Supplementary Information

Machine learning based national mapping of archaeological occurrence and landscape pressure in Tajikistan

Li, Nebbia, Cilio, Cuthbertson, Pang, Jorayev, Williams, Xu & Sun

Contents

Supplementary Tables **2**

Supplementary Figures **5**

Supplementary Tables

Supplementary Table S1: Environmental predictors used in the occurrence models. The table reports metadata supported by the available input files and preparation records.

Variable (group; type)	Unit	Description of supplied input
Elevation (`dem`) Terrain; continuous	m	Prepared elevation raster used as the DEM predictor.
Slope (`slope`) Terrain; continuous	°	Prepared slope raster; the national and regional inputs differ on shared cells, as reported in Methods.
Aspect (`aspect`) Terrain; continuous	°	Prepared DEM-derived aspect raster.
Topographic wetness index (`twi`) Hydrology; continuous	Unitless	Prepared DEM-derived wetness-index raster.[1]
Distance to high-order rivers (`hirivdist`) Hydrology; continuous	m	Prepared distance-to-rivers raster.
NDVI (`ndvi`) Surface context; continuous	Unitless	Sentinel-2-derived NDVI predictor for 2023.
Land cover (`worldcover`) Surface context; categorical	Class	ESA WorldCover 2021 (v200); original class codes (10–100) retained and treated as categorical in MaxEnt.
Night-time light (`nitlight`) Modern disturbance; continuous	$\mathrm{nW\,cm^{-2}\,sr^{-1}}$	VIIRS DNB-derived night-time-light predictor for 2023.

The environmental predictors were supplied as preprocessed raster layers, prepared using Google Earth Engine and other GIS workflows, and distributed in release v1.0.0 of the project repository (`https://github.com/sq-modelling/tajikistan-archaeological-risk-mapping`). All predictors were provided on a common 99.99 m (≈100 m) modelling grid in WGS 84 / UTM zone 41N (EPSG:32641). This grid resolution can differ from the native resolution of the source datasets. The NDVI predictor for 2023 is distinct from the agricultural-pressure indicator, which uses observations from 2020, 2021 and 2024, as described in Methods.

Supplementary Table S2: Test AUC of each fold of the random ten-fold cross-validation of the published-configuration models (MaxEnt defaults). Each cell gives the AUC reported by MaxEnt, whose background includes the training presences, followed by the AUC of the same model against 60,000 random cells of its domain. The last column is the national model refitted with 100,000 background points without adding presences, for which the two values nearly coincide.

Fold	Tajikistan	DRS	GBAO	Khatlon	Sughd	Tajikistan (100k)
1	0.687 / 0.864	0.771 / 0.902	0.858 / 0.907	0.726 / 0.765	0.841 / 0.920	0.856 / 0.856
2	0.668 / 0.854	0.790 / 0.915	0.865 / 0.913	0.689 / 0.728	0.857 / 0.929	0.874 / 0.875
3	0.677 / 0.861	0.792 / 0.916	0.870 / 0.919	0.714 / 0.752	0.855 / 0.929	0.867 / 0.867
4	0.685 / 0.868	0.763 / 0.895	0.854 / 0.904	0.740 / 0.777	0.862 / 0.933	0.851 / 0.851
5	0.694 / 0.870	0.790 / 0.916	0.863 / 0.910	0.709 / 0.747	0.840 / 0.916	0.863 / 0.864
6	0.682 / 0.865	0.770 / 0.900	0.873 / 0.917	0.722 / 0.760	0.860 / 0.934	0.866 / 0.866
7	0.693 / 0.867	0.775 / 0.906	0.840 / 0.889	0.721 / 0.760	0.845 / 0.920	0.871 / 0.872
8	0.683 / 0.866	0.775 / 0.908	0.871 / 0.917	0.737 / 0.775	0.831 / 0.908	0.861 / 0.862
9	0.683 / 0.865	0.776 / 0.906	0.857 / 0.905	0.703 / 0.743	0.861 / 0.931	0.859 / 0.860
10	0.676 / 0.859	0.780 / 0.908	0.860 / 0.909	0.761 / 0.798	0.857 / 0.930	0.856 / 0.857
Mean	0.683 / 0.864	0.778 / 0.907	0.861 / 0.909	0.722 / 0.760	0.851 / 0.925	0.862 / 0.863

Supplementary Table S3: Test AUC of each fold of the spatially blocked five-fold cross-validation (50 km blocks). Regional columns give the regional model, then the national model of the same fold scored on the same held-out regional presence and background locations, using their respective predictor values; numbers of held-out presence cells are in brackets. The last column is the national model with terrain and hydrological predictors only.

Fold	Tajikistan	DRS	GBAO	Khatlon	Sughd	Terrain only
1	0.892 (2,962)	0.909 / 0.859 (1,830)	0.888 / 0.835 (288)	0.849 / 0.682 (356)	0.801 / 0.805 (488)	0.884
2	0.804 (1,730)	0.807 / 0.824 (733)	0.908 / 0.893 (300)	0.574 / 0.626 (375)	0.965 / 0.875 (322)	0.781
3	0.835 (1,831)	0.786 / 0.756 (751)	0.892 / 0.854 (288)	0.688 / 0.693 (443)	0.894 / 0.887 (348)	0.822
4	0.830 (2,190)	0.904 / 0.892 (750)	0.880 / 0.873 (296)	0.652 / 0.583 (384)	0.848 / 0.807 (760)	0.826
5	0.870 (2,205)	0.887 / 0.822 (1,182)	0.929 / 0.865 (288)	0.564 / 0.509 (384)	0.895 / 0.893 (351)	0.837
Mean	0.846	0.859 / 0.830	0.899 / 0.864	0.665 / 0.619	0.881 / 0.853	0.830

Supplementary Table S4: Share (%) of cells with valid predictions from both national and regional models, and of valid presence cells, in each occurrence level within each region, for the national surface and for the surface of the regional model, using equal-interval classes of the ten-replicate average on the 100 m grid. Presence cells were used in model training, so these shares describe the mapped surfaces; independent performance is reported in Supplementary Tables S2 and S3.

		Area in level (%)					Presence cells in level (%)				
Region	Surface	1	2	3	4	5	1	2	3	4	5
DRS	national	11.2	15.7	23.0	36.9	13.2	0.0	0.3	3.6	31.7	64.4
	regional	51.6	21.2	16.3	8.3	2.7	2.1	6.6	18.6	33.4	39.3
GBAO	national	60.2	25.5	9.6	3.7	1.0	4.7	26.8	32.2	28.9	7.5
	regional	66.1	18.4	8.6	4.3	2.6	4.5	14.6	20.3	23.1	37.6
Khatlon	national	1.1	14.4	38.1	39.1	7.4	0.3	6.1	25.3	49.8	18.4
	regional	16.3	29.6	32.8	15.9	5.4	2.2	11.5	27.1	27.2	32.0
Sughd	national	13.9	29.1	32.3	21.3	3.3	0.1	1.8	14.0	49.7	34.5
	regional	69.9	14.1	8.9	5.1	2.0	5.3	7.1	15.9	33.5	38.2

Supplementary Table S5: Sensitivity of the classified occurrence surface and of the high–high class ($S \geq 4$ and $P \geq 4$) to model configuration and classification. All surfaces are sampled at the centres of the 259,062 integration-grid cells where P is defined and all eight predictors are valid (94,431 km^2); the original composite pressure raster P is held fixed. Equal-interval classes use the range of each surface on the 100 m grid; quintiles are computed over the analysed cells. 'Retained': share of the baseline high–high cells that are also high–high in the alternative. 'In baseline': share of the alternative high–high cells that are also high–high in the baseline. 'Jaccard': intersection over union. The predictor-set comparison uses two full-data models fitted to the same 10,918 presence cells and 10,000 background cells; the last row of each block checks that the eight-predictor full-data fit matches the ten-fold average.

Baseline	Alternative	Spearman ρ	Same level (%)	$S \geq 4$ (km^2) baseline / alt.	High–high (km^2) baseline / alt.	Retained (%)	In baseline (%)	Jaccard (%)
Equal-interval classes								
Default, ten-fold mean	100,000 background, no added presences	0.96	28.8	32,472 / 12,782	180 / 75	42	99.5	42
Eight predictors, full-data fit	Terrain and hydrology only, full-data fit	0.89	67.5	35,539 / 42,466	188 / 248	87	66	60
Default, ten-fold mean	Eight predictors, full-data fit	0.998	88.8	32,472 / 35,539	180 / 188	98	94	92
Quintile classes								
Default, ten-fold mean	100,000 background, no added presences	0.96	71.8	37,773 / 37,773	226 / 244	85	78	68
Eight predictors, full-data fit	Terrain and hydrology only, full-data fit	0.89	60.8	37,773 / 37,773	203 / 223	76	69	57
Default, ten-fold mean	Eight predictors, full-data fit	0.998	93.3	37,773 / 37,773	226 / 203	89	99	88

Supplementary Table S6: Regional versus national models with the same predictor inputs. The regional fold models were refitted with the national predictor rasters, using the same training presence cells, background sample and 50 km folds as presented in Supplementary Table S3, and scored, together with the national fold model, on the same held-out regional presence and background cells. Each cell gives regional / national test AUC, with the number of held-out presence cells in brackets.

Region	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Regional higher
DRS	0.905 / 0.859 (1,830)	0.800 / 0.824 (733)	0.779 / 0.756 (751)	0.899 / 0.892 (750)	0.883 / 0.822 (1,182)	0.853 / 0.830	4/5
GBAO	0.887 / 0.835 (288)	0.910 / 0.893 (300)	0.886 / 0.854 (288)	0.877 / 0.873 (296)	0.922 / 0.865 (288)	0.897 / 0.864	5/5
Khatlon	0.858 / 0.682 (356)	0.566 / 0.626 (375)	0.686 / 0.693 (443)	0.650 / 0.583 (384)	0.559 / 0.509 (384)	0.664 / 0.619	3/5
Sughd	0.798 / 0.805 (488)	0.960 / 0.875 (322)	0.894 / 0.887 (348)	0.867 / 0.807 (760)	0.896 / 0.893 (351)	0.883 / 0.853	4/5

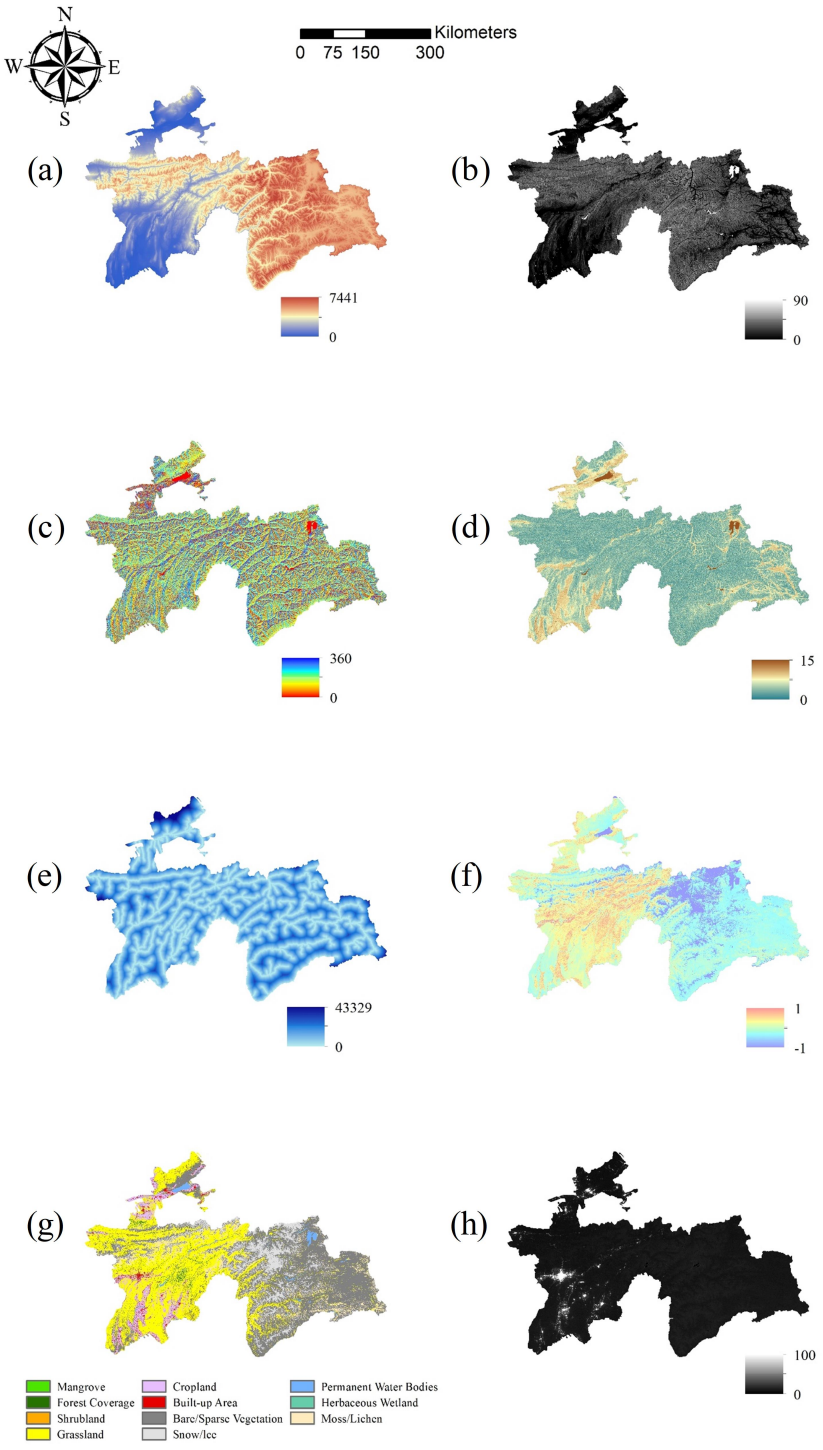
Supplementary Table S7: Sensitivity of the spatially blocked evaluation of the national model across block sizes and buffered evaluation designs. Blocks containing presences were assigned to five folds with the same greedy rule balancing regional presence counts; buffers remove training presences and background within the stated distance of any held-out block. Buffered runs use the same folds, held-out presences and held-out background as the main 50 km design, so they differ from it only in the removed training data. Test AUC per fold, mean ± s.d., and ranges of held-out and training presence cells per fold. With 100 and 200 km blocks the folds become unbalanced and each removes large contiguous areas.

Design	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Held-out presences	Training presences
25 km	0.889	0.830	0.837	0.867	0.863	0.857 ± 0.024	2,089–2,421	8,497–8,829
50 km (main design)	0.892	0.804	0.835	0.830	0.870	0.846 ± 0.035	1,730–2,962	7,956–9,188
50 km, 10 km buffer	0.887	0.799	0.817	0.828	0.861	0.839 ± 0.035	1,730–2,962	6,062–7,585
50 km, 25 km buffer	0.874	0.804	0.818	0.835	0.832	0.833 ± 0.026	1,730–2,962	2,570–5,108
100 km	0.896	0.754	0.761	0.730	0.743	0.777 ± 0.068	455–5,098	5,820–10,463
200 km	0.899	0.720	0.880	0.841	0.721	0.812 ± 0.086	34–4,493	6,425–10,884

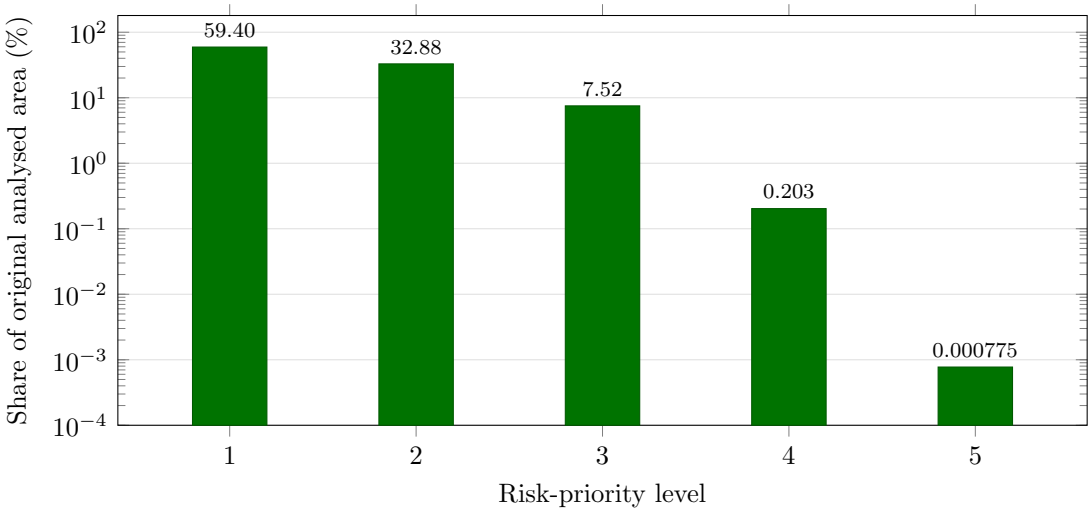
Supplementary Figures

References

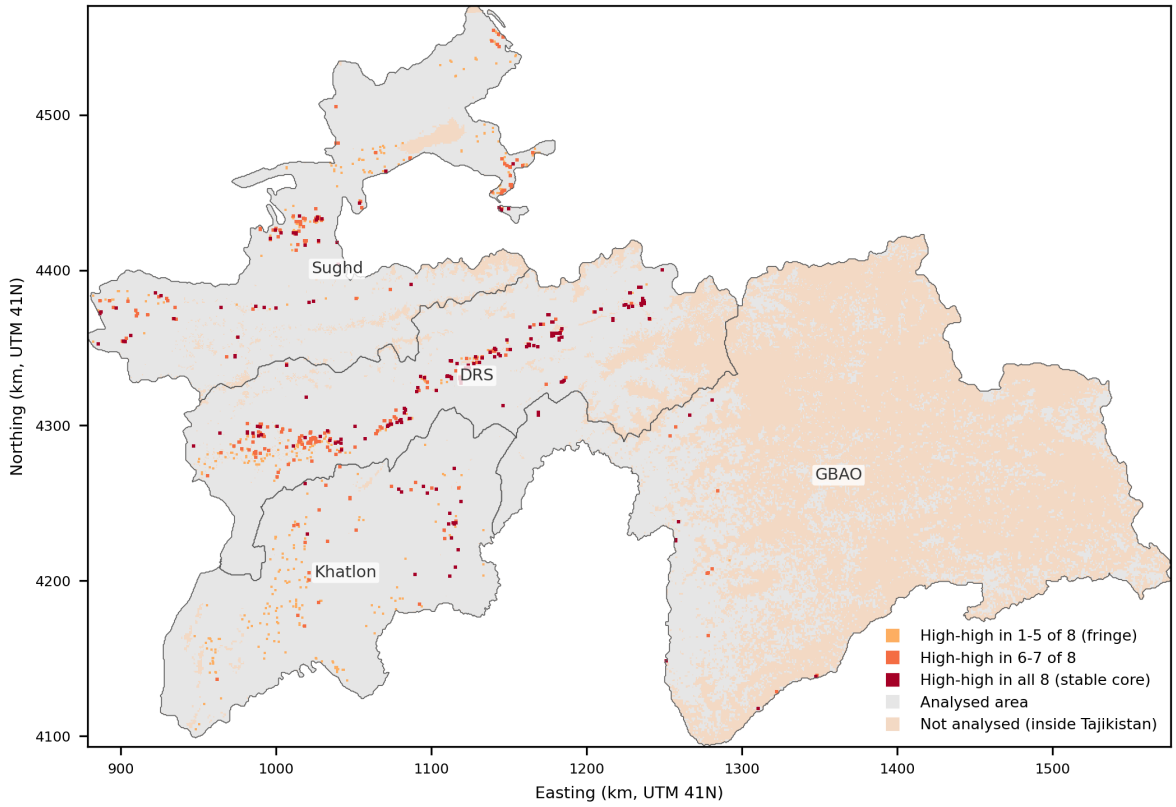
[1] Böhner, J. & Selige, T. Spatial prediction of soil attributes using terrain analysis and climate regionalisation. In *SAGA-Analyses and modelling applications*, 13–28 (Goltze, 2006).



Supplementary Fig. S1: Environmental predictors of the national occurrence model: (a) elevation (m); (b) slope (°); (c) aspect (°); (d) topographic wetness index; (e) distance to high-order rivers (m); (f) NDVI; (g) land cover (ESA WorldCover); (h) night-time light radiance.



Supplementary Fig. S2: Share of the original 94,243.34 km^2 analysed area in each risk-priority level (R), calculated from the original S–P cross-tabulation presented in Table 4 of the main text. The vertical axis is logarithmic, and the shares refer to the original analysed area.



Supplementary Fig. S3: Stability of the high–high class ($S \geq 4$ and $P \geq 4$) across eight combinations of four occurrence surfaces (ten-fold default, ten-fold with 100,000 background points, full-data eight-predictor and full-data five-predictor models) and two classification schemes (equal intervals and quintiles), with the original composite pressure raster held fixed, as presented in Supplementary Table S5. Dark red: high–high in all eight (stable core, 75 km^2; DRS 49, Sughd 15, Khatlon 7 and GBAO 4 km^2). Orange: six or seven (86 km^2). Light orange: one to five (configuration-dependent fringe, 181 km^2). Grey: analysed area; beige: inside Tajikistan, outside the analysed area. Cells of the 603.75 m integration grid are enlarged for visibility. Lines: regional modelling units. Projection: WGS 84 / UTM zone 41N.