}^{n} \mathrm{I}(\boldsymbol{x}_i \in A(\boldsymbol{x}))}$$

- What are good **response-homogeneous (R-H)** strata?

Partitioning estimate converges to $f(\boldsymbol{x})$

Methodology

- Data: $(\boldsymbol{x}_i, y_i)$ iid $\sim F$, $i = 1, \ldots, n$, with $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$

- Generate k clusters on the Y-space. Then form k **R-H strata** on the X-space.
  - Clusters (Y-space): $\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_k$
  - R-H strata (X-space): $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k$

- $\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_k$ are constructed by minimizing

$$\sum_{l=1}^{k} \int_{\mathcal{J}_l} \{y - \mathrm{E}(Y|Y \in \mathcal{J}_l)\}^2 g(y) dy$$

  $g(y)$: marginal density function of the response

- $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k$ are constructed by $f^{-1}(\mathcal{J}_l)$s (inverse images of $\mathcal{J}_l$)

Methodology

- Assume: **(i)** $f(\boldsymbol{x})$ is bounded; **(ii)** $\mathrm{Var}(Y|x)$ is bounded; **(iii)** $g(y)$ is bounded, defined on a compact support, and has 1^{st} to 4^{th} bounded derivatives. Then, the MISE for the **partitioning estimate** $\hat{f}(\boldsymbol{x})$ is:

$$\mathrm{E}\left\{\int \left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 \mu(d\boldsymbol{x})\right\} = \mathrm{O}\left(\frac{k}{n} + \frac{1}{k^2}\right)$$

- Suggest $k = n^{\frac{1}{3}}$ ➔ Convergence rate: $\mathrm{O}\left(n^{-\frac{2}{3}}\right)$

$\log_{10}(n)$	2	3	4	5	6	7	8	9	10
k	4	9	21	46	99	215	464	999	2154

Methodology

- $f(\boldsymbol{x})$ is unknown

- $\mathcal{J}_1, \mathcal{J}_2, \dots, \mathcal{J}_k$ are constructed by the sample *k*-means clustering (Pollard, 1981):

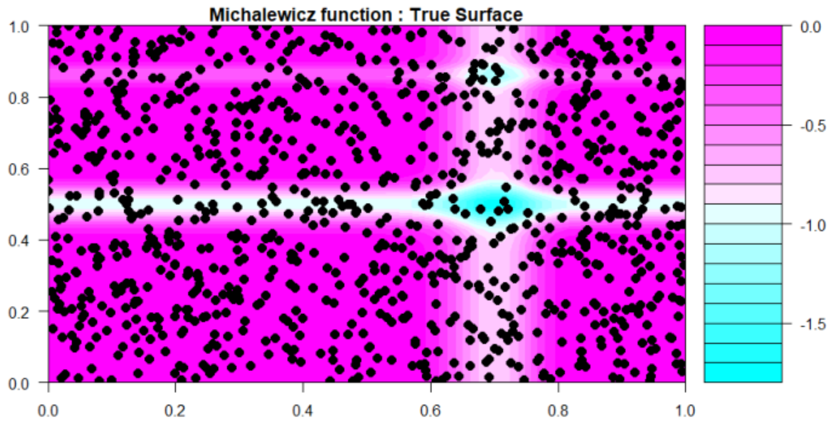
$$\text{Minimize } \sum_{i=1}^{n} \min_{1 \le l \le k} |y_i - c(\mathcal{J}_l)|^2 \text{ over } \mathcal{J}_1, \mathcal{J}_2, \dots, \mathcal{J}_k$$

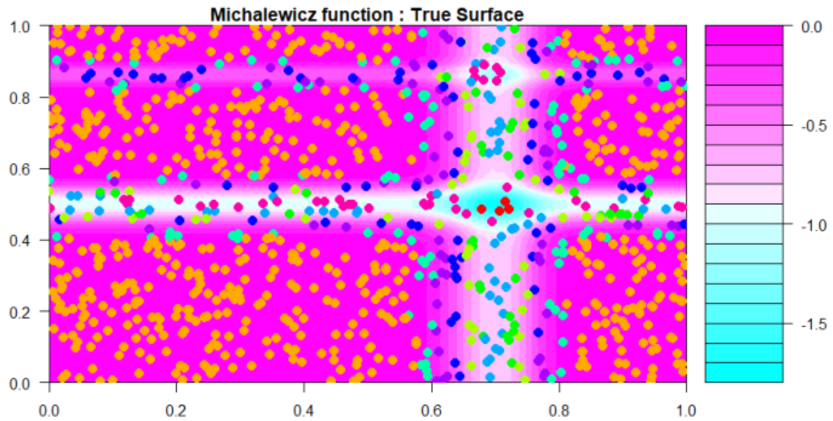
- $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_k$ are constructed by

$$\mathcal{A}_l = \left\{ \mathbf{x} \in \mathcal{X} : \min_{1 \le i \le n : Y_i \in \mathcal{I}_l} |\mathbf{x} - \mathbf{x}_i| \le |\mathbf{x} - \mathbf{x}_j| \text{ for all } Y_j \notin \mathcal{I}_l \right\}$$

Methodology

- Michalewicz function $(d = 2)$
    - $n = 1000$
    - $k = 9$





Methodology: SSS

- We refer to the proposed method as
  Supervised **S**tratified **S**ubsampling (**SSS**)
  - Decide subdata size n_S
  - Randomly select n_j data points in $\mathcal{A}_j$ **without replacement**
  - Repeat B times and aggregate the predictions

- Optimal allocation of n_js: $n_j \propto$ MISE due to $\mathcal{A}_j$

- Using partitioning estimate, the usual bagged prediction $\bar{\hat{f}}_S(\boldsymbol{x})$ is **unbiased** for $\hat{f}(\boldsymbol{x})$, and

$$\mathrm{E}\left\{\int \left(\bar{\hat{f}}_S(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 d\boldsymbol{x}\right\} = \frac{1}{B}\mathrm{E}\left\{\int \mathrm{Var}_S(\hat{f}_S(\boldsymbol{x}))d\boldsymbol{x}\right\} + \mathrm{E}\left\{\int \left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2 d\boldsymbol{x}\right\}$$

Methodology: SSS

- Two aggregation methods for $\bar{\hat{f}}_S(\boldsymbol{x})$:

$$\hat{f}_{\mathcal{S}_b}(\mathbf{x}) = f(\mathbf{x}) + \gamma_b$$

1 $$\overline{\hat{f}}_{\text{GLS}}(\mathbf{x}) = \left\{\mathbf{1}_B^\intercal \mathbf{\Sigma}^{-1} \mathbf{1}_B\right\}^{-1} \mathbf{1}_B^\intercal \mathbf{\Sigma}^{-1} \hat{\mathbf{f}}(\mathbf{x})$$

2 $$\overline{\hat{f}}_{\text{OLS}}(\mathbf{x}) = \left\{\mathbf{1}_B^\intercal \mathbf{1}_B\right\}^{-1} \mathbf{1}_B^\intercal \hat{\mathbf{f}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{\mathcal{S}_b}(\mathbf{x})$$

rSSS: robustified SSS

- **Algorithm**
  - Apply *k-means on Y* with $k = \lfloor n^{1/3} \rfloor$
  - Form the clusters $\{y_{i1}, \dots, y_{ik_i} : i = 1, \dots, \lfloor n^{1/3} \rfloor\}$
  - Form the sets $\{\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{ik_i} : i = 1, \dots, \lfloor n^{1/3} \rfloor\}$
  - Form the nearest regions on the X-space using $\{\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{ik_i} : i = 1, \dots, \lfloor n^{1/3} \rfloor\}$
    - If #(some region)>10, then apply *k-means on X* to that region with $k = \min\{k^*: \frac{\text{SS}_{\text{between}}}{\text{SS}_{\text{within}}} > 0.95\}$
  - Form strata $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_{k'}$ $(k' \geq k)$
  - Randomly sample data points in each $\mathcal{A}_l$
  - Repeat B times. Then aggregate

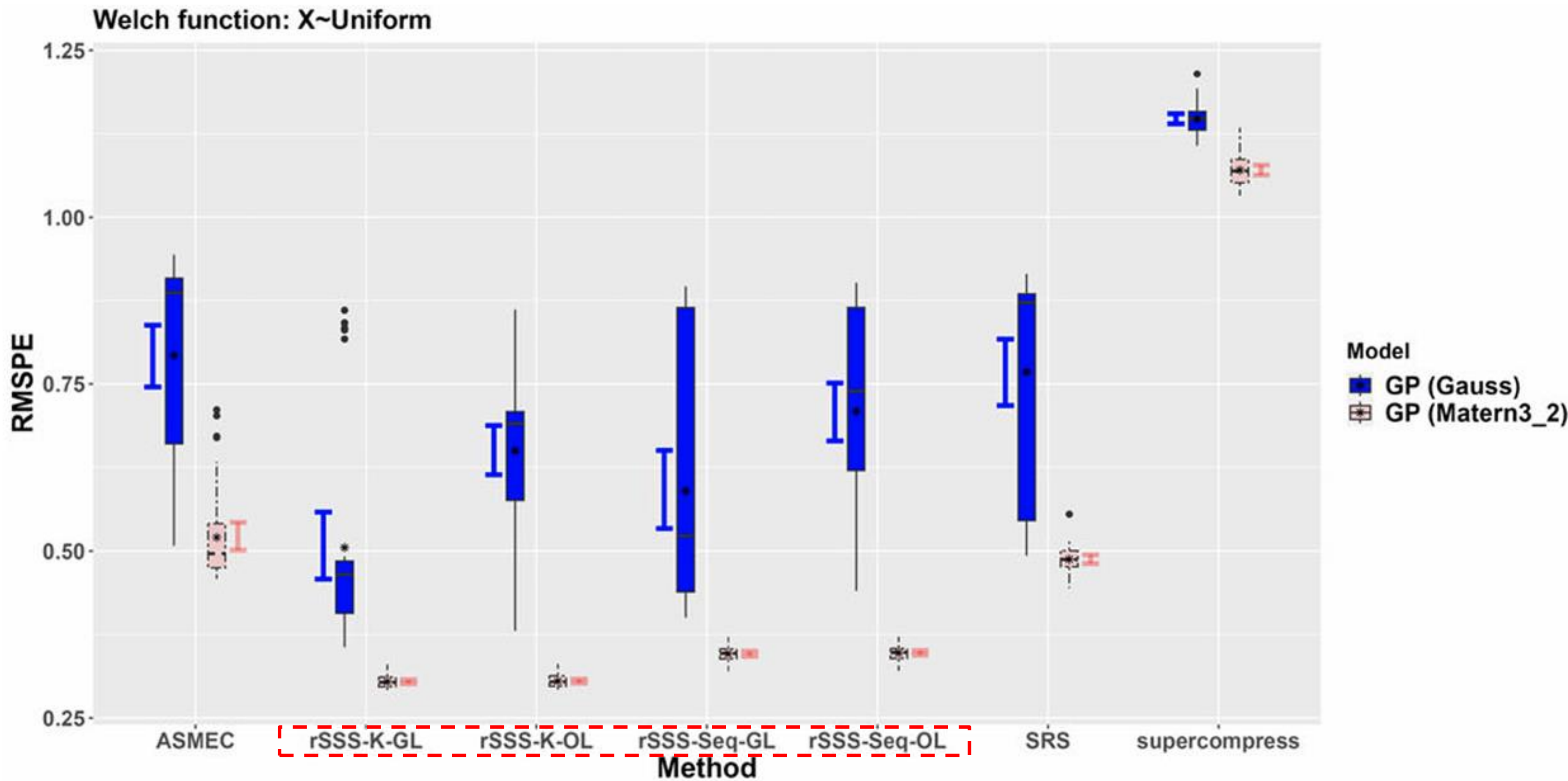
Simulation: Welch function

- $d = 20$
- Distribution of X: (1) uniform; (2) mixture normal (3) T
- 100,000 training + 10,000 testing
- Subdata size: 1,000
- SNR = 5
- $B = 5$

40 replications → 40 RMSPEs

- Methods: **rSSS-K-GL/OL**, **rSSS-Seq-GL/OL**, supercompress, ASMEC, SRS

- Models
  - Gaussian process regression (mleHomGP)
    - Gaussian correlation function
    - Matern32 function
  - k-NN ($k = 1$ and $k = 2$, knn.reg)


Welch function: X~Uniform
RMSPE
1.25
1.00
0.75
0.50
0.25
ASMEC
rSSS-K-GL
rSSS-K-OL
rSSS-Seq-GL
rSSS-Seq-OL
SRS
supercompress
Method
Model
GP (Gauss)
GP (Matern3_2)

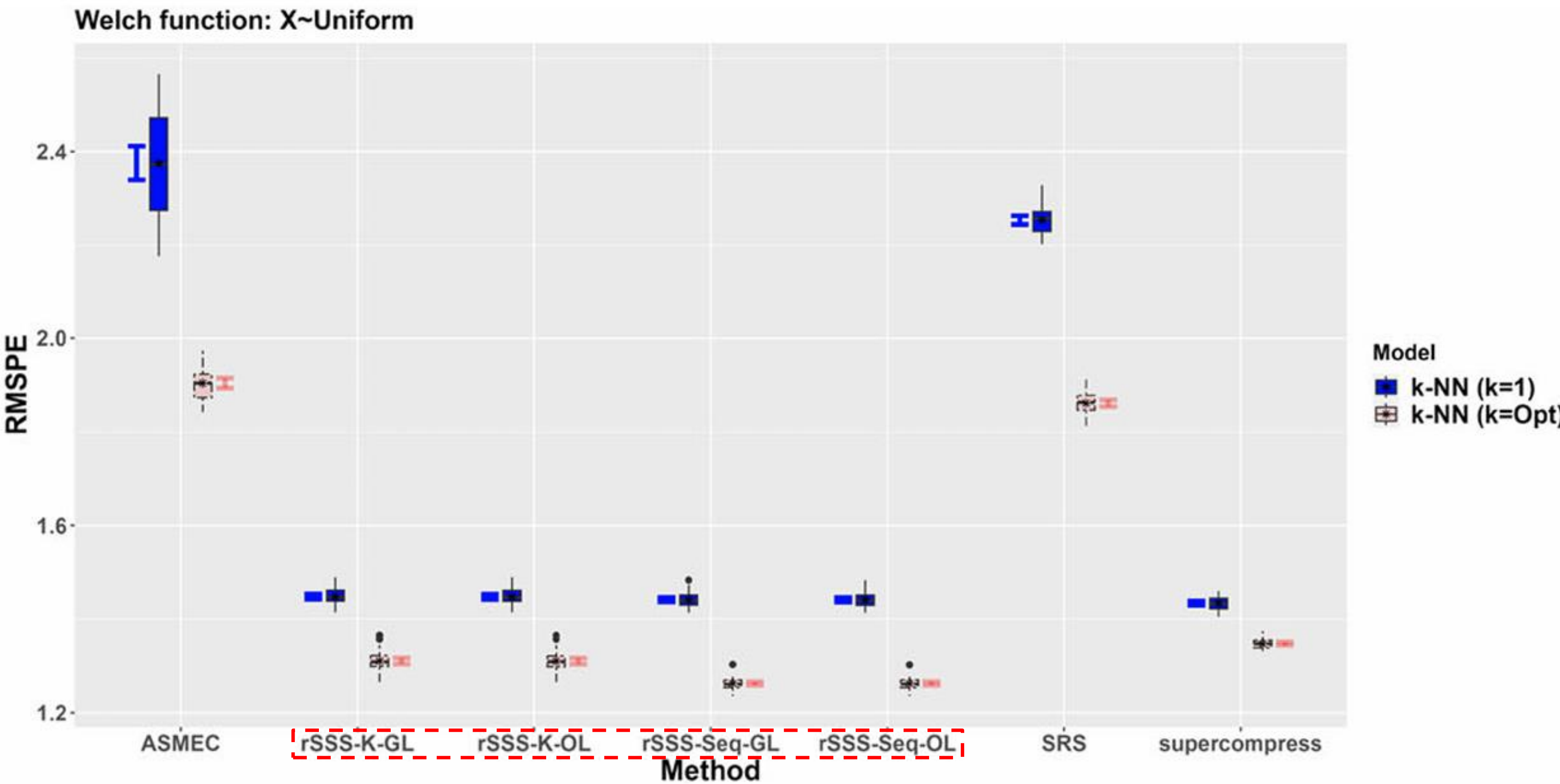

Welch function: X~Uniform
RMSPE
2.4
2.0
1.6
1.2
ASMEC
rSSS-K-GL
rSSS-K-OL
rSSS-Seq-GL
rSSS-Seq-OL
SRS
supercompress
Method
Model
k-NN (k=1)
k-NN (k=Opt)

Table 5. Medians of the 40 RMSPEs (bold for the minimum): Welch function under (X-1).

	GP (Gauss)	GP (Matern)	k-NN (k=1)	k-NN (k=Opt)
ASMEC	0.8861	0.4962	2.373	1.904
SRS	0.8715	0.4875	2.257	1.863
rSSS-Kmeans-GLS	**0.4647**	0.3036	1.447	1.308
rSSS-Kmeans-OLS	0.6906	**0.3043**	1.446	1.309
rSSS-Seq-GLS	0.5231	0.3469	1.439	**1.261**
rSSS-Seq-OLS	0.7394	0.3483	1.439	**1.261**
supercompress	1.1480	1.0700	**1.433**	1.348
Full data	Infeasible	Infeasible	1.934	1.556

Table 6. Average computation time (in minutes) over the 40 replicates under the three mechanisms.

(X-1)/(X-2)/(X-3)	rSSS-Kmeans	rSSS-Seq	supercompress	ASMEC
Piston	8.27/6.09/4.07	8.19/6.15/4.05	8.27/8.16/6.45	1.70/1.27/1.00
Borehole	7.87/6.87/6.95	8.35/6.93/7.16	9.56/9.21/11.82	2.01/1.33/1.68
Wing Weight	7.30/7.44/8.09	7.51/7.78/7.87	14.91/13.64/18.62	1.83/1.54/2.20
Welch	10.70/12.50/12.68	9.74/12.85/10.04	40.84/19.18/35.45	3.33/3.18/3.67

NOTE: Piston ($d = 7$); Borehole ($d = 8$); Wing Weight ($d = 10$); Welch ($d = 20$)

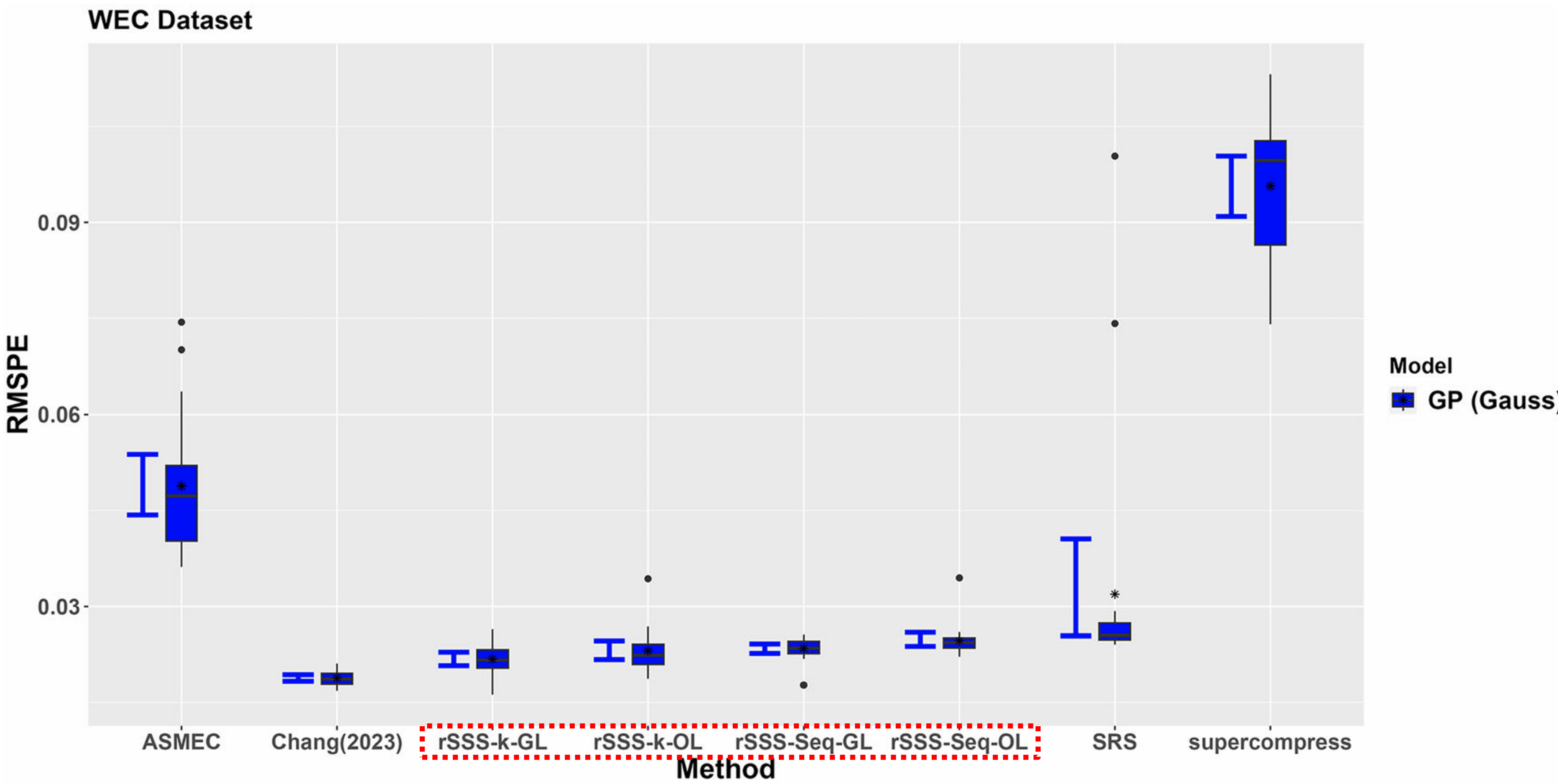
Desktop computer with a 3.20GHz Intel Corei9 CPU and 128GB of RAM

WEC Dataset

- The Wave Energy Converters (WEC) dataset, provided by UCI Machine Learning Repository (Dua and Graff, 2019)
  - Y: total power output
  - X: 32 location variables and 16 absorbed power variables ($d = 48$)
  - 288,000 = 252,000 for training + 36,000 for testing (divided by SRS)
  - Subdata size: 1,000
  - $B = 5$

40 replications ➔ 40 RMSPEs

- Methods: **<u>rSSS-K-GL/OL</u>**, **<u>rSSS-Seq-GL/OL</u>**, <u>supercompress</u>, <u>ASMEC</u>, <u>SRS</u>, <u>Chang(2023)</u>
- Models
  - Gaussian process regression (`mleHomGP`)
    - Gaussian correlation function
  - k-NN ($k = 1$ and $k = 5$, `knn.reg`)


WEC Dataset
RMSPE
0.09
0.06
0.03
ASMEC
Chang(2023)
rSSS-k-GL
rSSS-k-OL
rSSS-Seq-GL
rSSS-Seq-OL
SRS
supercompress
Method
Model
GP (Gauss)

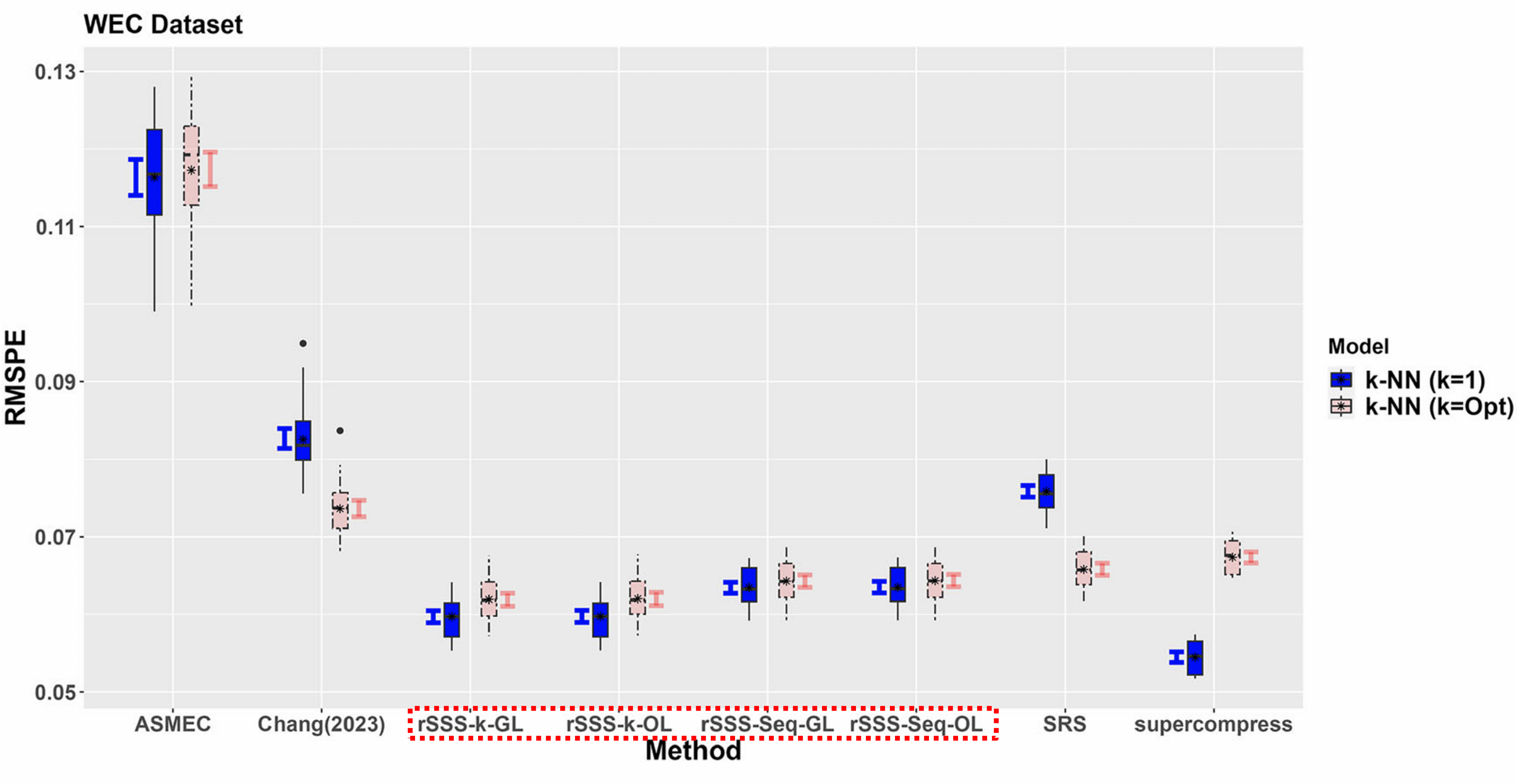

WEC Dataset
0.13
0.11
0.09
0.07
0.05
RMSPE
ASMEC
Chang(2023)
rSSS-k-GL
rSSS-k-OL
rSSS-Seq-GL
rSSS-Seq-OL
SRS
supercompress
Method
Model
k-NN (k=1)
k-NN (k=Opt)

Table 8. Medians of the 40 RMSPEs (bold for the minimum): WEC data.

	GP (Gauss)	k-NN (k=1)	k-NN (k=Opt)
ASMEC 8.59 minutes	0.04767	0.11672	0.11921
Chang(2023) 30.57 minutes	**0.01889**	0.08165	0.07357
SRS	0.02594	0.07549	0.06566
rSSS-Kmeans-GLS	0.02211	0.05973	**0.06169**
rSSS-Kmeans-OLS	0.02294	0.05974	0.06172
rSSS-Seq-GLS 19.95 minutes	0.02417	0.06317	0.06411
rSSS-Seq-OLS	0.02457	0.06322	0.06420
supercompress 75.20 minutes	0.10003	**0.05451**	0.06754

Desktop computer with a 3.20GHz Intel Corei9 CPU and 128GB of RAM

Conclusion

- Aim at a **model-free** and **-robust** subsampling method

- Propose **S**upervised **S**tratified **S**ubsampling (**SSS**)
  - Form response-homogeneous (R-H) strata
  - Sampling from every R-H stratum

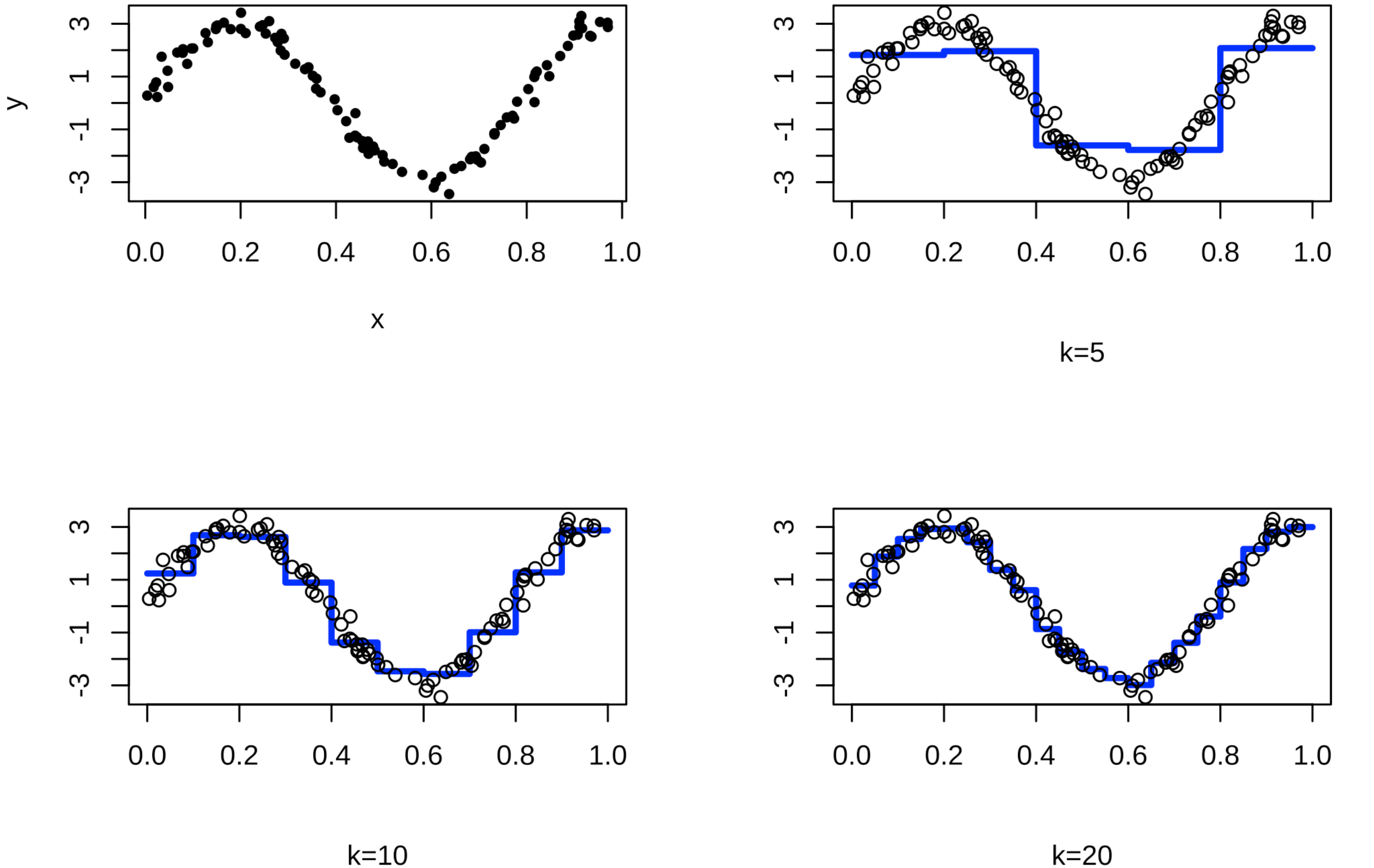
- Large B ➔ High computational cost ($B = 5$ seems fine)

- Observations from the numerical studies:
  - *Chang(2023)* not good for k-NN
  - *supercompress* usually better for 1-NN (non-smooth model)
  - *SSS* seems more robust

Thank you for your attention

Ming-Chung Chang (2024): Supervised Stratified Subsampling for Predictive Analytics, *Journal of Computational and Graphical Statistics*, DOI: 10.1080/10618600.2024.2304075

Appendix


y
x
k=5
k=10
k=20