

METHODOLOGY FOR SALINITY MAPPING

1. Soil collection methodology

Soil sampling was conducted using a systematic field protocol designed to obtain representative measurements while accounting for the natural spatial variability of soils across Karakalpakstan.

At each sampling location, a **composite soil sampling approach** was applied. The sampling site was represented by a **30 m × 30 m plot**, within which five subsamples were collected. One subsample was taken from the center of the plot, and four additional subsamples were collected from its corners, approximately 21 m from the center. Five subsamples were collected to characterize each sampling location. Measurements obtained from these subsamples were subsequently averaged to derive representative value for each sampling location. This approach reduces the influence of small-scale soil heterogeneity and provides a more representative characterization of local soil conditions than a single-point observation.

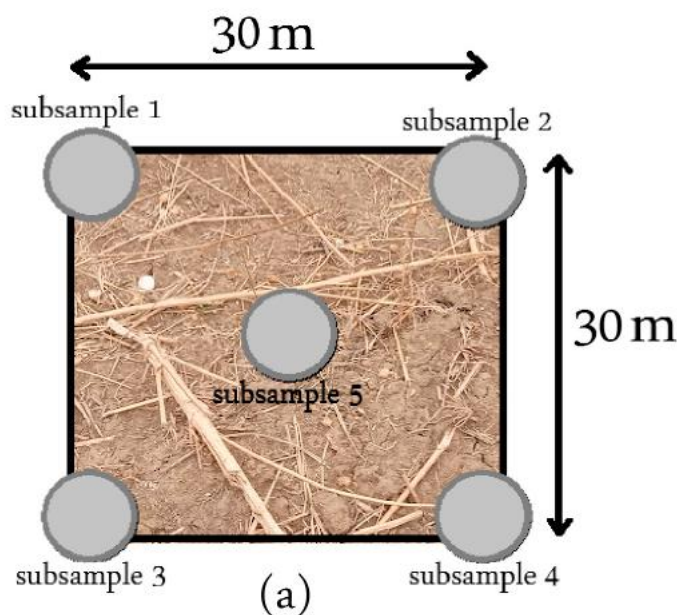


Figure 1. Sampling design

The sampling procedure was designed to support the spatial assessment of key soil properties across the diverse agricultural and degraded landscapes of Karakalpakstan. Field observations generated through this standardized sampling framework subsequently formed the primary dataset for spatial analysis and digital soil mapping.

2. From Field Sampling to Spatial Soil Information

The field sampling campaign provides the foundation of the **Karakalpakstan Soil Geoportal**. Each sampling location represents a geographically referenced observation of soil conditions, allowing field measurements to be linked directly with GIS-based spatial analysis.

The sampling strategy recognizes that soils are inherently heterogeneous and that sampling design should correspond to the intended mapping approach.

Through the combination of standardized field sampling, georeferenced observations and subsequent spatial modelling, individual soil measurements can be transformed into continuous information on the spatial distribution of soil properties across Karakalpakstan.

Field Sampling → Georeferenced Soil Measurements → Soil Database → Spatial Analysis → Soil Property Maps → Karakalpakstan Soil Geoportal

This provides a transparent methodological connection between what was measured in the field and what users ultimately see on the soil maps.

3. GIS-Based Creation of Soil Sampling Points in ArcGIS Pro

A stratified and spatially balanced sampling approach was implemented in ArcGIS Pro to generate representative soil sampling locations across the study area. The classified Land Use/Land Cover (LULC) raster was used as the main spatial framework for the sampling design.

First, the classified LULC raster was processed so that spatially adjacent raster cells belonging to the same LULC class were grouped into contiguous regions. Each contiguous region was treated as an individual sampling stratum. This allowed the sampling design to account for both land cover type and its spatial distribution across the study area.

The area of each stratum was calculated using the following equation:

$$Area(k) = N_k \times A_{cell}$$

where:

Area(k) = area of stratum k

N_k = number of raster cells within stratum k

A_{cell} = area of one raster cell

The total study area was calculated as the sum of the areas of all strata:

$$A_{total} = \sum Area(k)$$

The number of sampling points assigned to each stratum was then determined proportionally to its area:

$$n_k = [Area(k) / A_{total}] \times N_{total}$$

where:

n_k = number of sampling points allocated to stratum k

$Area(k)$ = area of stratum k

A_{total} = total study area

N_{total} = total number of sampling points

This proportional allocation ensured that larger contiguous regions received more sampling points, while smaller regions were also represented in the sampling design.

Within each stratum, sampling points were generated using a spatially balanced random sampling approach based on the Generalized Random Tessellation Stratified (GRTS) algorithm. Candidate sampling points were generated within the corresponding strata until the required number of points was reached.

To reduce clustering of sampling locations and improve their spatial distribution, a minimum-distance criterion was applied. The distance between any two sampling points had to satisfy the following conditions:

$$Distance(p_i, p_j) \geq D_{min}$$

where:

p_i = newly generated candidate sampling point

p_j = previously accepted sampling point

D_{min} = minimum required separation distance

Candidate points that did not satisfy the minimum-distance requirement were rejected and regenerated. A minimum separation distance of 2 km was applied in the final sampling design.

The resulting sampling network therefore combined LULC-based stratification, proportional allocation of sampling points, spatially balanced random sampling, and minimum-distance control. This approach provided a systematic and spatially representative distribution of soil sampling locations for subsequent field data collection and soil mapping in Karakalpakstan.

Final field sampling locations were selected from the candidate locations generated while accounting for field accessibility and logistical constraints. Where predefined locations were inaccessible, sampling points were relocated to suitable nearby

locations while maintaining, as far as practicable, the original sampling design.

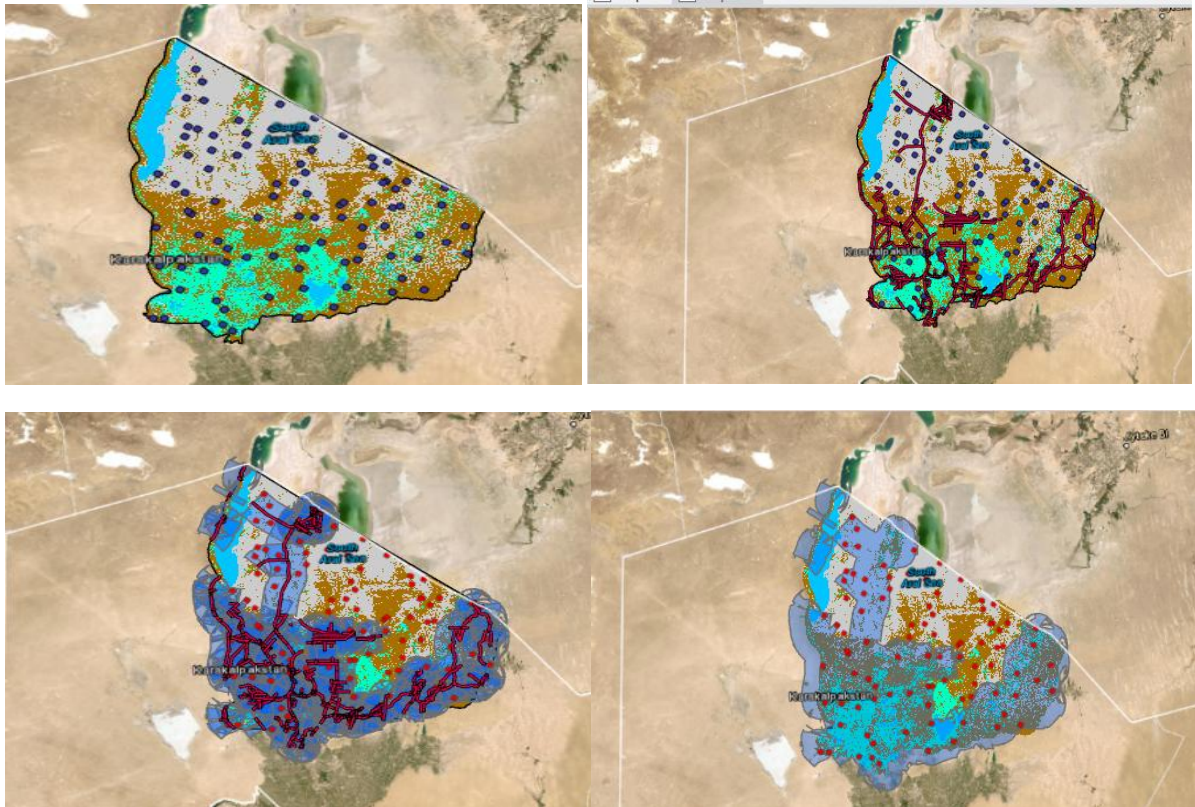


Figure 2. Spatial allocation of soil sampling locations based on LULC stratification and field accessibility

4. Soil Measurements

Soil measurements were conducted at two depths: 0–15 cm and 15–30 cm. Seven key soil parameters were assessed: pH, electrical conductivity (EC), soil moisture, soil temperature, nitrogen (N), phosphorus (P), and potassium (K). Field measurements were obtained using a portable soil measurement device.

At each sampling location, five subsamples were collected following a standardized spatial sampling pattern: one subsample from the center and four subsamples from the surrounding edges of the sampling quadrat. Measurements from the five subsamples were combined by calculating their arithmetic mean, which was used as the representative value for each sampling location and soil depth.

The resulting averaged soil data were subsequently processed in a GIS environment and spatially interpolated to generate continuous soil-property maps for pH, EC, soil moisture, soil temperature, N, P, and K across the study area.



Figure 3. Soil device



Figure 4. Field work

5. Interpolation and cross validation results

Spatial interpolation was used to transform discrete soil observations into continuous soil-property surfaces. Empirical Bayesian Kriging (EBK) was tested as the principal geostatistical interpolation method, while Inverse Distance Weighting (IDW) was evaluated as an alternative where EBK produced unsuitable predictions or where comparison between methods was required.

A 500 m output cell size was used for the final interpolation workflow. For the documented EBK runs, the input Z field was the average measured value. The EC workflow, for example, used no data transformation, a Power semivariogram model, 100 points in each local model, an overlap factor of 1, 100 simulated semivariograms, and a standard circular search neighborhood with a maximum of 15 and minimum of 10 neighbors.

Table 1. Empirical Bayesian Kriging parameters in ArcGIS Pro

Parameter	Setting / Value
Interpolation method	Empirical Bayesian Kriging (EBK)
Maximum number of points in each local model	100
Local model area overlap factor	1
Number of simulated semivariograms	100
Search neighborhood	Standard Circular
Maximum neighbors	15
Minimum neighbors	10
Radius	1.55270445324274
Output surface type	Prediction

IDW was also tested using the Geostatistical Analyst tools. The documented configuration used a 500 m cell size, the average soil-property value as the Z value, a maximum of 15 neighbors, a minimum of 10 neighbors, and a one-sector search neighborhood.

Table 2. IDW parameters in ArcGIS Pro

Parameter	Setting / Value
Interpolation method	Inverse Distance Weighted (IDW)
Output cell size	500 m
Power	2
Search neighborhood	Standard
Maximum neighbors	15
Minimum neighbors	10
Sector type	1 Sector
Angle	0
Major semiaxis	1.55270445324274
Minor semiaxis	1.55270445324274
Weight field	None

The interpolation method was selected separately for each soil property based on model behavior, cross-validation statistics, and the plausibility of the resulting prediction surface. This was particularly important for highly variable parameters such as potassium.

Cross-validation was used to evaluate the predictive performance and uncertainty of the interpolation models before the final soil surfaces were selected. The assessment considered Mean Error, Root-Mean-Square Error (RMSE), Mean Standardized Error, Root-Mean-Square Standardized Error, Average Standard Error, Average Continuous Ranked Probability Score (CRPS), and the percentages of observations falling within the 90% and 95% prediction intervals.

In general, a Mean Error and Mean Standardized Error close to zero indicates low prediction bias. A Root-Mean-Square Standardized Error close to 1 and an Average Standard Error close to the RMSE indicates that prediction uncertainty is being estimated appropriately. For probabilistic EBK outputs, 90% and 95% interval coverage values close to their nominal levels provide additional evidence of model calibration.

Table 3. Soil Depth: 0–15 cm

Parameter	Method	n	RMSE	Mean error	RMS std.	Avg. SE	90%	95%	CRPS
pH	EBK	280	0.289	0.00039	0.892	0.320	94.27%	96.42%	0.150
EC	EBK	280	3.168	0.0285	0.981	3.259	90.65%	92.45%	1.679
Soil moisture	EBK	280	7.480	-0.0028	1.001	7.374	92.45%	97.48%	3.908
Soil temperature	EBK	280	5.981	-0.091	0.853	7.145	92.09%	96.40%	3.307
K	IDW	280	139.9	77.980	—	—	—	—	—
N	IDW	280	115	74.2	—	—	—	—	—
P	IDW	280	127	75.3	—	—	—	—	—

The 0–15 cm EBK results showed low bias for pH, EC, soil moisture, and soil temperature. Potassium required a different treatment: the documented EBK trials produced implausible negative values or very poor error statistics. IDW at 500 m was therefore preferred for K because it avoided negative predictions, despite its relatively high RMSE.

Table 4. Soil Depth: 15–30 cm

Parameter	Method	n	RMSE	Mean error	RMS std.	Avg. SE	90%	95%	CRPS
EC	EBK	280	2.987	0.035	1.010	2.971	89.86%	92.03%	1.611

pH	EBK	280	0.262	0.00016	0.894	0.294	89.49%	95.29%	0.141
Soil temperature	EBK	280	7.044	-0.074	0.963	7.502	92.39%	96.74%	3.637
Soil moisture	EBK	280	7.670	-0.0765	1.010	7.595	91.30%	95.29%	3.975
K	IDW	280	201.785	84.004	—	—	—	—	—
N	IDW	280	187	81.6	—	—	—	—	—
P	IDW	280	166.4	75.1	—	—	—	—	—

For the 15–30 cm layer, EBK produced particularly strong calibration for EC, pH, and soil moisture. The pH model had an RMSE of 0.262 pH units and 95% interval coverage of 95.29%. For EC, the RMS standardized value was 1.010, and the Average Standard Error (2.971) was very close to the RMSE (2.987). The soil-moisture model also showed an RMS standardized value close to 1 and close agreement between Average Standard Error and RMSE.

For potassium at 15–30 cm, the documented final comparison used IDW, with an RMSE of 201.785.

Following interpolation and cross-validation, the selected prediction surfaces were exported as raster datasets at 500 m spatial resolution and prepared for publication. The final products represent the spatial distribution of the measured soil properties for the two soil-depth intervals.

The final workflow therefore follows a transparent sequence from sampling design to map publication:

LULC-based sampling design → Field sampling → Soil measurements → Spatial interpolation → Cross-validation → Final soil maps → Geoportal