Spatial Clustering of Citizen Science Data Improves Downstream Species Distribution Models

Supplementary Material

Nahian Ahmed[1], **Mark Roth**[1], **Tyler A. Hallman**[3],
W. Douglas Robinson[2], **Rebecca A. Hutchinson**[1, 2]

[1]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, USA
[2]Department of Fisheries, Wildlife, and Conservation Sciences, Oregon State University, Corvallis, OR 97331, USA
[3]School of Natural Sciences, Bangor University, Bangor LL57 2DG, UK
ahmedna@oregonstate.edu, roth.markh@gmail.com, t.hallman@bangor.ac.uk,
{douglas.robinson, rah}@oregonstate.edu

Species	Abbreviation	Prevalence	Prevalence Level	Habitat	Generalist/ Specialist	Home Range	Territory Size (ha)
Cooper's Hawk	COHA	0.36%	l	e	s	l	1000
Northern Pygmy-Owl	NOOW	0.60%	l	f	s	l	50
Mountain Quail	MOQU	1.04%	l	e	s	l	1000
Bald Eagle	BAEA	1.36%	l	e	g	l	2500
Hammond's Flycatcher	HAFL	1.72%	l	f	s	s	1
Yellow Warbler	YEWA	1.84%	l	e	s	s	0.5
Bushtit	BUTI	2.20%	l	e	g	s	1
Hairy Woodpecker	HAWO	2.64%	l	f	s	m	2.5
Pileated Woodpecker	PIWO	2.64%	l	f	g	l	250
Olive-sided Flycatcher	OLFL	2.80%	l	f	s	l	40
Red-tailed Hawk	REHA	2.84%	m	e	g	l	2300
Brown Creeper	BRCR	4.04%	m	f	s	m	5
Yellow-breasted Chat	YEBCHA	4.57%	m	e	s	s	2
MacGillivray's Warbler	MAWA	6.09%	m	e	s	s	2
Pacific Wren	PAWR	7.17%	m	f	s	m	2.5
Wrentit	WRENTI	9.01%	m	e	s	s	2
Northern Flicker	NOFL	9.29%	m	e	g	m	100
Hermit Warbler	HEWA	10.13%	m	f	s	s	0.5
California Scrub-Jay	CASC	11.25%	m	e	g	m	4
Chestnut-backed Chickadee	CHBCHI	13.14%	m	f	s	s	1
Pacific-slope Flycatcher	PAFL	14.42%	h	f	s	m	2.5
Western Tanager	WETA	15.14%	h	f	g	m	5
Warbling Vireo	WAVI	15.54%	h	f	s	s	1.5
Wilson's Warbler	WIWA	16.18%	h	f	g	s	1
Western Wood-Pewee	WEPE	17.10%	h	f	g	s	2
Spotted Towhee	SPTO	18.94%	h	e	g	s	2
Black-headed Grosbeak	BKHGRO	25.67%	h	f	g	m	3
Swainson's Thrush	SWTH	27.51%	h	f	g	s	2
American Robin	AMRO	28.19%	h	e	g	s	1
Song Sparrow	SOSP	30.88%	h	e	g	s	0.5
American Crow	AMCR	33.04%	h	e	g	l	2500

Table S1: Species, abbreviations, and traits. The prevalence values were calculated from the training checklists of 2017. "l", "m", and "h" refer to low, medium, and high prevalence respectively. Species which reside in early seral and grassland habitats are denoted by "e", and those in forested habitats are denoted by "f". Generalist and specialist species are denoted by "g" and "s" respectively. "s", "m", and "l" refer to small, medium, and large territory sizes respectively. We used these species traits to analyze how interaction groups of these traits and clustering algorithms affect overall performance on downstream task of maximizing AUC of occupancy models. Prevalence rates (class balance) ranges from just 0.36% detections (positives) for the Cooper's Hawk (COHA) and up to 33.04% detections for the American Crow (AMCR).

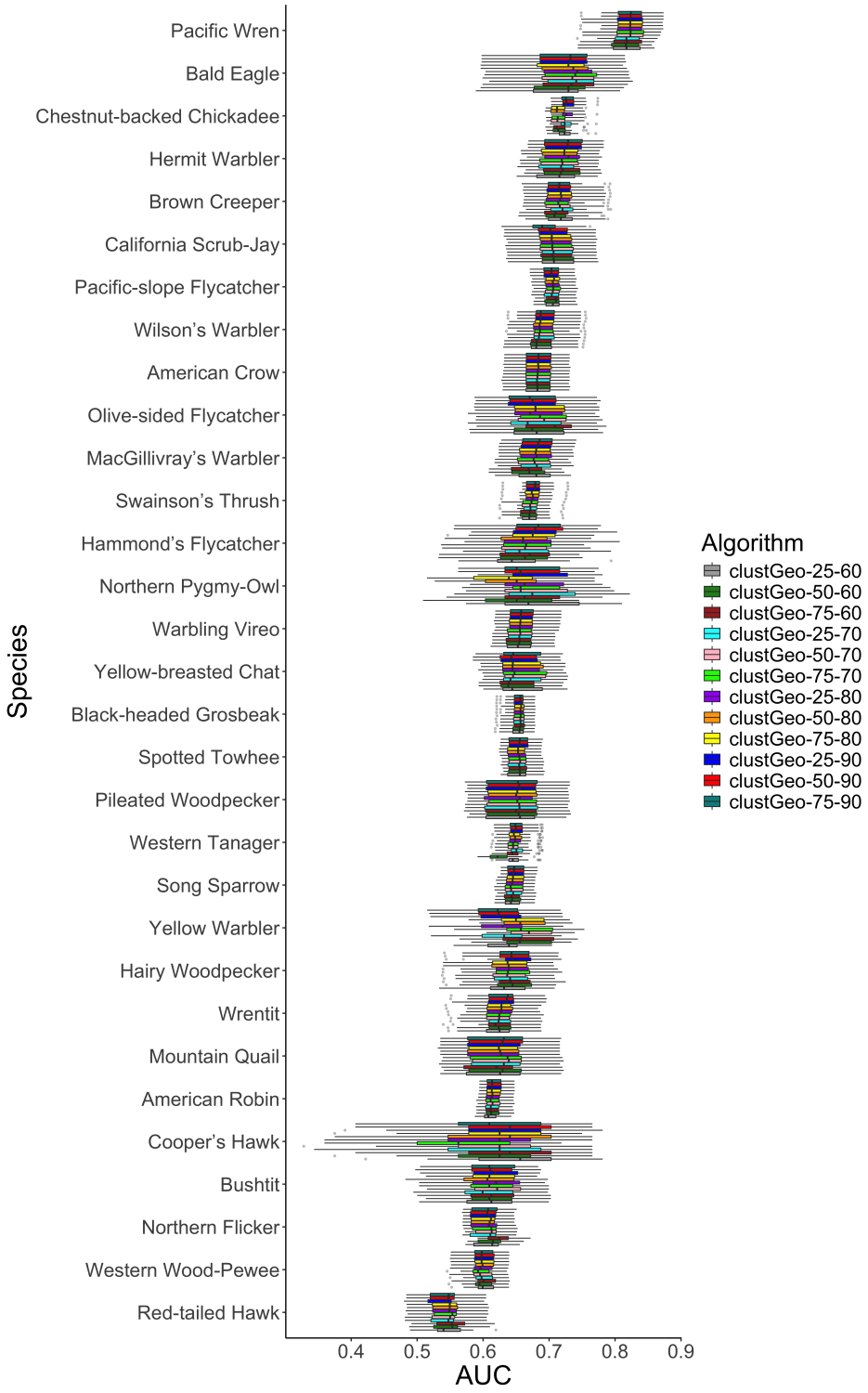

Figure S1: AUC of models built on sites clustered by 12 variants of clustGeo, for all 31 species. Species on the y-axis are in ascending order of mean species AUC (going from the bottom to the top). Choice of α and λ affects the clustering constructed by clustGeo and performance of subsequent occupancy models.

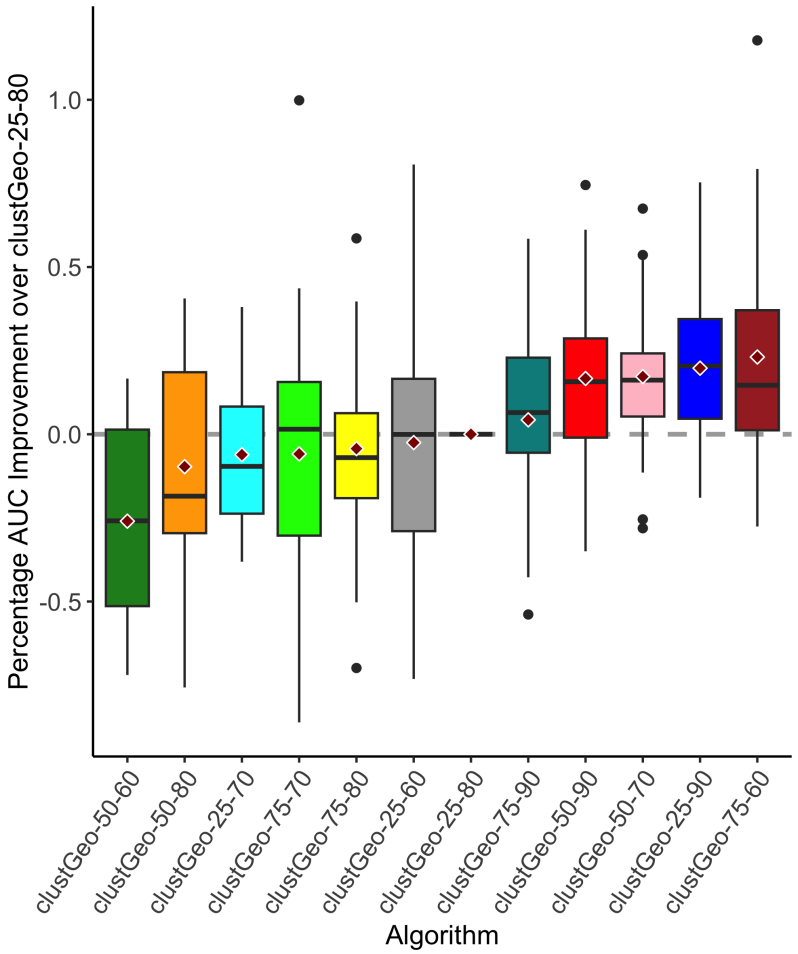

Figure S2: Boxplots show the percentage AUC improvement of clustGeo parameterizations over clustGeo-25-80. Higher positive values indicate better performance compared to clustGeo-25-80. Performance varies across parameter combinations of clustGeo.

Species	**Best α**	**Best λ**	$\Delta\alpha$	$\Delta\lambda$	**AUC (mean)**	**AUC (std. dev.)**
AMCR	0.5	80	-0.0714	18	0.6878	0.0268
AMRO	0.25	90	-0.3214	28	0.6178	0.0155
BAEA	0.25	70	-0.3214	8	0.7322	0.0527
BKHGRO	0.75	60	0.1786	-2	0.6549	0.0153
BRCR	0.25	70	-0.3214	8	0.7179	0.0338
BUTI	0.5	70	-0.0714	8	0.6195	0.0501
CASC	0.25	60	-0.3214	-2	0.7145	0.0356
CHBCHI	0.5	90	-0.0714	28	0.7279	0.0168
COHA	0.25	60	-0.3214	-2	0.6462	0.0845
HAFL	0.25	90	-0.3214	28	0.6778	0.0547
HAWO	0.25	90	-0.3214	28	0.6483	0.0415
HEWA	0.75	90	0.1786	28	0.725	0.0349
MAWA	0.75	90	0.1786	28	0.6836	0.0298
MOQU	0.5	70	-0.0714	8	0.6271	0.054
NOFL	0.75	60	0.1786	-2	0.625	0.0212
NOOW	0.25	90	-0.3214	28	0.6824	0.0548
OLFL	0.75	60	0.1786	-2	0.6964	0.0476
PAFL	0.75	70	0.1786	8	0.7078	0.0165
PAWR	0.75	80	0.1786	18	0.8207	0.0284
PIWO	0.5	60	-0.0714	-2	0.6528	0.0454
REHA	0.75	60	0.1786	-2	0.5493	0.0303
SOSP	0.5	90	-0.0714	28	0.65	0.018
SPTO	0.5	60	-0.0714	-2	0.655	0.0183
SWTH	0.25	90	-0.3214	28	0.6786	0.019
WAVI	0.75	90	0.1786	28	0.6611	0.0263
WEPE	0.75	60	0.1786	-2	0.6016	0.0221
WETA	0.25	70	-0.3214	8	0.6522	0.0185
WIWA	0.25	90	-0.3214	28	0.6938	0.0276
WRENTI	0.25	90	-0.3214	28	0.631	0.0337
YEBCHA	0.75	70	0.1786	8	0.6604	0.036
YEWA	0.5	70	-0.0714	8	0.6691	0.0372

Table S2: Species-specific parameters for clustGeo. α = 0.5 implies uniform weighting of geospatial and environmental habitat features. λ = 80% implies that the number of resultant clusters/sites equals 80% of the number of unique locations of points. The table shows species-specific parameter combinations of α and λ which led to the best model evaluated post hoc on test data. We selected those parameter combinations to capture species-specific characteristics, leading to best-clustGeo. BayesOptClustGeo selected $\alpha = 0.57139$ and $\lambda = 62.1339$. Differences between parameters selected by best-clustGeo and BayesOptClustGeo are shown in $\Delta\alpha$ and $\Delta\lambda$. The last two columns show the performance of the selected best-clustGeo parameterizations.

Method	No. of points	No. of clusters	Min. cluster size	Max. cluster size	Mean cluster size	Std. dev. of cluster size	Percentage single-visit sites
2to10	552	139	2	10	3.9712	2.6208	0
2to10-sameObs	531	134	2	10	3.9627	2.6198	0
1-kmSq	2497	728	1	618	3.4299	23.2246	51.6484
lat-long	2497	1315	1	612	1.8989	17.0597	89.4297
rounded-4	2497	1305	1	612	1.9134	17.1248	88.7356
SVS	2497	2497	1	1	1	0	100
1-UL	1315	1315	1	1	1	0	100
clustGeo-25-60	2497	788	1	632	3.1688	22.7723	51.0152
clustGeo-50-60	2497	788	1	621	3.1688	22.3766	50.8883
clustGeo-75-60	2497	788	1	618	3.1688	22.2684	50.6345
clustGeo-25-70	2497	920	1	632	2.7141	21.0484	60.6522
clustGeo-50-70	2497	920	1	621	2.7141	20.7023	61.087
clustGeo-75-70	2497	920	1	618	2.7141	20.6029	62.5
clustGeo-25-80	2497	1051	1	631	2.3758	19.6592	71.4558
clustGeo-50-80	2497	1051	1	617	2.3758	19.2454	71.6461
clustGeo-75-80	2497	1051	1	617	2.3758	19.2401	71.7412
clustGeo-25-90	2497	1183	1	612	2.1107	17.9919	80.896
clustGeo-50-90	2497	1183	1	612	2.1107	17.9922	81.1496
clustGeo-75-90	2497	1183	1	612	2.1107	17.992	80.896
DBSC	2497	946	1	619	2.6395	20.3699	71.1416
BayesOptClustGeo	2497	816	1	621	3.06	21.9842	52.9412

Table S3: Descriptive statistics of clustered sites. Percentage single-visit sites refer to percentage of clusters that have a single point.

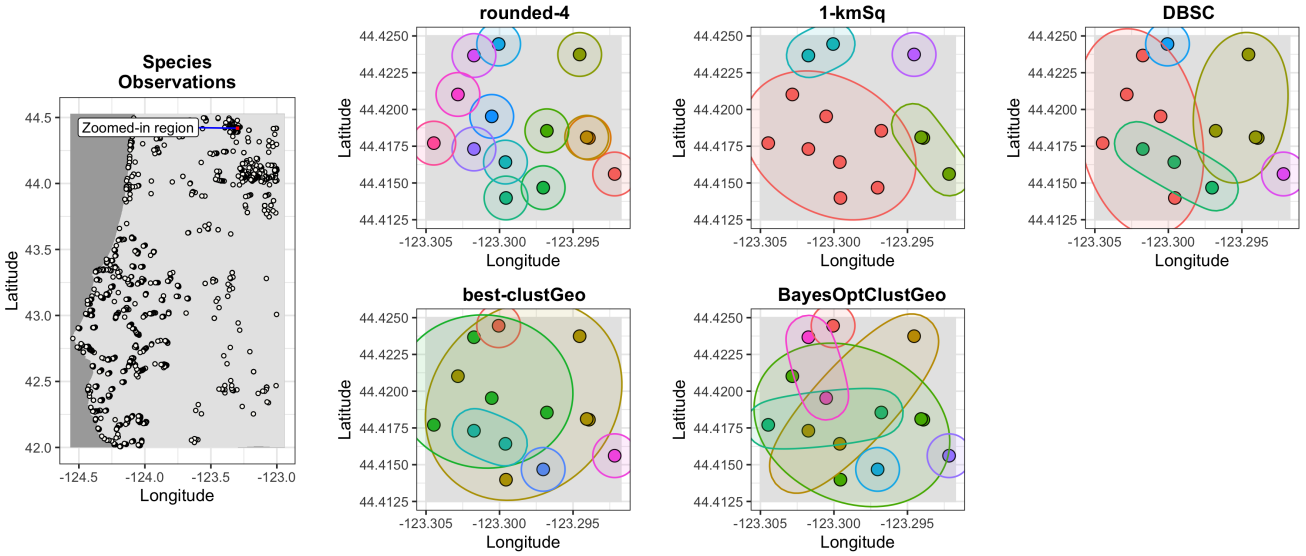

Figure S3: Visualized clusters for Northern Flicker (*Colaptes auratus*). Methods which are able to group points with different geospatial coordinates are shown. Points in the same cluster have identical colors.

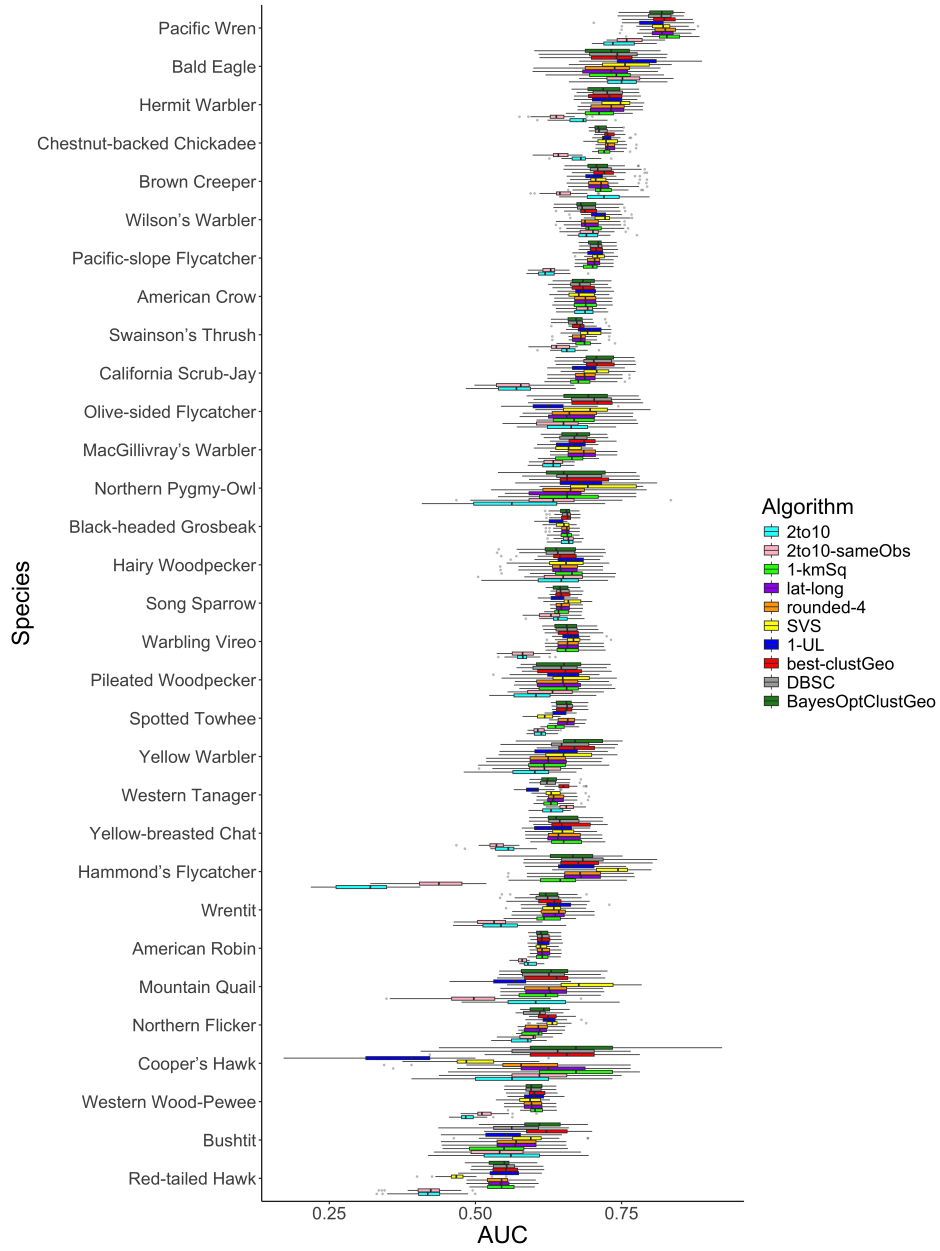

Figure S4: AUC of occupancy models based on sites/clusters produced by ten clustering algorithms over 31 species. Species on the y-axis are in ascending order of mean species AUC (going from the bottom to the top). Though we can see some trends of clustering algorithms having similar performance for each species, further aggregation is necessary to study the intricate effects of clustering algorithm choice on performance, shown as the percentage AUC improvement over lat-long in the main paper.

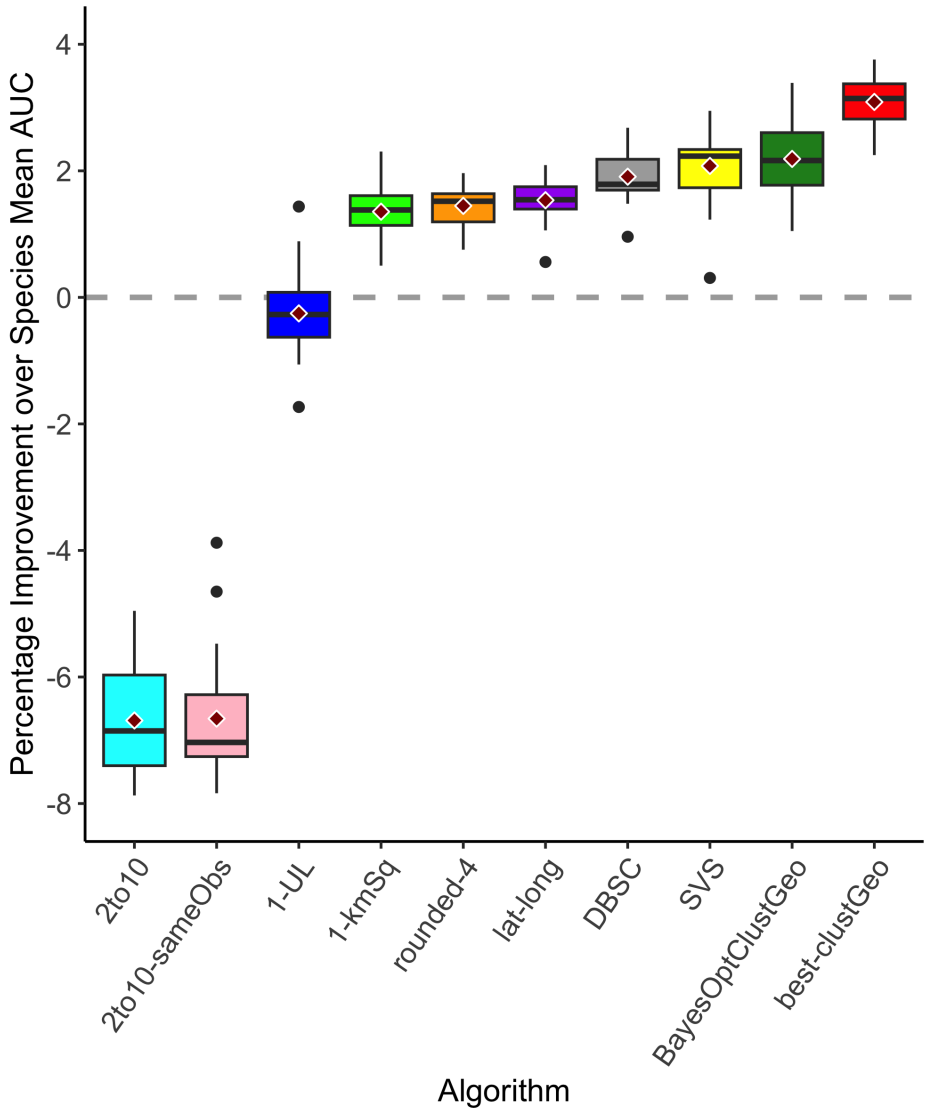

Figure S5: Boxplots show the percentage improvement of each method over the average species AUC. Zero indicates average performance; larger positive values indicate better-than-average performance; negative values indicate worse-than-average performance.

Algorithm 1	Algorithm 2	*p*-value	Adjusted *p*-value
1-UL	SVS	1.41e-08	**6.34e-08**
1-UL	lat-long	0.00272	**0.0051**
1-UL	2to10	0.0449	0.0532
1-UL	2to10-sameObs	0.0427	0.0519
1-UL	rounded-4	0.00605	**0.00972**
1-UL	best-clustGeo	1.5e-15	**2.24e-14**
1-UL	BayesOptClustGeo	1.58e-08	**6.46e-08**
1-UL	DBSC	3.21e-07	**1.03e-06**
lat-long	SVS	0.00748	**0.0116**
lat-long	rounded-4	0.801	0.858
2to10	SVS	1.62e-14	**1.21e-13**
2to10	lat-long	5.64e-07	**1.59e-06**
2to10	2to10-sameObs	0.983	1
2to10	rounded-4	2.03e-06	**4.8e-06**
2to10	best-clustGeo	1.81e-23	**4.08e-22**
2to10	BayesOptClustGeo	1.88e-14	**1.21e-13**
2to10	DBSC	1.11e-12	**5.54e-12**
2to10-sameObs	SVS	1.37e-14	**1.54e-13**
2to10-sameObs	lat-long	5.04e-07	**1.51e-06**
2to10-sameObs	rounded-4	1.82e-06	**4.55e-06**
2to10-sameObs	best-clustGeo	1.46e-23	**6.56e-22**
2to10-sameObs	BayesOptClustGeo	1.59e-14	**1.43e-13**
2to10-sameObs	DBSC	9.48e-13	**5.33e-12**
rounded-4	SVS	0.00342	**0.00616**
1-kmSq	SVS	0.00235	**0.0048**
1-kmSq	1-UL	0.00854	**0.0124**
1-kmSq	lat-long	0.713	0.783
1-kmSq	2to10	3.56e-06	**7.64e-06**
1-kmSq	2to10-sameObs	3.21e-06	**7.22e-06**
1-kmSq	rounded-4	0.908	0.95
1-kmSq	best-clustGeo	8.92e-08	**3.34e-07**
1-kmSq	BayesOptClustGeo	0.0025	**0.0049**
1-kmSq	DBSC	0.0131	**0.0184**
best-clustGeo	SVS	0.0212	**0.0272**
best-clustGeo	lat-long	6.37e-07	**1.69e-06**
best-clustGeo	rounded-4	1.68e-07	**5.8e-07**
best-clustGeo	DBSC	0.00415	**0.00692**
BayesOptClustGeo	SVS	0.984	0.984
BayesOptClustGeo	lat-long	0.00793	**0.0119**
BayesOptClustGeo	rounded-4	0.00364	**0.0063**
BayesOptClustGeo	best-clustGeo	0.0201	**0.0266**
BayesOptClustGeo	DBSC	0.588	0.661
DBSC	SVS	0.574	0.663
DBSC	lat-long	0.0346	**0.0432**
DBSC	rounded-4	0.018	**0.0245**

Table S4: Pairwise statistical significance testing based on percentage AUC improvement over lat-long. We ran Dunn's test with *p*-value adjustment for multiple testing by Benjamini-Hochberg method. Significant differences denoted by adjusted *p*-values < 0.05 are in bold.

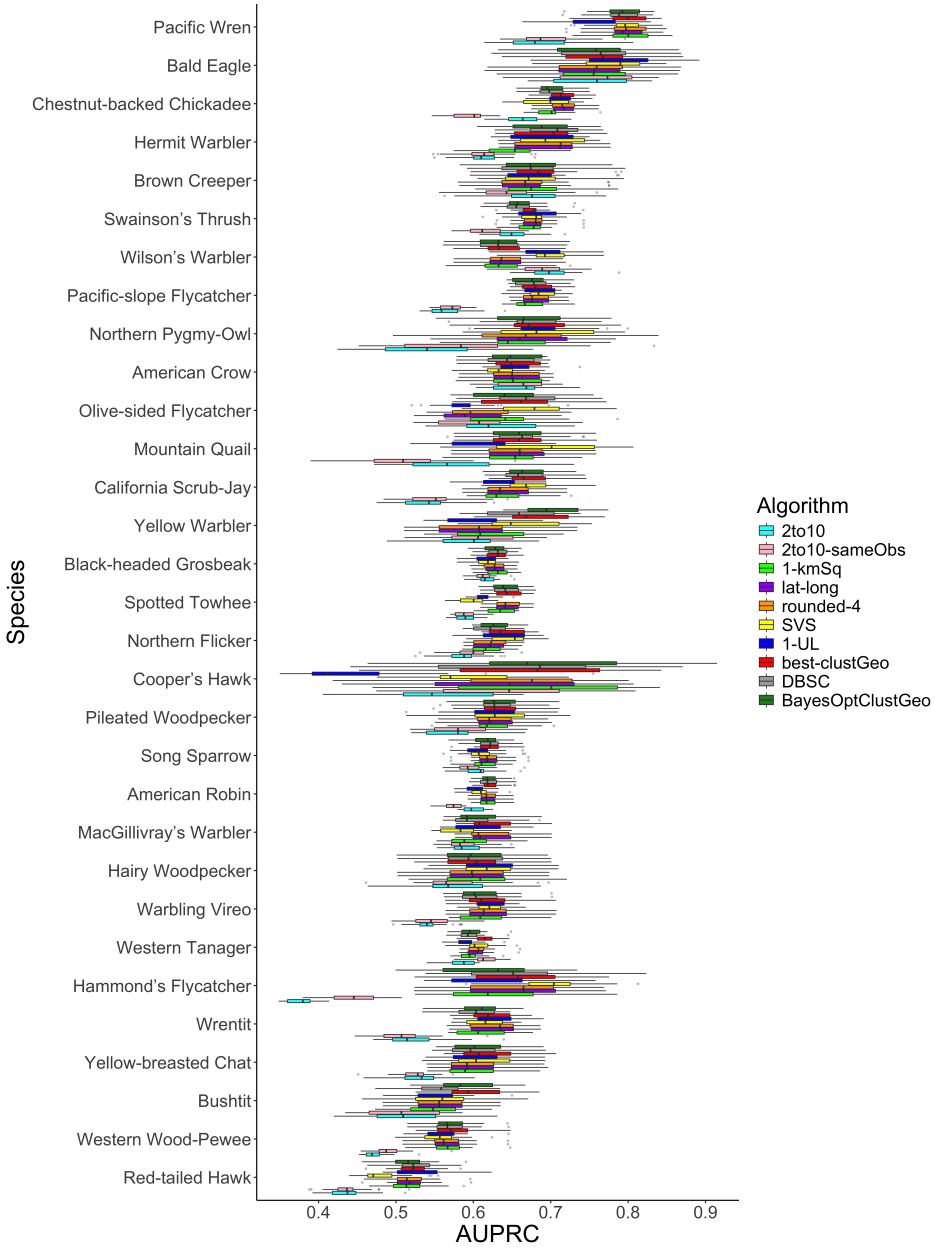

Figure S6: AUPRC of occupancy models based on sites/clusters produced by ten clustering algorithms over 31 species. Species on the y-axis are in ascending order of mean species AUPRC (going from the bottom to the top).

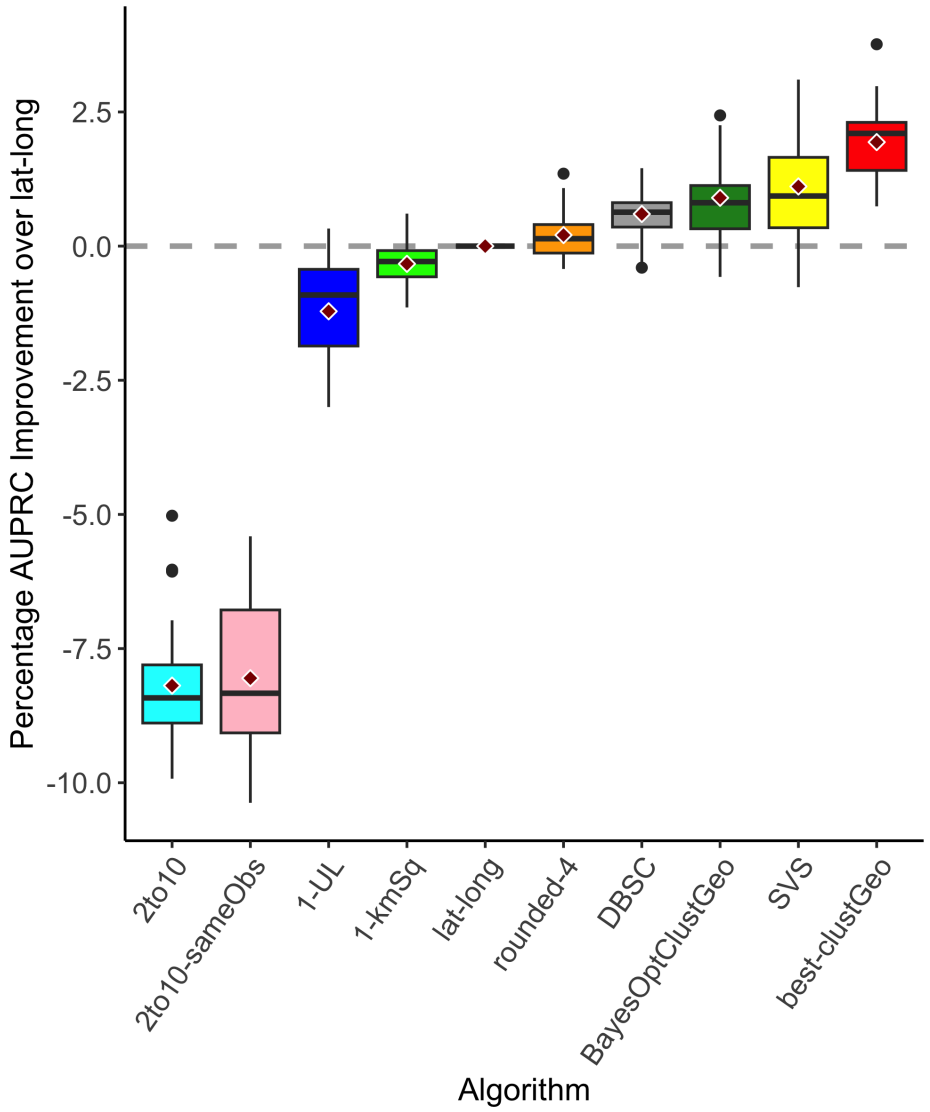

Figure S7: Boxplots show the percentage AUPRC improvement of each method over lat-long. Larger positive values indicate better performance than lat-long; negative values indicate worse performance than lat-long.

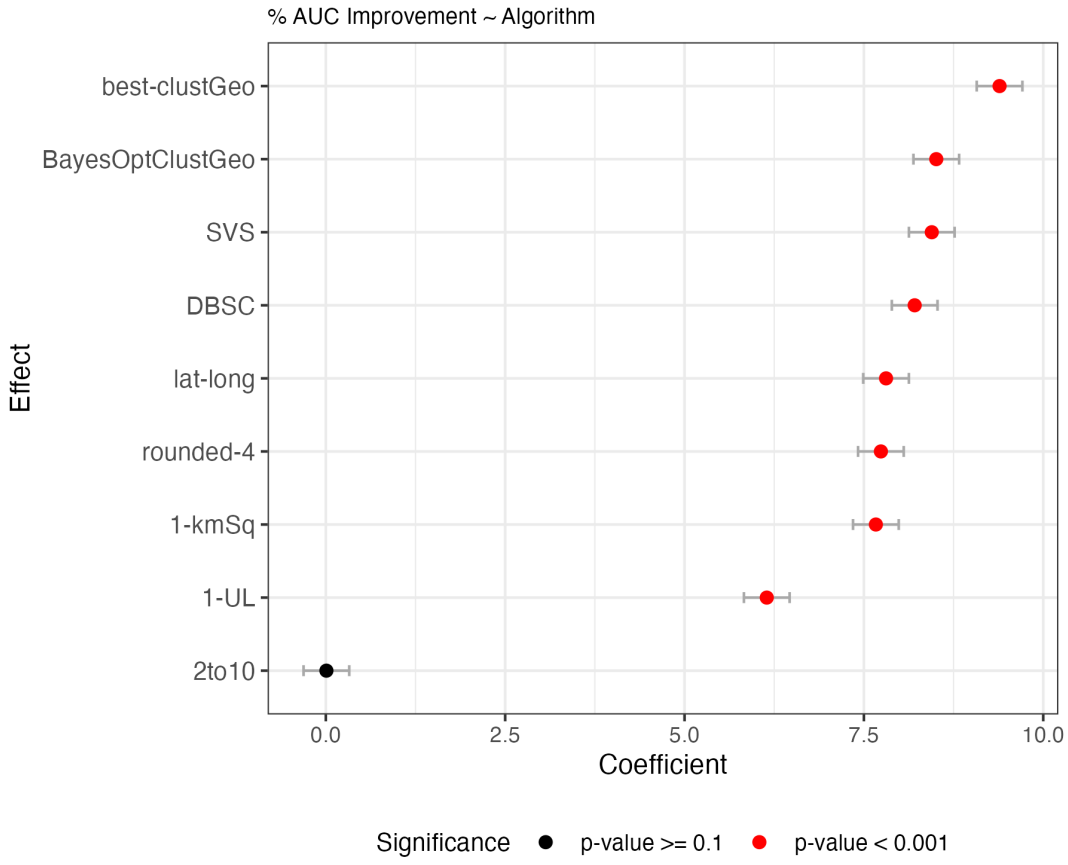

Figure S8: Non-intercept coefficients of linear mixed-effect model for measuring the effects of clustering algorithm on percentage AUC improvement over lat-long. 2to10-sameObs is the reference level.

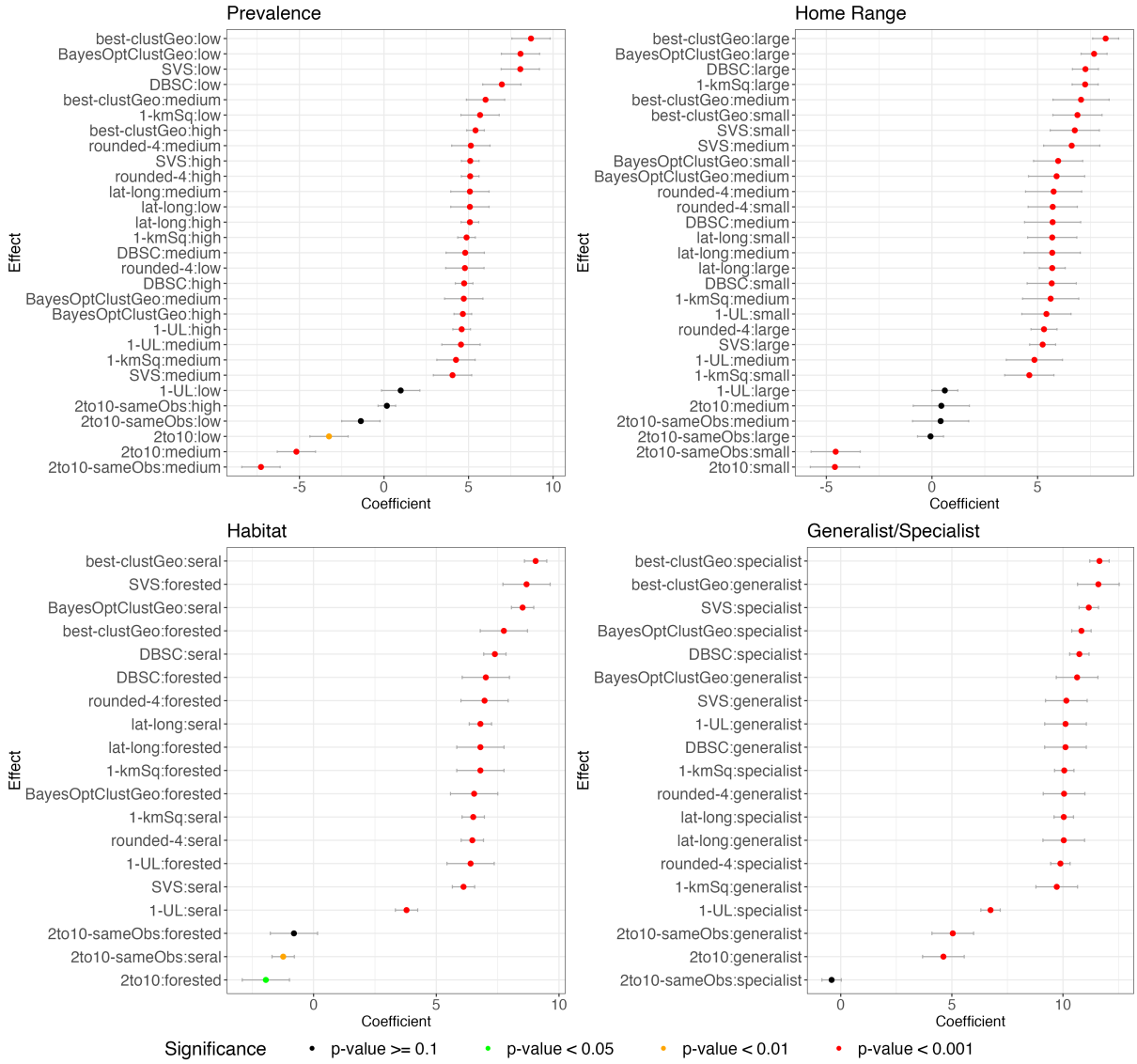

Figure S9: Non-intercept coefficients of linear mixed-effect models for measuring the effects of species traits on impact of clustering algorithms on percentage AUC improvement over lat-long.

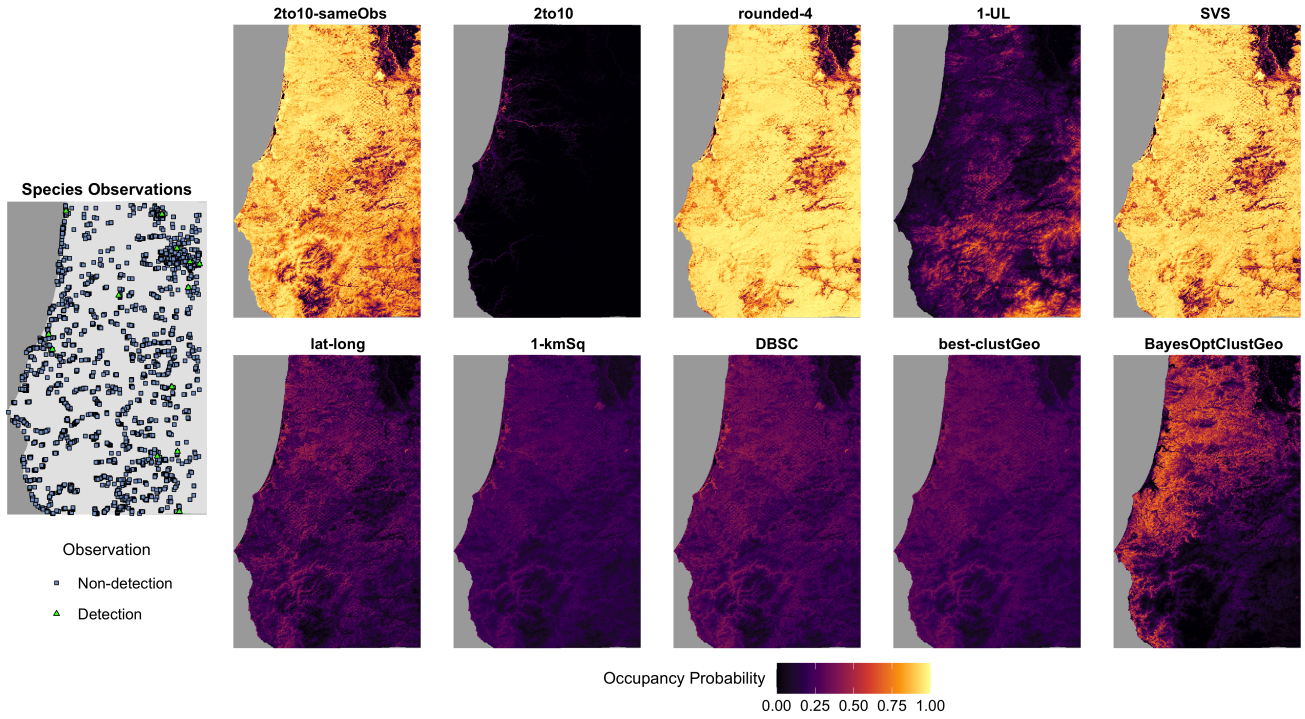

Figure S10: Another mapped species example: Occupancy probability of Cooper's Hawk (*Accipiter cooperii*) over southwestern Oregon, United States predicted by species distribution models built from sites produced by ten clustering algorithms.