

University of Greifswald  
Faculty of Mathematics and Natural Sciences  
Institute of Botany and Landscape Ecology  
Landscape Ecology and Nature Conservation program

**Thesis for the Degree of Master of Science:**  
**Analysis of MaxEnt efficiency for mapping the distribution of  
the *Marmota bobak* in Kazakhstan.**

**Submitted by:** Irina Grigoryeva

Student ID: 166806

**Supervisors:**

Prof. Dr. Sebastian van der Linden

University of Greifswald - Department of Geography and Geology

ACBK Science fellow, PhD Alyona Koshkina

Association for the Conservation of Biodiversity of Kazakhstan

## Content

Introduction	1
The largest rodent in Kazakhstan	1
Remote sensing of burrowing mammals	3
Aerial and satellite monitoring of <i>Marmota bobak</i>	4
Overview of species distribution modelling	5
Materials and methods	8
Study site	8
Visibility of marmot mounds on satellite images	9
Modelling <i>Marmota bobak</i> distribution	12
Collection of observation data	12
Field studies as sources of species presence data	12
Choice of environmental variables	14
Climatic predictors	15
Geographic predictors	16
Landscape predictors	17
Anthropogenic predictors	17
Preprocessing	18
Step 1: Preparing the data for MaxEnt in RStudio	18
Step 2: Cross-validation and model training in RStudio	21
Step 3. Final model training and construction of the distribution map of the <i>Marmota bobak</i>	25
Filtering of predictors	27
Results	31
Validation map for MaxEnt models	38
Discussion	41
Conclusion	47
Acknowledgments	48
References	49
Appendices	55

## **List of abbreviations:**

**MaxEnt** - Maximum Entropy Model - A species modeling algorithm based on the principle of maximum entropy.

**QGIS** - Quantum GIS is a free geographic information system for working with spatial data.

**SRTM** - Shuttle Radar Topography Mission - NASA mission that conducted radar mapping of the Earth's surface, the data from which are used to build digital elevation models.

**CRS** - Coordinate Reference System - A system that defines how geographic coordinates are referenced on a map.

**ROC** - Receiver Operating Characteristic - A graph used to evaluate the quality of binary classifiers, including species distribution models.

**SDM** - Species Distribution Model - Models that predict species range based on known locations and environmental predictors.

**AIC** - Akaike Information Criterion - An index used to compare models and select the best among them.

**AUC** - Area Under the Curve is a measure of model quality based on the ROC curve. In species distribution models, AUC is used to assess the accuracy of predictions.

**CSV file** - Comma-Separated Values - A tabular data format where values are separated by commas (or other separators).

**NDVI** - Normalized Difference Vegetation Index - An index calculated from satellite imagery that reflects the density and condition of vegetation.

**SAVI** - Soil-Adjusted Vegetation Index is a modification of NDVI that takes into account the influence of soil.

**GBIF** - Global Biodiversity Information Facility is an international platform for sharing biodiversity data.

**USSR** - Union of Soviet Socialist Republics - a former socialist state that existed from 1922 to 1991.

**ACBK** - Association for the Conservation of Biodiversity of Kazakhstan.

## Abstract

Steppe marmot (*Marmota bobak*) is a large, social burrowing rodent native to the grass-forb steppes of Eurasia. Currently, nearly the entire population is concentrated in Kazakhstan, occupying vast steppe landscapes, many of which are under cultivation for crop production, potentially affecting the species' range. However, tracking population trends and suitable habitats across such extensive areas remains highly challenging.

In this study, the MaxEnt modeling approach was employed to predict the potential habitat of *Marmota bobak* in the Kazakh steppes. Model validation was performed using a distribution map of the species in Kazakhstan, generated through manual interpretation of freely available satellite imagery from Google and Bing, focusing on the presence of active burrows. The model predicted the potential distribution of the *Marmota bobak* in Kazakhstan with an area under the curve (AUC) of 0.925, indicating high predictive accuracy and substantial agreement with the satellite-based distribution map. However, some discrepancies were observed when comparing the MaxEnt maps with the satellite map. MaxEnt failed to detect a relatively large area in the northern part of the species' range, which is most likely due to incomplete presence data. This suggests that the model relied too heavily on the available occurrence records, making it overly restrictive and preventing it from predicting the species' presence in areas with slightly different environmental conditions.

The analysis highlighted the significant contribution of climatic and landscape factors, with snow cover being among the top five factors influencing distribution. MaxEnt modeling also suggested that the suitable habitat for *Marmota bobak* might be slightly larger than indicated by current presence records, necessitating further research to assess potential range expansion. We suggest that future studies should incorporate grazing pressure maps to evaluate the impact of livestock grazing, as the species frequently inhabits such areas, as well as including more points of presence and points of real absence.

This study contributes to the understanding of spatial distribution modeling for large burrowing rodents across extensive open landscapes.

## Introduction

There is a global trend of declining populations of burrowing rodents due to various drivers: including climate change, hunting, habitat loss, and pesticide use (Davidson et al 2012, Fleming et al 2014, Gabrielle Beca 2021).

Burrowing rodents are key species involved in the functioning of grassland ecosystems. They transport the lower soil layers to the surface thereby mixing soil horizons with each other as well as mixing in their excrement (Villarreal et al. 2008), enhance water infiltration (Miranda et al. 2019), facilitate the distribution of chemical elements (Valentine et al. 2018, Fleming et al 2014), and influence carbon accumulation (Eldridge and Koen 2021, Martínez-Estévez et al. 2013, White et al. 2000).

Such areas create an environment distinct from the rest of the landscape thereby supporting certain vegetation and animal species, as well as increasing biodiversity and plant biomass around the burrows (Davidson and Lightfoot 2006, Davidson et al. 2008, Yoshihara et al. 2009, Hogan 2010). Additionally, burrows of digging mammals are also used by other animals (Murdoch et al. 2009, Davidson et al. 2012, Davidson and Lightfoot 2007, Valkó et al 2020), and the rodents themselves are prey for predators (Davidson et al 2012, Wheeler et al. 2015). Thus, it is evident that the decline in burrowing rodent populations may cause a negative cascading effect (Davidson et al 2012, Fleming et al 2014, Valiente-Banuet et al 2015).

### The largest rodent in Kazakhstan

*Marmota bobak* (Fig. 1) is considered an edicator, an ecosystem engineer and the largest species of social burrowing rodents in the steppes of Kazakhstan. (Zimina 1978, Dudnikov 2021). Similarly to ground squirrels, prairie dogs, other marmots, kangaroo rats, etc., they create steppe patches that differ from the surrounding landscape and vegetation, increasing habitat heterogeneity by forming micro-ecosystems. They influence water infiltration and bring deeper soil layers to the surface, thereby improving soil properties. Additionally, they prevent shrub encroachment, thus supporting the sustainability of pasture systems. (Abaturov B. 1984, Branch et al 1999).

Additionally, the *Marmota bobak* is among three numerous marmot species population in Eurasia (Zimina 1978, Kolesnikov 2011). Detailed information about the

*Marmota bobak*'s range begins to appear in the 18th century. Since then, a decline in the *Marmota bobak* population has been noted due to increased anthropogenic impact (Bibikov 1989, Kirikov 1983).

In the first half of the 20th century, the continued harvesting of pelts and the utilization of marmot meat for subsistence contributed to the decline of the population. (Bibikov 1989, Formozov 1963, Zimina and Isakov 1980). Though the Kazakh steppe was not subjected to large-scale ploughing (Zarubin 1997), the marmot population suffered from overexploitation, which undermined its numbers. In 1932, a hunting ban allowed the population to recover (Shubin 1978).

In the middle of the 20th century, the large-scale ploughing associated with the "Virgin Lands Campaign" in Kazakhstan led to the disruption and fragmentation of marmot settlements, resulting in a population decline of no less than 75% (Rumyantsev 1991). The marmots in the areas of plowed upland steppes suffered the most (Bibikov 1989, Zimina and Isakov 1980).

However, by the mid-1980s, there is evidence of the adaptation of marmots to plowed lands where their settlements were found (Rumyantsev 1991). According to various estimates the population of the *Marmota bobak* in Kazakhstan during that period ranged from 383,300 to 4,239,800 individuals. Although there were significant discrepancies in estimates, cartometric analysis of those years and data on the discovery of settlements on ploughland suggest that the *Marmota bobak* population in Kazakhstan numbered in the hundreds of thousands (Kolesnikov 2011, Rumyantsev 1991). Also, a recent research casts doubt on the perceptions of the significant decline in marmot populations during the period of active agricultural development in the country. According to the latest estimates the population size of the Kazakh *Marmota bobak* is estimated at 6.1 ( $\pm 2.4$ ) million individuals (Koshkina A. et al 2019).



Figure 1. *Marmota bobak* in Kazakhstan. Credit by Alyona Koshkina

### **Remote sensing of burrowing mammals**

Traditionally, the monitoring of burrowing mammals and the traces of their activity is conducted using terrestrial methods based on direct observation and counting of individuals, families, and burrows (Hoffmann et al. 2010). In order to cover the entire habitat of the subject it is necessary to conduct fieldwork across the entire range or extrapolate data which requires significant resources. Not all areas of the range can be surveyed due to remoteness and the accuracy of the obtained data is often low (Kolesnikov 2011). Therefore, many scientists have encountered challenges when studying extensive areas and the emergence of remote sensing methods for monitoring environmental changes has provided a solution to these difficulties.

Since the 1930s, aerial photography has been conducted and after World War II the use of aerial photography for research purposes has increased (Daniel J. 1953).

Use of satellite imagery in wildlife research has been developing since the 1960s as a form of telemetry. In 1971, a study conducted in Wyoming tracked elk movements, demonstrating the potential of this method for studying wildlife displacement (Buechner et al., 1971). Later, researchers utilized the Argos system to produce movement maps of

caribou and polar bears based on data collected through satellite telemetry (Fancy et al., 1988). Since the 1980s, satellite imagery has been used for mapping wildlife habitats (Saxon, 1983).

Given that burrowing mammals significantly alter the landscape they become easy to spot in contrast to their surroundings when using remote sensing methods.

For this reason, aerial photography has been used to survey prairie dog colonies in the United States (Dalsted et al. 1981). In the 1980s, one of the first studies of burrowing mammals, specifically the southern hairy-nosed wombats (*Lasiorhinus latifrons*) in South Australia, was conducted using large-scale Landsat imagery (Lciffler and Margules 1980). Later on, high-resolution satellite imagery was used for studying prairie dog colonies (Sidle et al. 2002).

The popularity of remote sensing of wildlife methods has been increasing. In 2012, this method was applied to study small rodent activity, such as voles and lemmings (Johan Olofsson et al. 2012). In 2018, a study focused on the range expansion of the North American beaver (Ken D. Tape et al. 2018). Additionally, in 2018, another study investigated the distribution and abundance of the southern hairy-nosed wombat (*Lasiorhinus latifrons*) in South Australia, using satellite tracking (Michael J. Swinbourne et al. 2018).

### **Aerial and satellite monitoring of *Marmota bobak***

In the second half of the 20th century, the first attempts to monitor the marmot population in Kazakhstan using aerial photography were made by Soviet scientists; however, they were not successful. (Bibikov, Chekalin 1959, Vinogradov, Leontiev 1985, Rummyantsev 1993).

Then in 2011, 10 freely accessible Google Earth images were used to study marmot settlements in Mongolia, confirming the capability of satellite imagery to identify marmot burrow mounds (Kolesnikov 2011).

However, the use of satellite monitoring methods at the scale of entire populations of *Marmota bobak* was implemented later in Kazakhstan. Based on Google Earth and Bing imagery, 1300 random points were established within the species' range (~950 000 km<sup>2</sup>). A total of 7425 points were mapped via satellite and the reliability of settlement

identification was confirmed through field trips. Based on these results, the population of steppe ground squirrels across their entire range was estimated at 6.1 +/- (2.4) million individuals (Koshkina A. et al 2020). The method demonstrated that the probability of detecting the presence of colonies using satellite imagery was nearly 100%, whereas the probability of detecting individual dens averaged approximately 40%.

Thus, satellite imagery successfully addresses the challenges of analysing species habitats based on continuous raster geographic data (Dubinin M., Kostikova A., 2008). However, it is not always possible to find high-quality open-access satellite images for all research areas. Therefore, over the past forty years, the use of computer algorithms, such as Species Distribution Modelling, has also become widely used for determining the distribution of species in geographic and temporal space (J. Elith et al. 2017). Moreover, habitat modelling appears to be the only feasible approach for studying burrowing rodents whose burrows are not visible in satellite images. Therefore, this study contributes to understanding which predictors are key for modelling the habitats of burrowing species.

### **Overview of species distribution modelling**

Species Distribution Models (SDMs) are predictive tools used to determine the relationships between observed species and environmental predictors through modelling techniques (Srivastava et al., 2019). The method has been evolving since the early 1970s and continues to improve to this day, including advancements in the classification of the niche concept, model formulation, model selection, model evaluation, and variable selection methods (Zimmermann et al., 2010).

Despite existing issues, the method is gaining popularity in the fields of ecology, biogeography, conservation policy, as well as in evolutionary biology, invasive species management, protected area design, and climate change impact forecasting (Guillera-Arroita et al. 2015, Ehrlén and Morris 2015, Ferrier et al. 2016, Zurell et al. 2020). Its widespread use is also attributed to the increased availability of digital data (Franklin et al. 2017, Wüest et al. 2020), user-friendly software packages (Golding et al. 2018), and accessible guides, tutorials, and textbooks (Merow et al. 2013).

Typically, two categories of data are required for modelling:

1. Species data, which can be nominal (presence/absence records), ordinal (ranked abundance), or ratio-based (abundance and richness).
2. Environmental data, which includes information on biotic and abiotic conditions. The most commonly used variables are climatic and topographic, as these datasets represent multi-scale conditions important for modelling (Miller, 2010).

Among the available ecological models, we selected the most popular method for modelling the spatial distribution of living organisms (Lissovsky and Dudov, 2020) – the Maximum Entropy method (MaxENT). MaxEnt is a machine learning algorithm that operates with presence-only data to generate predictions on a geographic scale (Mitchell, 2006). The algorithm is particularly popular not only due to its user-friendly interface, relative ease of use, and ability to generate clear visual results, but also because MaxEnt maximally avoids the uncertainties and uses only presence points, thereby reducing bias in the results (Dhyani et al. 2020). Moreover, the method surpasses similar algorithms in terms of accuracy. Using environmental data as a background, the program maximizes the entropy of the target species. This is akin to maximizing the logarithmic likelihood and excluding the penalty term which is similar to the concept of AIC (T. Shameer, R. Sanil. 2023). However, MaxEnt has its vulnerabilities: 1) a strong dependence on the completeness of the provided presence points of the target species, incomplete data can lead to significant distortions in the results; 2) high sensitivity to background points; 3) a tendency to overestimate (S. Dhyani et al. 2023, A. Lissovsky, and S. Dudov. 2020).

Despite the popularity of the method, there is a notable lack of studies and recommendations on MaxEnt modelling for species with large ranges, which increases the risk of inaccurate predictions (Van Der Wal et al., 2009; Stockwell and Peterson, 2002). Considering that large-scale spatial predictions based on remote sensing data are nearly impossible in some regions (Longhui Lu et al., 2022), the use of SDMs remains the only available approach for studying, conserving, and managing animal populations.

However, understanding that accurate modelling is crucial for conservation efforts, and errors in analysis can lead to ineffective decisions — such as the improper design of protected areas or inadequate control of invasive species (Wiens et al., 2009) — we set

a goal, using verified data on the distribution of the *Marmota bobak*, to assess the reliability of MaxEnt analysis by studying the *Marmota bobak* population in Kazakhstan.

Therefore, in this study, the reliability of MaxEnt analysis is addressed for the range of the *Marmota bobak* population in Kazakhstan, followed by the validation of the obtained results using manual satellite mapping data based on high-resolution satellite images available in the public use.

We set the following research questions:

1. How accurate are the predictions of the MaxEnt model in estimating the range of *Marmota bobak*, and how critical are the discrepancies compared to the remote sensing results?
2. What are the key predictors most significant in defining the range of the *Marmota bobak* in Kazakhstan according to the MaxEnt model?

## Materials and methods

### Study site

The study region encompasses the current distribution range of *Marmota bobak* in Kazakhstan (Fig.2). The climatic conditions of the area are characterized by high continentality, with cold winters and hot, dry summers. Annual precipitation varies depending on the year, ranging from 250 to 400 mm. Temperatures fluctuate across regions; however, the average January temperature in the northern and central parts of the country is around -20°C, and the average July temperature is +19°C (Lydolph, 1965). The predominant soils include fertile chernozems in the north, transitioning to lighter and medium kastanozems to the south, with salt-affected soils such as solonchaks and solonetz commonly found in depressions. The steppe vegetation is primarily dominated by feathergrass (*Stipa*) and fescue (*Festuca*) species, accompanied by a variety of forbs and some *Artemisia* species (Brinkert et al., 2016).

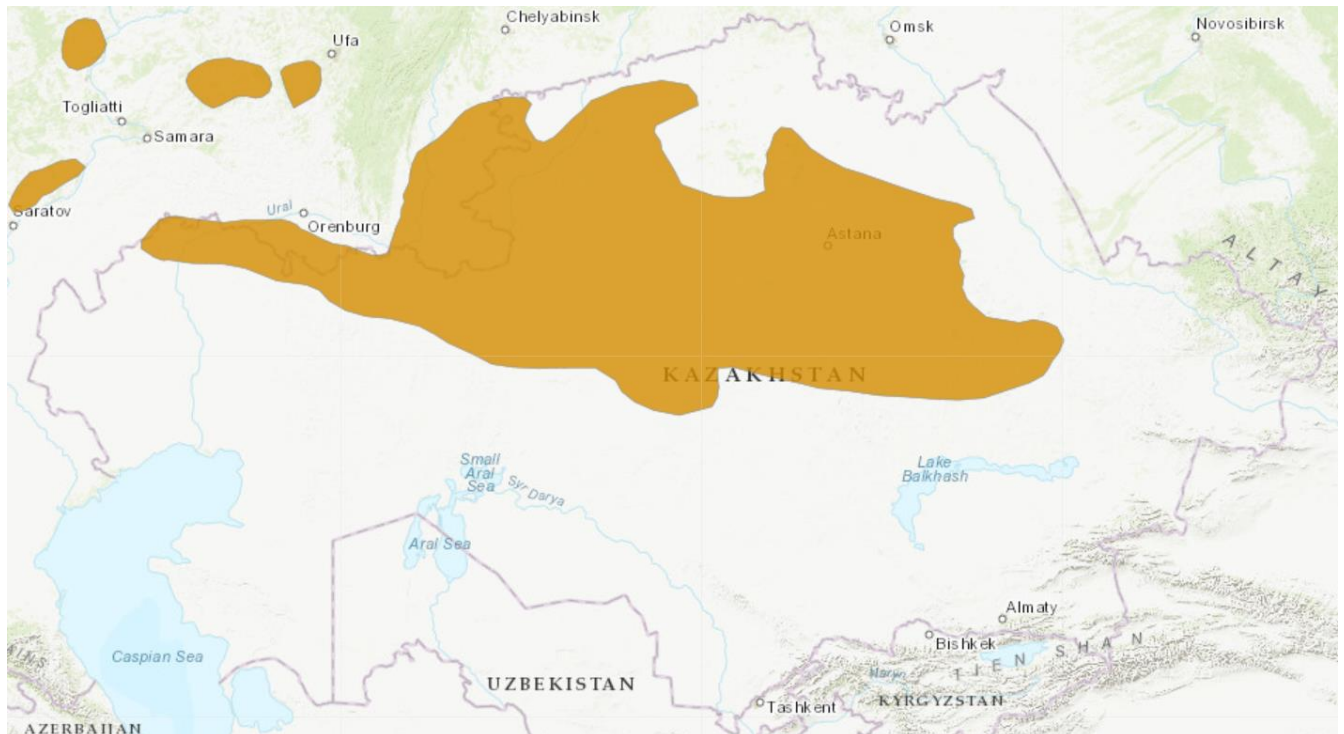


Figure 2. The IUCN distribution map of *Marmota bobak* in Kazakhstan

## Visibility of marmot mounds on satellite images

We utilized pre-existing data from previous studies on the distribution of the *Marmota bobak* researched by local specialists such as Kapitonov V.I. (1966), Sludskii A.A. (1969, 1980), Bibikov D.I. (1989), and Zimina (1967, 1980). In our study, we manually digitized occurrence points of *Marmota bobak* burrows in Kazakhstan and in border regions of southern Russia, where small marmot colonies were recorded in the past. Following the recommendations of Koshkina et al., we employed the OpenLayers plugin in QGIS 2.16.2 for mapping and performed map validation using version 3.28.11. We also utilized freely accessible satellite imagery from Google Earth ([www.maps.google.com](http://www.maps.google.com)) and Bing ([www.bing.com/maps](http://www.bing.com/maps)). The satellite images we used were from the spring and summer seasons, with a preference for either the most recent images or those with the highest resolution.

We overlaid a 5x5 km grid (25 km<sup>2</sup> per cell) onto the study area. This resolution enables more precise analysis of each section for the presence of active marmot burrows, minimizing the risk of missing areas and allowing for a more detailed delineation of the species' range (Fig. 3). The density of marmot colonies varies depending on habitat, with burrow density remaining stable in the northern part of the study area and showing a tendency to decrease towards the south. On average, the distance between burrows is approximately 120-150 meters (Vinogradov, 1985).

The next step involved a reviewer examining each grid cell. If active marmot mounds were detected, the cell was marked with a presence point; if no mounds were found or if colonies had been abandoned, the cell was skipped. To mitigate issues with burrow detection, we analyzed cells at a scale of 1:10,000, and in some cases, when identification was challenging, we used a scale of 1:5,000. This process is labour-intensive and time-consuming; in our case, digitizing the entire study area took approximately one year. In 2018, we validated 667 marmot presence points in the field to assess the methodology and train observers in identifying marmot burrows on satellite imagery (Fig. 4). The identification of marmots was based on examples from similar studies by other researchers utilizing aerial photographs (Vinogradov and Leontieva, 1985; Rummyantsev, 1989) and satellite imagery (Kolesnikov, 2011). Additionally, the methodology drew upon field deciphering data conducted by Koshkina et al. during their fieldwork.

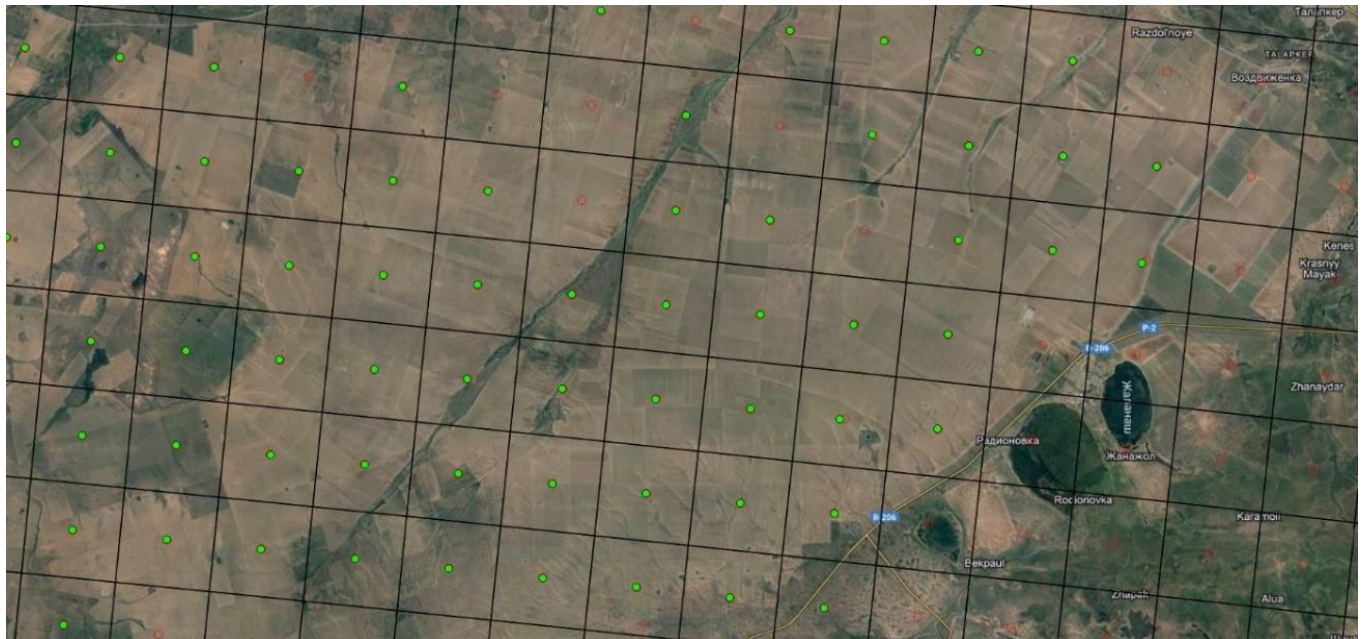


Figure 3. A grid with 5 km cell sides helps avoid double-counting the same family territory of marmots. The green points indicate the presence of active marmot burrows.

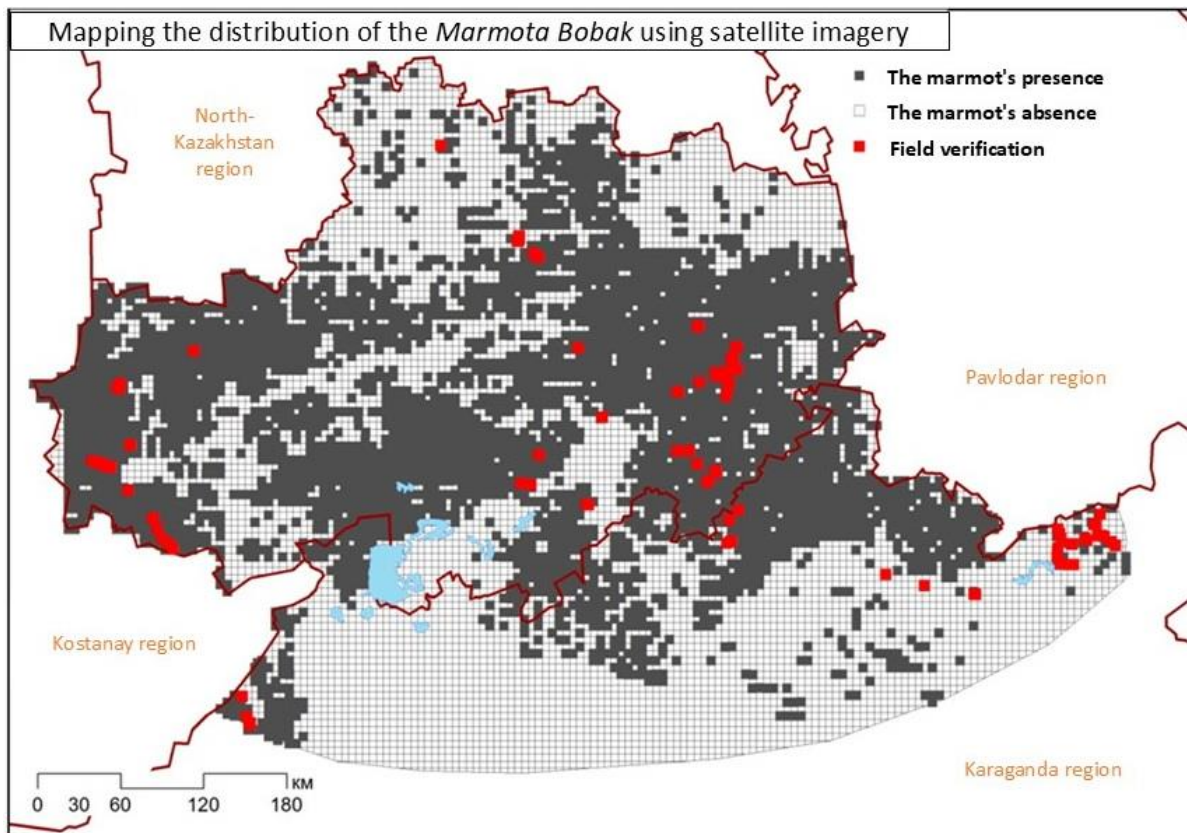


Figure 4. The method was tested in the territories of the Karaganda and Akmola regions.

The boundaries of colonies, the sizes of mounds, and the distinction between active and abandoned marmot burrows are clearly visible in satellite images, along with trails and the type of burrow (Vinogradov and Leontieva, 1985). Abandoned burrows appear darker in the images and exhibit blurred boundaries due to encroachment by annual and biennial vegetation, which is typically not characteristic of the steppe flora. In contrast, active burrows display a lighter hue against the steppe background and have well-defined boundaries owing to the light-coloured soils that marmots excavate during their burrowing activities (Fig. 5A and Fig. 5B) (Vinogradov and Leontieva, 1985).

Based on their dimensions, the mounds of the *Marmota bobak* can be classified as large (over 20 m in diameter), medium (10-20 m), and small (less than 10 m). Mounds with large and medium diameters typically serve as year-round habitats, whereas those with smaller diameters function as protective burrows and are primarily used during the spring and summer months (Vinogradov and Leontieva, 1985).



Figure 5. Examples of *Marmota bobak* burrows: Figure A illustrates abandoned burrows, while Figure B illustrates active burrows. The scale of the images is 1:10,000.

In mapping marmot colonies, it was crucial to document the presence of active burrows used year-round. Therefore, we marked only those grid cells containing large or

medium burrows with a presence point, while burrows with small diameters were either ignored or not visible at the scale we selected.

### **Modelling *Marmota bobak* distribution**

Manual high-density digitization of presence points provides an excellent opportunity for reliably visualizing the population boundaries of the *Marmota bobak*. However, this approach is highly resource-intensive. Therefore, the aim of this study is to attempt to predict the distribution of *Marmota bobak* using the widely recognized species distribution modelling software, MaxEnt.

The territory of Kazakhstan was selected for modelling because it hosts the majority of the global population of the *Marmota bobak*. This species predominantly inhabits steppe areas of the country and demonstrates a preference for grazed landscapes. Given that Kazakhstan historically supported nomadic pastoralism and was home to large wild ungulates, these factors have played a significant role in shaping the habitat of this steppe-dwelling rodent.

### **Collection of observation data**

#### **Field studies as sources of species presence data**

For the analysis, it was decided to use species presence data collected from field surveys and expedition trips starting in 2009. A major advantage of such data is the direct recording of species presence in its natural habitat by specialists equipped with GPS devices for precise georeferencing. Additionally, primary data are verified through the availability of photographs (Fig. 6) and records of habitat conditions (Fig. 7). The total number of presence points for the *Marmota bobak* is 901.

The choice was made in favour of field data without incorporating records from scientific sources or open-access databases such as GBIF or iNaturalist, as both of these sources often exhibit issues with coordinate accuracy, which can vary significantly.



Figure 6. Data collection on the *Marmota bobak* within the BALTRAK project in 2017. Photo. ACBK.



Figure 7. Recording of *Marmota bobak* burrows within the BALTRAK project. Photo. ACBK.

All data were collected during the spring-summer period, when marmots exhibit active surface activity, foraging intensively to accumulate fat reserves for the winter while using their burrows primarily as shelters from predators and for overnight rest. Consequently, researchers also relied on direct visual observations of the animals and their vocalizations for data collection.

A total of 219 points were used for the analysis (Fig.8), recorded in the WGS84 coordinate system using decimal degrees. The dataset was verified in QGIS Desktop 3.28.11 and structured into columns within a CSV file.

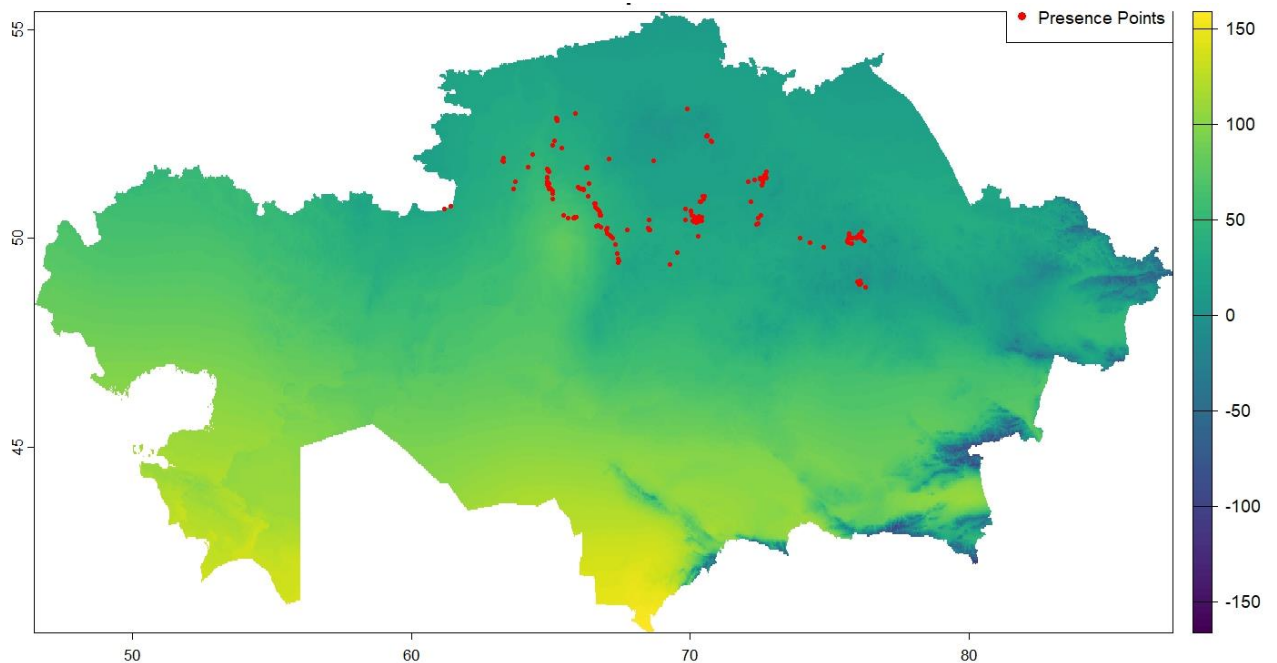


Figure 8. 219 species occurrence points on the map.

### **Choice of environmental variables**

From field research experience and literature sources, it was clear which predictors should be considered for the studied species and what each of them represents. However, it is crucial to assess the individual contribution of each predictor and identify the key variables.

The preparation for the analysis includes the following ecologically justified predictors.

## **Climatic predictors**

Climatic predictors were selected as the primary ecological constraints for species distribution and range formation. The data were obtained from WorldClim (<https://www.worldclim.org/>), which provides global climatic datasets with spatial resolutions (e.g., 30 arc-seconds, ~1 km<sup>2</sup>) and temporal coverage (e.g., 1970–2000) in GeoTIFF format. Additionally, we utilized snow cover data for the years 1982–2014 ([https://silvis.forest.wisc.edu/data/global\\_freeze/](https://silvis.forest.wisc.edu/data/global_freeze/)).

For the marmot, we selected BIO01, BIO05, BIO12, BIO18, and Snow Cover as these variables significantly influence survival, activity, and foraging.

BIO1 = Annual Mean Temperature

The *Marmota bobak* is highly dependent on annual temperature fluctuations, which influence its hibernation and activity periods, as well as the availability of forage vegetation during the spring-summer season. Therefore, temperature serves as a crucial variable shaping the overall climatic conditions within the species' habitat.

BIO5 = Max Temperature of Warmest Month

The predictor represents extreme summer temperatures and is crucial for the marmot due to its impact on vegetation cover. In Kazakhstan, grass often dries out and burns by mid-summer, which may affect the marmot, as this rodent relies on the summer months to accumulate fat reserves before hibernation.

BIO12 = Annual Precipitation

This predictor also has a significant impact on vegetation productivity. Insufficient precipitation limits plant growth and biomass accumulation, while excessive precipitation may lead to waterlogging, which, in turn, affects marmot settlements.

BIO18 = Precipitation of Warmest Quarter

Due to the marmot's activity during the warm season, precipitation in these months contributes to increased grass growth when present and, conversely, to plant desiccation

and burnout in its absence. This, in turn, influences the marmot's ability to accumulate sufficient fat reserves or may lead to foraging challenges.

#### Snow cover duration

The snow cover not only influences vegetation but also affects the rate of snowmelt on the ground surface, which may play a role in the timing of marmots' emergence from hibernation.

NDVI (Normalized Difference Vegetation Index) was obtained from MODIS via Google Earth Engine, with mean values calculated from early April to late August (the active growing season) over the past three years (2022, 2023, and 2024). This predictor is essential for capturing spatial and interannual variations in vegetation.

SAVI (Soil-Adjusted Vegetation Index) not only reflects vegetation changes but also accounts for bare soil areas, which is crucial for steppe ecosystems. SAVI was also derived from MODIS via Google Earth Engine and incorporated data from the active growing season over the past three years.

NDVI and SAVI were selected in addition to the climatic predictors from WorldClim. Despite some degree of correlation, these predictors are not mutually exclusive for use. NDVI and SAVI account for the variability of vegetation cover, which may be influenced by anthropogenic factors and non-climatic variables, whereas climatic predictors provide information on the fundamental environmental conditions (Ya Liu et al 2015, Telnova 2017).

#### **Geographic predictors**

The geographic variables selected for analysis included Elevation and Slope, derived from the Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global dataset. The spatial resolution is 3 arc-seconds for global coverage (approximately 90 m), with data corresponding to the year 2000. These data were obtained through Google Earth Engine.

Additionally, the soil map of the country was provided by the Association for the Conservation of Biodiversity of Kazakhstan.

### Elevation

The *Marmota bobak* prefers steppe habitats at elevations up to 500 meters above sea level and avoids mountainous areas. Additionally, it tends to avoid ravines and other low-lying depressions, as elevation directly influences burrow flooding risk, vegetation composition, and soil type and properties.

### Slope

Marmots prefer flat surfaces where the soil is stable and not subject to erosion. However, burrows are frequently found on road slopes. Therefore, slope was included as a predictor to assess its influence.

## **Landscape predictors**

### Soils

Soils are a crucial predictor for burrowing animals, determining habitat not only through their type, friability, and thermal insulation but also by influencing vegetation diversity.

## **Anthropogenic predictors**

Since the *Marmota bobak* prefers areas with an open view and often establishes its settlements near human habitation, particularly around villages where livestock grazing occurs, it was decided to use the 2020 WorldPop data to examine the correlation between human population density and marmot settlements. The data were accessed through Google Earth Engine.

Additionally, the model includes an important predictor—Land Cover. The data were obtained for the year 2014 from the EarthEnv resource (<https://www.earthenv.org/>) with a resolution of 1 km in GeoTIFF format. This predictor is directly related to the marmot's habitat, influencing food availability, as well as vegetation structure and density. The

marmot prefers stable areas, while human-induced land use changes may alter the species' distribution patterns.

## **Preprocessing**

The preparation of predictor layers was conducted using Google Earth Engine in JavaScript. Most maps were loaded directly, while layers such as soil data and national boundaries were uploaded manually via a personal account.

Vector data (such as the administrative boundary map of Kazakhstan) and raster data (such as soil maps) were uploaded to the platform. Predictors were clipped to the boundaries of the study area. All layers were converted to one of the most commonly used geographic projections, WGS84 (EPSG:4326), which is also widely supported by almost all GIS software. The output data were resampled using the Code Editor and saved at a resolution of 1000 meters (1 km) to avoid unnecessary detail, which could introduce noise, and to ensure optimal suitability for MaxEnt modelling. After processing the maps to match the study area, further data preparation, including the refinement of occurrence points for the target species, was conducted in RStudio version 4.4.2.

### **Step 1: Preparing the data for MaxEnt in RStudio**

Further data processing for analysis is conducted in R. This software was chosen due to its capabilities, including step-by-step process tracking enabled by open-source code, analytical flexibility, integration of tools from various packages, handling of large and complex datasets, relatively easy preparation of all necessary data—from point and raster processing to background point generation—analysis reproducibility, package version control to prevent changes in future processing or model replication, independent selection of optimal modelling analyses, excellent visualization, compatibility with GIS software, and the fact that both the software and its packages are free of charge.

Data preparation before running the modelling process is a critically important step and consists of the following stages:

1. The installation of packages is performed using the “install.packages” function. In this analysis, it is necessary to install the terra package, which is a modern package designed for working with raster data and spatial objects.

2. The Raster package is also designed for working with raster data. Although it was used before the development of the modern terra package, it is still frequently utilized for converting objects between different formats and integrating with terra.
3. Dismo – an R package for ecological modelling, generating background points, and visualization in MaxEnt.
4. Corrplot – the primary function of this package is to construct and visualize a correlation matrix.
5. The Sp package is a widely used library for spatial data. It can be loaded separately, but more modern alternatives are already available within the terra package. Additionally, the sf package provides the same functionality with updated features.
6. Dplyr – a package for fast preprocessing of tabular data before analysis.
7. spThin – an R package for filtering occurrence points.

Setting the working directory for data retrieval and storage of results during the analysis process using the setwd function.

Step one: Processing species occurrence points:

1. Loading species presence points within the study area in CSV table format, structured with two columns: *lon* and *lat*, as our analysis requires only the coordinates from the points.
2. Thinning of presence points using the "thin" function based on distance, with a filtering distance of 500 meters in this case. This step removes spatially clustered points to reduce data redundancy, preventing model overfitting and thereby improving its performance. The 500-meter filtering distance is optimal as it increases the likelihood of capturing a single burrow per family territory, helps avoid spatial averaging, and preserves a sufficient number of data points for analysis. In contrast, increasing the distance threshold may lead to a significant reduction in the number of presence points, potentially making the model unstable.

3. Convert the points into a SpatVector object and set the coordinate reference system to WGS84/EPSSG:4326.

Step Two: Loading 19 Selected Raster Predictors Clipped to the Borders of Kazakhstan in the WGS84/EPSSG:4326 Coordinate System in GeoTIFF Format.

Stages:

1. After adding all raster predictors, one file is selected as the reference (ensuring it has the required projection, resolution, and extent) to align the other predictor layers. Even if all datasets were pre-processed with identical parameters, it is recommended to perform this step again to avoid potential layer misalignment issues during analysis and ensure accurate modelling.
2. Using the functions crop (clipping to a reference), resample (adjusting resolution to a reference), and extend (expanding to a reference), we precisely align the layers to a uniform value.
3. Aligned raster layers are merged into a single stack to facilitate analysis in MaxEnt.

Step three: Removal of sublayers and correlation check of key layers.:

1. The "names" function is utilized to extract layer names, ensuring the correct loading of rasters, as climatic variables often contain sublayers, thereby increasing the number of predictors. Consequently, instead of 19 primary layers, 91 were generated, necessitating filtration to restore the original state.
2. Predictor filtering is performed using the "grep" expression, retaining only the primary key layers.
3. Creation of a unified stack from filtered key layers.
4. To assess layer correlation, standard values are extracted from 10,000 pixels using the "spatSample" function. A correlation matrix is then computed with the "cor" function and subsequently visualized. The final step involves removing highly correlated layers based on the standard ecological threshold of  $|r| > 0.7$ . (J. Elith et al. 2017). In this study, the decision was made to retain these data layers due to their diverse ecological information, which is essential for the species. Moreover,

MaxEnt is capable of handling highly correlated predictors by selecting the most relevant ones and reducing the influence of less informative variables.

Step four: alignment of the Coordinate Reference System (CRS), filtering of presence points with NA values, and generation of background points. It is essential to ensure that the assembled stack of predictors and the filtered presence points are in the same CRS. If discrepancies are present, the points are reprojected to match the coordinate system of the rasters:

1. Removal of NA points: in this case, five points are located outside the study area. To prevent errors in modelling, these points are excluded, leaving 219 points for analysis.
2. MaxEnt is capable of generating background points independently; however, it is preferable to generate the background manually. A set of 10,000 points is also an optimal choice for large-scale areas. Background points play a crucial role in estimating species presence probability and enhancing the representativeness of the model (J. Elith et al. 2017).

Step five: the final set of points is generated.

1. Coordinates for actual presence points and background points are obtained and subsequently used to create separate data tables.
2. An additional verification of points containing NA is performed, and removal is conducted as necessary.

The implementation of these steps results in a balanced set of essential data for modelling the distribution of the *Marmota bobak* in Kazakhstan.

## **Step 2: Cross-validation and model training in RStudio**

Cross-validation is a tool employed to evaluate the performance of a model and determine its effectiveness in accomplishing the specified task. Cross-validation is divided into several components, with one component dedicated to training the model and another to evaluating the model.

Cross-validation encompasses several types, each characterized by specific approaches to handling data depending on volume and repetition cycles. In this study, the **K-Fold Cross-Validation** method was employed, following the algorithm outlined below:

- A certain integer  $k$  (typically ranging from 5 to 10) is defined.
- The dataset is divided into  $k$  equal parts, referred to as folds.
- Next,  $k$  iterations are performed, during each of which one-fold serves as the test set, while the combination of the remaining folds acts as the training set. The model is trained on  $k-1$  folds and evaluated on the remaining one. (Phillips, 2021).

Before initiating the cross-validation cycle, an additional verification of the presence and background point coordinates is performed to ensure that the absence of columns with correct data does not lead to errors. Next, the dataset is partitioned into folds ( $k = 5$ ), and to enable result reproducibility, the initial state of the random number generator is set (`set.seed(123)`). The use of five folds is a standard approach to balancing result accuracy and computational time; however, even this configuration remains resource- and time-intensive, requiring substantial system capabilities, which limits its frequent application. In this study, the chosen number of folds ensures a sufficient amount of data in each subset for model training and validation. However, the randomization of data usage presents a drawback, as certain data points may be omitted (Robert J. Hijmans, 2012). The division of presence points and background into folds is performed using code

```
presence_df$fold <- kfold(presence_df, k = k)
background_df$fold <- kfold(background_df, k = k)
```

After executing the command, each point is assigned a number corresponding to one of the five folds. At each iteration, one group of data (four folds) is used for model testing, while the remaining fold is used for training. This process is repeated five times, corresponding to the number of folds.

At the next stage, an array is created to store AUC values (`auc_values <- numeric(k)`). The vector `auc_values` is specifically designed to store numerical values obtained during the cross-validation process, with a separate AUC value calculated and stored for each fold  $k$ . Subsequently, the array utilizes the mean AUC from five folds. The mean value

serves as a key metric for evaluating the overall performance of the model (Cory Merow et al.2013).

AUC (Area Under the Curve) is a metric that quantifies model performance with a single value by measuring the area under the Receiver Operating Characteristic (ROC) curve (Melo 2013). It evaluates the model's ability to distinguish between two categories, in this case, presence points and background points.

AUC ranges from zero to one, providing an assessment of the quality of species distribution predictions:

- 0.5 – The model performs at a random guessing level;
- < 0.5 – The model performs worse than random guessing;
- 0.7 – 0.9 – The model demonstrates reasonably good predictive performance;
- > 0.9 – The model demonstrates high predictive accuracy (Cory Merow et al.2013).

In this study, the use of AUC remains crucial for an accurate model assessment due to a significant imbalance in the dataset, consisting of 219 presence points and 10,000 background points. Upon execution of the code line, the array becomes ready for further computations.

The cross-validation cycle is then initiated through the execution of the corresponding line of code:

```
# Loop for performing cross-validation
for (i in 1:k) {
  cat("Starting processing of fold", i, "\n")
  # Splitting train/test data for presence points
  train_p <- presence_df[presence_df$fold != i, c("x", "y")]
  test_p <- presence_df[presence_df$fold == i, c("x", "y")]
  # Splitting train/test data for background points
  train_b <- background_df[background_df$fold != i, c("x", "y")]
  test_b <- background_df[background_df$fold == i, c("x", "y")]
}
```

After executing this code, the data is partitioned into four training folds ("train\_p") and one test fold ("test\_p"). Upon completion of the loop, the MaxEnt model training process can commence.

```
# MaxEnt model training
me <- tryCatch({
  maxent(
    x = filtered_stack_raster,
    p = train_p,
    a = train_b
  )
}, error = function(e) {
  cat("Error occurred during model training on the fold", i, ":", e$message, "\n")
})
```

```

return(NULL)
})

```

At this stage of the code, a model is trained using four training folds, where:

- x = This is a set of environmental variables compiled into a raster stack;
- p = This variable represents the coordinates of species presence detection points;
- a = These are the coordinates of background points, primarily serving as inputs for modelling species presence probability.

At this stage, the foundation for further analysis of the distribution of the *Marmota bobak* is established, with a predictive map integrating presence points and the surrounding environment.

The code proceeds with model evaluation:

```

# Evaluation of the model on test data
eval_me <- tryCatch({
  evaluate(
    p = test_p,
    a = test_b,
    model = me,
    x = filtered_stack_raster
  )
}, error = function(e) {
  cat("Error in model evaluation on the fold ", i, ":", e$message, "\n")
  return(NULL)
})

```

- x = This is the same set of environmental variables compiled into a raster stack;
- p = This variable represents the coordinates of species presence points that were not used for training;
- a = These are background point coordinates not used for training;
- model = The trained MaxEnt model.

This step enables the calculation of the AUC value, which indicates model performance and serves as a key criterion for determining the suitability of the data for further analysis.

Further refinement of the code focuses on evaluating the model for each fold, storing the AUC values in an array, and calculating the mean AUC across all folds as an indicator of model performance (Cory Merow et al.2013).

The first part of the code verifies the model evaluation results. In case of an error, the output is set to NULL. If the process completes successfully, the metric data are stored in an array, while in the event of an error, the values are assigned as NA

```

if (!is.null(eval_me)) {
  auc_values[i] <- eval_me@auc
  cat("Fold =", i, "AUC =", eval_me@auc, "\n")
}

```

```

} else {
  auc_values[i] <- NA
}
}

```

The second part of the code is dedicated to calculating the mean AUC across all folds and displaying the obtained value.

```

# The mean AUC is calculated
mean_auc <- mean(auc_values, na.rm = TRUE)
cat("Mean AUC across ", k, " folds:", mean_auc, "\n")

```

After obtaining the value, the analysis proceeds to the next step.

### **Step 3. Final model training and construction of the distribution map of the *Marmota bobak*.**

The final model training represents a crucial step in species distribution mapping, requiring maximum accuracy and completeness. This step involves utilizing the full dataset without partitioning into training and test subsets, as done in cross-validation. Such an approach is essential not only for generating the species distribution map but also for estimating its probability of occurrence in adjacent areas. At this stage, the AUC metric reflects the model's performance when applied to the complete dataset, whereas in cross-validation, AUC indicates model stability. The outcome of the full model training is a unified final map illustrating the species distribution across the area of interest (Cory Merow et al.2013, M. Ahmad et al.2023).

```

final_maxent_model <- tryCatch({
  maxent(
    x = filtered_stack_raster, # Raster stack of environmental predictors
    p = presence_df[, c("x", "y")], # Coordinates of presence points
    a = background_df[, c("x", "y")] # Coordinates of background points
  )
}, error = function(e) {
  cat("Error during the training of the final model:", e$message, "\n")
  return(NULL)
})

```

In the lines "X, P, A," the final model processes the complete dataset. The tryCatch function safeguards the program against failures caused by errors, while the cat function outputs a warning in the console in the event of such an error.

Upon completion of the training process, it is necessary to obtain results on the importance of variables, which serve as an additional assessment of model accuracy

alongside AUC. Therefore, Contribution, % (Fig. 10) is calculated using the following code:

```
var_contribution <- final_maxent_model@results[grep("contribution",  
rownames(final_maxent_model@results)), , drop = FALSE]
```

This presents data on the contribution of each variable to model construction. Permutation importance (%) serves as the second metric for assessing variable significance and is calculated using the following code:

```
var_importance <- final_maxent_model@results[grep("permutation.importance",  
rownames(final_maxent_model@results)), , drop = FALSE]
```

Each variable is sequentially "shuffled," followed by an assessment of the resulting decline in model performance. A greater reduction in model accuracy indicates a stronger contribution of the variable to the prediction.

The analysis results collectively indicate which of the 19 predictors exert the greatest influence on the model and assess their statistical significance.

Next, the AUC is calculated to assess the quality of the final model using the same evaluation code as applied during cross-validation:

```
eval_final <- evaluate( # Model evaluation  
  p = presence_df[, c("x", "y")],  
  a = background_df[, c("x", "y")],  
  model = final_maxent_model,  
  x = filtered_stack_raster  
)  
cat("AUC (Final Model) =", eval_final@auc, "\n") #AUC of Final Model.
```

The next step involves constructing response curves for each predictor using the response command. The principle of operation is based on displaying the probability of species presence as a function of the variable. This serves as an additional analysis for assessing predictor significance and enables the elimination of less influential predictors.

Response curves become a crucial addition to previous analyses. However, it is also essential to assess the number of presence errors omitted by the model. If the omission rate is high, the input data must be adjusted accordingly.

The calculation code for omission and predicted area is used for this purpose:

```
for (i in seq_along(thresholds)) { # Iteration over all threshold values  
  binary_map_temp <- prediction_map_terra >= thresholds[i] # Temporary binary map testing threshold  
  values
```

```

omission[i] <- mean(terra::extract(binary_map_temp, presence_points_spat)[, 2] == 0, na.rm = TRUE) #
Extraction of values from presence points, identification of unsuitable points, and calculation of their mean
proportion
predicted_area[i] <- sum(values(binary_map_temp), na.rm = TRUE) / ncell(binary_map_temp) #
Estimation of the proportion of predicted suitable habitat
}

```

The analysis, after evaluating all probability thresholds, provides information on the proportion of the area of interest classified as suitable at a given threshold and assesses the model's accuracy in predicting known occurrence points. The plot is obtained through a direct link to the MaxEnt website, displaying information via the function `print(final_maxent_model)`.

The final steps involve constructing a habitat suitability probability map for the *Marmota bobak*. This process is carried out using the following code:

```

prediction_map <- predict(final_maxent_model, filtered_stack_raster) - # A raster object with suitability
values is being created.
plot(prediction_map, main = "MaxEnt habitat suitability (Final Model)") - # Constructing the map itself.
points(presence_points_spat, col="red", pch=16, cex=0.5) -# Adding occurrence points to the map

```

The final step in constructing a "hard" model is the binary map, which transforms the prediction map into a clear delineation between suitable and unsuitable areas for the marmot. The code:

```

writeRaster(prediction_map, "MaxEnt_prediction_final.tif", overwrite=TRUE) # The prediction map is saved
for the construction of a binary map.
binary_map <- prediction_map >= threshold(eval_final, "spec_sens") # A raster object is created with
habitat suitability threshold calculation and a logical raster for delineating clear boundaries between zones.
plot(binary_map, main = paste("Binary map (Threshold =", round(thr, 3), ")")) # Map visualization is
performed
writeRaster(binary_map, "MaxEnt_prediction_binary.tif", overwrite=TRUE) # Save in GeoTIFF format

```

Upon completion of the code, the output is a binary map with well-defined boundaries of the most suitable areas for the *Marmota bobak*.

### Filtering of predictors

Upon obtaining the initial final distribution map of the target species, additional adjustments to the selected predictors are required. Typically, this involves excluding variables with zero contribution and/or incorporating additional variables if available. Five rounds of eliminating low-significance predictors were conducted to assess the presence and nature of any changes (M. Ahmad et al.2023).

Initially, data availability must be verified. To enable the filtering of significant predictors, it is essential to ensure that variables are present and correctly recorded; otherwise, the script cannot proceed:

```
if (!exists("var_contribution") || !exists("var_importance")) {  
  stop("Variable contribution or importance is not defined. Please check the MaxEnt model")  
} - # Error return upon its presence
```

The data have been recorded correctly, and the next step requires setting a threshold value to filter out variables with zero contribution to the model. In this case, a threshold of 1 is applied.

```
threshold_importance <- 1 # Contributions or significance below 1% will be filtered out
```

The process of identifying predictors exceeding the specified threshold is initiated.

```
significant_by_contribution <- rownames(var_contribution)[var_contribution[, 1] >= threshold_importance]  
# A list of predictors with values exceeding 1 is generated, and their names are extracted  
significant_by_permutation <- rownames(var_importance)[var_importance[, 1] >= threshold_importance] #  
Generating a list of predictors with permutation importance greater than 1 and extracting their names.  
significant_predictors <- unique(c(significant_by_contribution, significant_by_permutation)) # combine the  
results
```

The following provides names in a standardized format to ensure correct matching and integration into a unified stack:

```
clean_predictors <- gsub("\\.contribution$|\\.permutation.importance$", "", significant_predictors)  
clean_predictors <- gsub("_", " ", clean_predictors)  
clean_predictors <- gsub("\\.tif$", "", clean_predictors) # Clearing the names  
stack_names <- names(filtered_stack) # Putting it together in one stack  
library(stringdist) # Metrics are compared through this package to avoid errors in case of mismatched  
names  
match_table <- sapply(clean_predictors, function(cp) {  
  distances <- stringdist::stringdist(cp, stack_names)  
  best_match <- stack_names[which.min(distances)]  
  return(best_match)  
})  
valid_predictors <- unique(match_table) # A stack is created after verification.
```

The final filtering is performed to eliminate variables close to zero from the stacks.

```
filtered_stack <- filtered_stack[[valid_predictors]]
```

The variables that have passed the filtration process are being checked to ensure that the procedure has been executed correctly.

```
cat("The remaining layers after filtering:\n")  
print(names(filtered_stack))  
cat("The number of remaining layers:", nlyr(filtered_stack), "\n")
```

The format is changed from SpatRaster to RasterStack, as MaxEnt performs better with data from the raster package than with terra.

```
filtered_stack_raster <- raster::stack(filtered_stack)
```

Once the variables are filtered and properly saved, the MaxEnt model must be retrained. The script code thereafter is identical to the one used during the initial training and visualization:

### Training the model

```
final_maxent_model <- maxent(  
  x = filtered_stack_raster, # Stack of significant predictors;  
  p = presence_df[, c("x", "y")], #Presence points;  
  a = background_df[, c("x", "y")] #Background points.  
)
```

Then, the contribution of variables and their permutation importance are assessed:

```
var_contribution <- final_maxent_model@results[grep("contribution",  
rownames(final_maxent_model@results)), , drop = FALSE]  
var_importance <- final_maxent_model@results[grep("permutation.importance",  
rownames(final_maxent_model@results)), , drop = FALSE]
```

Visualize the obtained result:

```
library(ggplot2) # Package for visualization, in this case constructing a diagram.  
var_contribution_df <- data.frame(Variable = rownames(var_contribution), Contribution = var_contribution[,  
1])  
ggplot(var_contribution_df, aes(x = reorder(Variable, Contribution), y = Contribution)) +  
  geom_bar(stat = "identity", fill = "steelblue") +  
  coord_flip() +  
  labs(title = "Contribution of Variables in Final Model", x = "Variables", y = "Contribution (%)") +  
  theme_minimal()
```

The AUC is being evaluated:

```
eval_final <- evaluate(  
  p = presence_df[, c("x", "y")],  
  a = background_df[, c("x", "y")],  
  model = final_maxent_model,  
  x = filtered_stack_raster  
)  
cat("AUC (Final Model) =", eval_final@auc, "\n")
```

In addition, response curves are generated to visualize the influence of each remaining variable:

```
predictor_names <- names(filtered_stack_raster)  
par(mfrow = c(ceiling(sqrt(length(predictor_names))), ceiling(sqrt(length(predictor_names)))),  
    mai = c(0.6, 0.6, 0.6, 0.4))  
for (predictor in predictor_names) {
```

```

response(final_maxent_model, var = predictor, col = "blue", rug = TRUE, main = paste("Response Curve
for", predictor))
}

```

The Omission plot is obtained via a direct link to the MaxEnt website, with the information displayed using the function `print(final_maxent_model)`.

The first suitability map for the marmot is constructed considering the current variables, using the same approach as for the first final map of the team, and the raster is saved for further use:

```

prediction_map <- predict(final_maxent_model, filtered_stack_raster)
plot(prediction_map, main = "MaxEnt habitat suitability (Model_2)")
points(presence_points_spat, col="red", pch=16, cex=0.5)
writeRaster(prediction_map, "MaxEnt_prediction_final.tif", overwrite=TRUE)

```

The final step involves creating a binary map with clear boundaries of the predicted suitable areas for the marmot, visualizing and saving it.

```

thr <- threshold(eval_final, "spec_sens")
binary_map <- prediction_map >= thr # The optimal threshold for habitat suitability is determined
plot(binary_map, main = paste("Binary map (Threshold =", round(thr, 3), ")"))
points(presence_points_spat, col="red", pch=16, cex=0.5)
writeRaster(binary_map, "MaxEnt_prediction_binary.tif", overwrite=TRUE)

```

Following a similar structure and code, three additional rounds of filtering predictors with the smallest contribution to the model are performed. Variables with an importance value less than 1 are excluded; however, in the 3rd and 4th rounds, the threshold is increased to 2 (`threshold_importance <- 2`), and in the 5th round, the threshold is set to 4 (`threshold_importance <- 4`) since the model shows the smallest contribution from variables with values of 1 and 3. The code remains unchanged.

## Results

This section presents the results of the *Marmota bobak* distribution modelling in Kazakhstan using MaxEnt analysis in RStudio, as well as a map generated from manual processing of marmot presence data derived from freely available satellite images from Google and Bing, which were processed using QGIS. This map will serve as the validation for the MaxEnt result.

Using the dismo package, we obtained a trained model, assessed model quality using AUC, calculated omission rates, determined variable importance, and constructed and visualized suitability and binary maps. Initially, 19 predictors were processed (Fig. 9). Although the correlation plot indicates that some variables exhibit both positive and negative correlations, all predictors were retained, as each provides essential information for the analysis. Using the sp, sf, and terra packages, 219 species occurrence points were processed (Fig. 10A). Additionally, 10,000 background points were manually generated across the entire country using the randomPoints function from the dismo package, as this approach is considered more reliable than the automatic function (Fig. 10B).

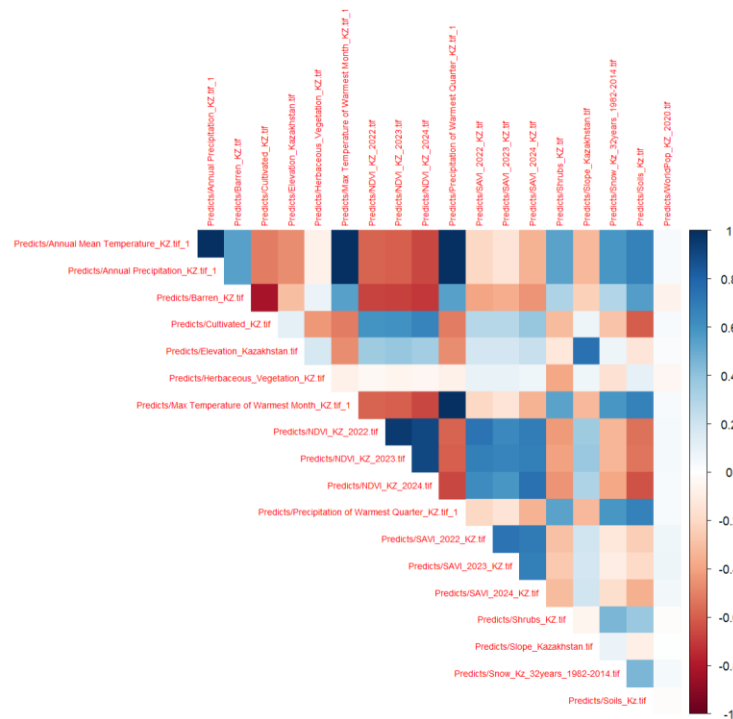


Figure 9. Correlation of predictor layers

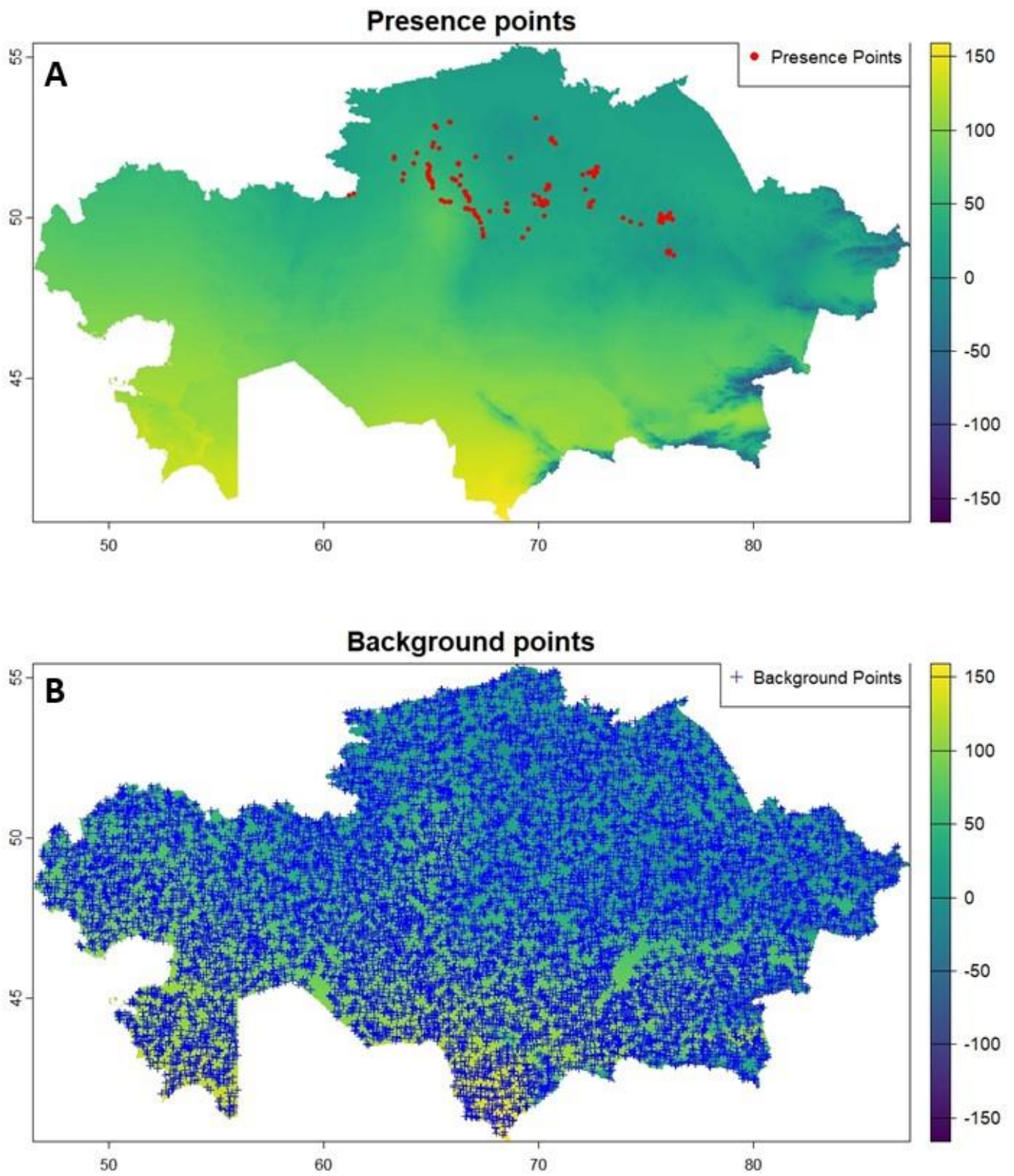


Figure 10. A) Map displaying 219 presence points. B) Map displaying 10,000 background points.

The model is now undergoing training, with cross-validation results across 5 folds as follows:

Processing of fold 1 started  
Fold = 1 AUC = 0.9257955  
Processing of fold 2 started  
Fold = 2 AUC = 0.9137841  
Processing of fold 3 started  
Fold = 3 AUC = 0.9177907  
Processing of fold 4 started  
Fold = 4 AUC = 0.9151932  
Processing of fold 5 started  
Fold = 5 AUC = 0.9230682

Calculating the mean AUC  
Mean AUC across 5 folds: 0.9191263

The result ranges from 0.91 to 0.93 indicates data stability, while the average AUC of 0.919 suggests that the model effectively distinguishes between species presence and absence, demonstrating high discriminatory power.

The final training is initiated, and it shows a result exceeding 0.9 in the first round of analysis. To track changes in the model, predictors with the least contribution are gradually excluded while monitoring the AUC. The metric remains within a stable range, with an average value of 0.923 across all analysis rounds (Table 1). Graphs can be found in Appendix A.

<b>Rounds of model</b>	<b>AUC</b>
Model 1	0,932
Model 2	0,925
Model 3	0,925
Model 4	0,920
Model 5	0,916
Mean value	0,923

Table 1. AUC for 5 MaxEnt analyses and their average value

The greatest contribution to the model was made by average annual temperature with a mean value of 24.5%, arable land at 17.4%, elevation at 14.1%, soil type at 13%, and snow cover at 9.9%. The least significant variables in the model were NDVI and SAVI data, contributing no more than 2% during the initial rounds of analysis. Temperature for

the warm month and precipitation for the warm quarter were excluded immediately after the first round, as were shrubs and slope, which showed minimal significance. By the end of the fifth analysis cycle, barren land was identified as the least significant variable, contributing only 2.2% in the fifth round, with an average contribution of 2.28% (Table 2).

Predictors	Round (%)				
	1	2	3	4	5
Annual.Mean.Temperature_KZ	22	24	25	26	25.6
Cultivated_KZ	17	17.3	17.1	18.1	17.9
Elevation_Kazakhstan	12.7	13.8	14.5	14.9	15
Soils_Kz	11.7	12.7	12.8	13.6	14.5
Snow_Kz_32years_1982.2014	9.7	9.9	10.6	9.5	9.8
NDVI_KZ_2022	7.6	7.3	7.5	7.4	9.2
Herbaceous_Vegetation_KZ	5.3	5	5.1	5.2	5.4
Annual.Precipitation_KZ	3.1	0.7	excluded	excluded	excluded
Barren_KZ	2.6	2.1	2.1	2.4	2.2
NDVI_KZ_2023	2	1.7	2.1	2	excluded
WorldPop_KZ_2020	1.9	1.7	1.9	excluded	excluded
SAVI_2023_KZ	1.5	1.6	excluded	excluded	excluded
SAVI_2022_KZ	1.3	1.3	excluded	excluded	excluded
Slope_Kazakhstan	1	excluded	excluded	excluded	excluded
Max.Temperature.of.Warmest.Month_KZ	0.4	excluded	excluded	excluded	excluded
NDVI_KZ_2024	0.2	excluded	excluded	excluded	excluded
Shrubs_KZ	0.1	excluded	excluded	excluded	excluded
SAVI_2024_KZ	0	excluded	excluded	excluded	excluded
Precipitation.of.Warmest.Quarter_KZ.	0	excluded	excluded	excluded	excluded

Table 2. Percent Contribution of Variables across 5 Rounds

Response curves generally demonstrate stability, with standard deviations not exceeding 0.13, which is acceptable. The most stable variable is soil ( $0.282 \pm 0.062$ ),

followed by elevation, which remains relatively stable ( $0.710 \pm 0.088$ ). Annual Mean Temperature ( $0.440 \pm 0.094$ ) and Annual Precipitation ( $0.505 \pm 0.095$ ) exhibit some fluctuations, while arable land shows the greatest variability among the variables presented ( $0.628 \pm 0.126$ ), although still slight (Table 3). Response curve graphs can be found in Appendix A.

<b>Predictor</b>	<b>Mean Response</b>	<b>Standard Deviation</b>
Annual Mean Temperature	0.44	0.094
Annual Precipitation	0.505	0.095
Barren	0.312	0.064
Cultivated	0.628	0.126
Elevation	0.71	0.088
Herbaceous Vegetation	0.54	0.074
Max Temperature of Warmest Month	0.405	0.065
NDVI_2022	0.198	0.042
NDVI_2023	0.2175	0.0525
NDVI_2024	0.229	0.063
Precipitation of Warmest Quarter	0.39	0.075
SAVI_2022	0.27	0.045
SAVI_2023	0.27	0.06
SAVI_2024	0.28	0.07
Shrubs	0.35	0.09
Slope	0.3	0.08
Snow 32 years	0.412	0.078
Soils	0.282	0.062
WorldPop 2020	0.3	0.07

Table 3. Averaged response curves data for 5 iterations

Model stability assessment through omission testing shows that the mean value ranges from 0.376 to 0.382. The model is optimal but exhibits a conservative bias, with approximately 30% of presence points not predicted by the model. The mean Predicted Area value ranges from 0.266 to 0.273, representing approximately 30% of the predicted suitable area. However, this is a reasonable value, indicating that the delineation of the range is not overestimated. Average results for each model are provided in Table 4. Graphs are available in Appendix A.

<b>Model</b>	<b>Omission</b>	<b>Predicted Area</b>
Model 1	0.379	0.273
Model 2	0.382	0.270
Model 3	0.376	0.271
Model 4	0.379	0.269
Model 5	0.381	0.266

Table 4. Averaged values of model throughput capacity

The final probability presence map derived from the obtained values clearly demonstrates the overall trend with minimal deviations. Upon removal of predictors with negligible or zero contribution, the map consolidates areas of suitability that were already identified in the initial modelling. However, with the binarization of maps, it becomes apparent that the threshold fluctuates from 0.395, increasing to 0.401 in the third model, and then becomes "softer" in the fifth model, decreasing to 0.265, thereby expanding the predicted area of suitable habitats for the *Marmota bobak* (Fig 11, Fig 12). A more detailed view of all maps is provided in Appendix B.

**Binary map (Threshold = 0.395 )**

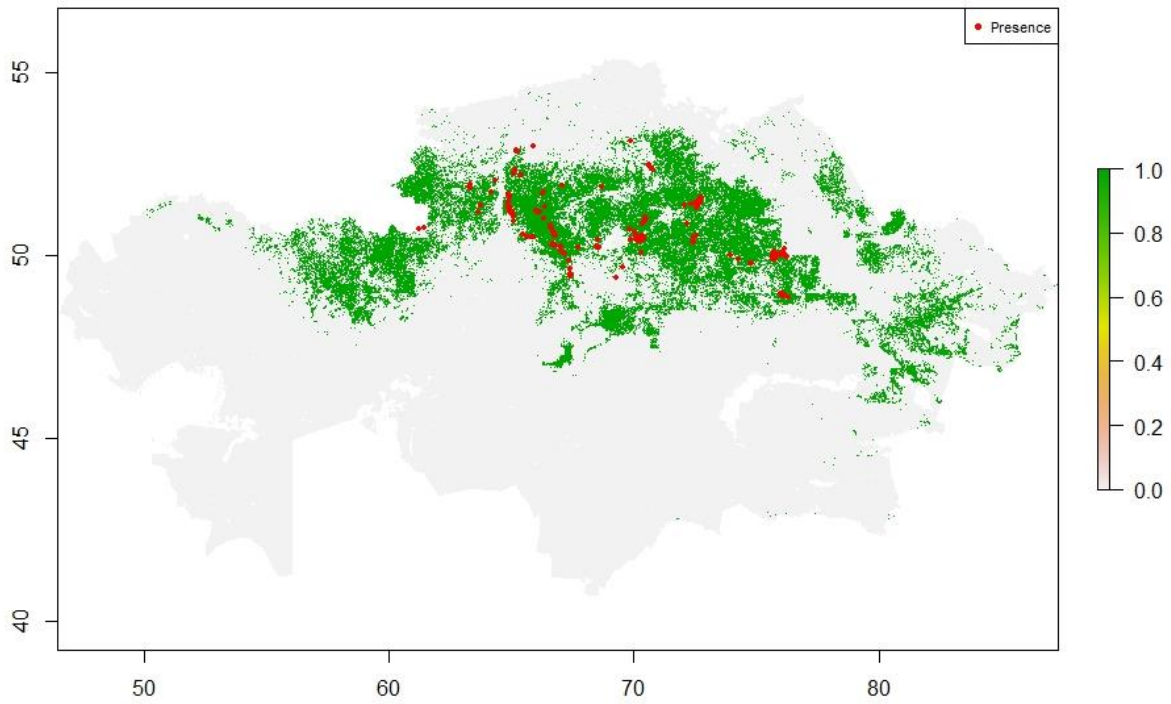


Figure 11. Binary map of the first model

**Binary map (Threshold = 0.265 )**

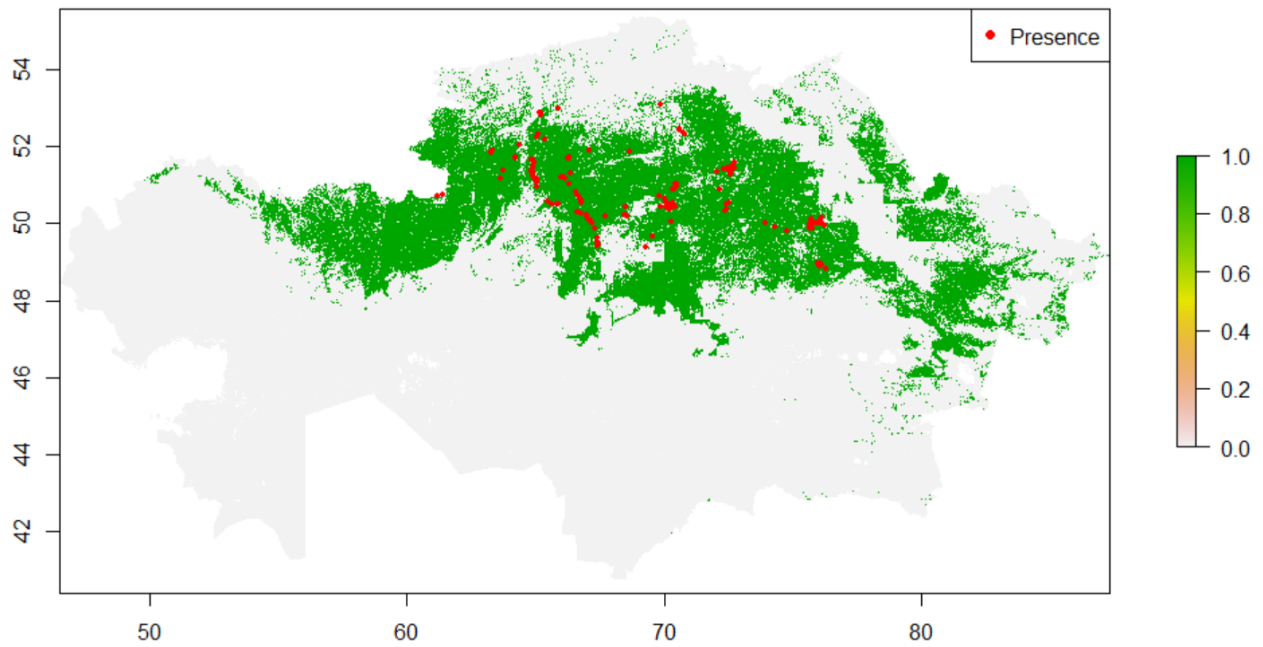


Figure 12. Binary map of the fifth model

### Validation map for MaxEnt models.

A total of 8,741 grid cells representing burrows of the *Marmota bobak* were identified through manual processing of satellite images. The presence of the species is marked by green points, indicating its occurrence in the area (Fig. 13). A subset of these points was verified through ground-truthing during fieldwork.

During the manual map processing, we marked 1,974 red points as absence locations. Within the presence areas, absence cells were placed densely between presence zones to track the boundaries of settlements within the presence territory. Beyond the limits of green points and historical boundaries, the red points become more dispersed, indicating the general verification of the territory (Fig. 14).

The satellite-based method, validated by the research of Koshkina et al. in the paper "Marmots from space: assessing population size and habitat use of a burrowing mammal using publicly available satellite images," is used as the basis for creating a map that is employed to validate MaxEnt models.

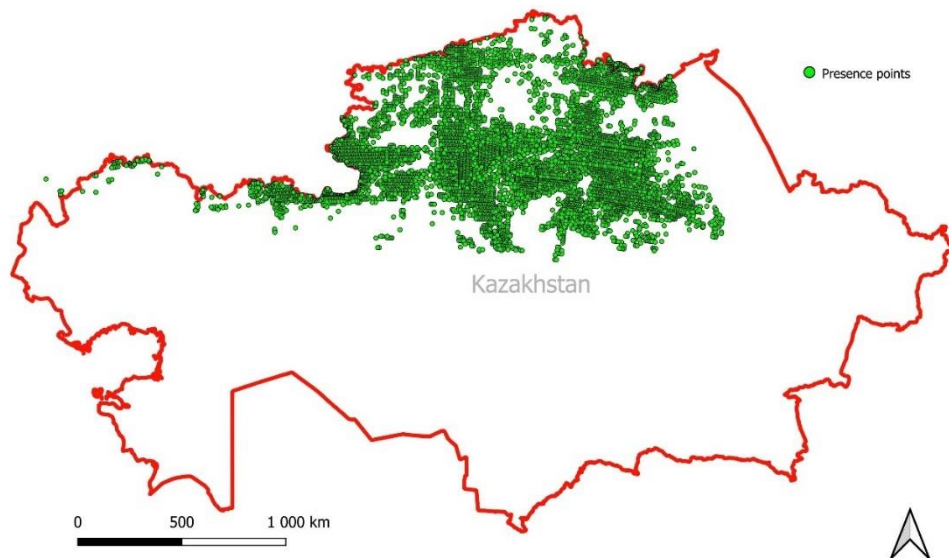


Figure 13. The final result of marmot range based on manual identification (Green – Marmot's present, red line – the Kazakhstan border).

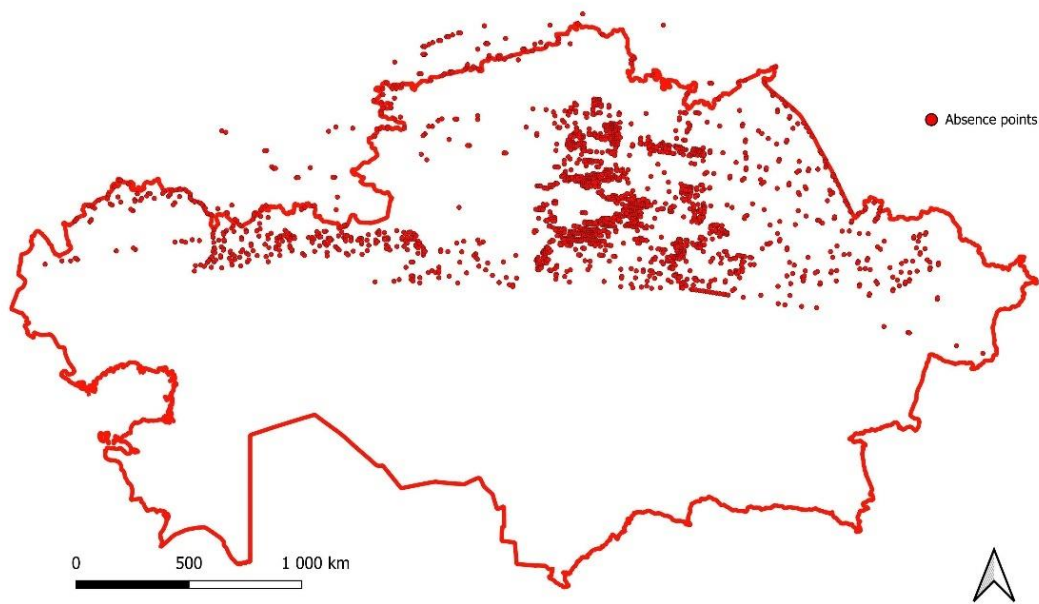


Figure 14. The result of marmot range based on manual identification (Red – Marmot’s absence, red line – the Kazakhstan border)

Comparing the modelled distribution map (red areas) with the manually mapped distribution (green areas) (Figure 15), we found that the majority of the manually mapped range overlaps with the predicted model, particularly in central Kazakhstan and slightly further north. However, there are northern areas where the model failed to predict the species' presence, despite confirmed occurrences in those regions. This discrepancy may be due to an insufficient number of presence points in the dataset, leading the model to overlook areas that differ slightly from those included in the training data. Additionally, the model predicted the species' presence further east, but this region is unsuitable for marmots due to changes in vegetation, the predominance of forest-steppe and mountainous landscapes and the presence of major rivers (e.g., the Irtysh), which make the species' occurrence in this area improbable.

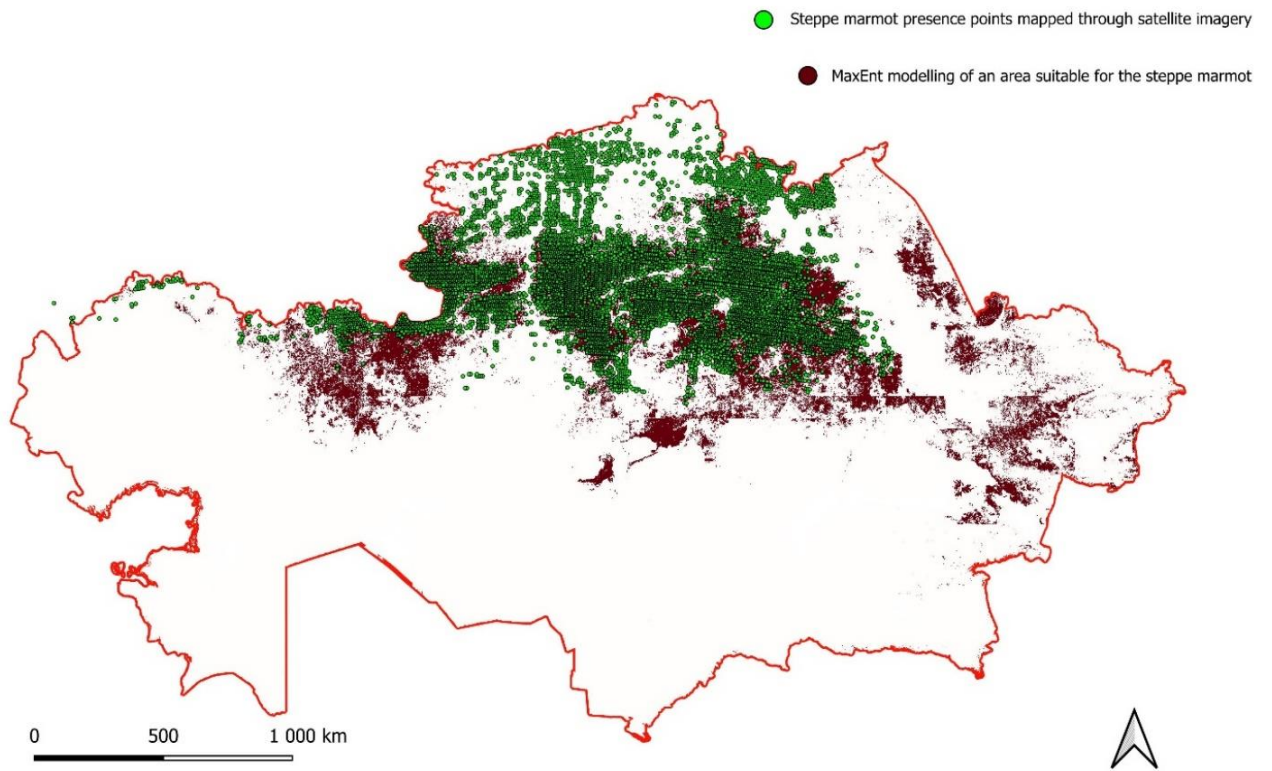


Figure 15. Comparison of satellite mapping and MaxEnt modelling maps

## Discussion

MaxEnt analysis was conducted in this study to develop a spatial habitat model for the *Marmota bobak* across Kazakhstan. The primary objective was to assess the model's effectiveness in distinguishing between presence and absence areas over large spatial scales and to identify the predictors contributing most significantly to the model. The MaxEnt model results were subsequently validated using field-mapped presence data, obtained through manual mapping based on freely available satellite imagery from Bing and Google. The comprehensive satellite-based species presence map was originally compiled in 2019–2020 and thoroughly revalidated prior to the current study.

Overall, MaxEnt demonstrated a high AUC for all five models, indicating consistently strong predictive performance. However, based on the thresholds of the binary maps, Model 3 appears to provide the most realistic prediction due to the strictest threshold for delineating suitable areas. The key variables contributing most significantly to the model are climatic factors (approximately 30%) and landscape features (around 10%), as expected. Additionally, the models reveal a substantial contribution from snow cover (9%), which is nearly equivalent to the contribution of soil variables.

In the first model, 19 variables were included, 18 of which are ecologically relevant for the *Marmota bobak*. The predictor *shrubs* represent only a part of land cover and does not exert a significant influence on the species. This is immediately reflected in the contribution of the data to the model, which excluded *shrubs*, *SAVI 2024*, and precipitation during the warmest quarter due to their zero contribution to the species distribution. The low contribution of these variables may be attributed either to the model deriving the necessary information from other predictors that already contain relevant data or to the uniformity or minimal variation of these variables across the study area, limiting the model's ability to detect patterns.

The most significant predictors are mean annual temperature, croplands, elevation range, soil types, and snow cover, while the remaining predictors exhibited moderate to minimal influence on the first model. Nevertheless, the model demonstrates high predictive performance with an AUC of 0.932. The map of suitable habitats accurately includes areas in Central and partially Northern Kazakhstan, but it also extends to Eastern

and slightly Western Kazakhstan, exceeding the known boundaries of the *Marmota bobak*'s habitat.

The second model, with 13 filtered predictors, also demonstrates a high AUC of 0.925. The dominant variables remain the same, maintaining the highest contribution to model construction. The maximum temperature of the warmest months shows a zero contribution, likely due to the model incorporating information from the mean annual temperature or assessing a low contribution of this variable. A reduction in the number of predictors increases the binary map threshold to 0.389, indicating that the model applies stricter criteria for site selection. This is reflected in the suitability map, where suitable areas become more fragmented, better corresponding to the actual distribution pattern of the marmot.

For the third model, the analysis focuses on the 10 most important predictors, with no zero contribution values observed. Each variable contributes to the model, while mean annual temperature, cropland, elevation, soil, and snow cover maintain consistently high contributions. In contrast, the influence of human settlements and NDVI for 2023–2024 decreases. The model performance remains robust with an AUC of 0.925; however, the binary threshold increases to 0.401, enhancing the model's conservativeness and reducing the likelihood of selecting areas with low habitat suitability for the marmot.

In the fourth analysis with nine variables, the contribution of all predictors increases, indicating greater reliance of the model on these factors. However, the AUC tends to decrease to 0.920, along with the binary threshold dropping to 0.328, making the model more liberal in selecting suitable areas.

In the fifth model, an additional predictor with the lowest contribution is excluded, resulting in increased model "softness" according to the binary threshold of 0.265. The AUC decreases to 0.916, while still remaining within the range of high-quality performance. The marmot distribution map retains the predicted range boundaries but becomes significantly denser compared to the previous models due to an increased number of areas classified as suitable by the model.

Overall, all models exhibit a similar general trend, maintaining high performance metrics despite a gradual decrease in AUC values and binary thresholds. The response

curves display relatively consistent patterns, with sharp fluctuations observed only when the number of variables is reduced in the later models. These models also show increased deviations, leading to a higher likelihood of errors and reduced accuracy.

All models indicate that the distribution of the *Marmota bobak* is strongly influenced by climatic factors. According to the response curves, the species is sensitive to deviations in temperature beyond its optimal range. In terms of anthropogenic impact, the *Marmota bobak* can inhabit areas along the edges of agricultural lands; however, the plateau observed in the response curves suggests that intensified cultivation may create unfavorable conditions, prompting the species to abandon such areas.

Elevation appears to be a significant factor, as the *Marmota bobak* tends to avoid lowlands and high elevations, likely due to climatic conditions and food availability, showing a preference for moderate altitudes. Soil characteristics are also crucial, affecting the marmot's distribution due to their suitability for burrowing and the availability of specific vegetation types. Regarding snow cover, the response curve indicates that the *Marmota bobak* avoids areas with deep snow, possibly due to the challenges associated with surface access after hibernation.

However, the third model appears to be the most balanced among all, demonstrating the best trade-off between accuracy and realism in marmot distribution modelling. It also exhibits a relatively high degree of fragmentation, highlighting the "core" area of the range more distinctly. In contrast, the fifth model provides clearer boundary delineation but tends to smooth the outputs, resulting in a more extensive filling of areas within the predicted range. Additionally, a reduction in the number of predictors simplifies the model while increasing the risk of overestimating the influence of the remaining factors on the species. Conversely, a large number of variables introduces "noise," leading to model overfitting.

However, even the best-performing third MaxEnt model does not fully correspond to the current distribution of the *Marmota bobak* in Kazakhstan. A comparison of the modelling maps with the manually created map based on satellite imagery reveals noticeable discrepancies between the model and the satellite-derived map.

When comparing the remote sensing map and the MaxEnt model, it is important to note that the model effectively identified the "core" of the range and the "gaps" within the presence area. However, it failed to detect areas in Northern Kazakhstan (closer to the Russian border), overestimated arid regions (Western Kazakhstan), and exhibited an eastward shift, despite these areas not being confirmed as habitats of the *Marmota bobak* by either the satellite-derived map or field surveys.

Modelling limitations and errors are likely associated with the number of predictors included in the training process of the initial model, potentially leading to overfitting and selection errors in subsequent analyses (Merow et al., 2013; Phillips, 2021). Additionally, the presence of highly correlated predictors may have introduced inaccuracies, causing the model to overestimate their contribution. The analysis also involved an insufficient number of local predictors, which could have affected model performance (Merow et al., 2013; Ahmad et al., 2023). Errors may also stem from model parameter settings; therefore, it is necessary to re-evaluate the MaxEnt regularization multiplier, which in this case is approximately 0.37. Although this value falls within acceptable limits, it remains sufficient to influence model sensitivity (Merow et al., 2013).

A similar study was conducted by Koshkina A. et al. (2019), where the authors explored the potential of satellite imagery for studying the *Marmota bobak* and incorporated an additional component in the form of MaxEnt analysis. Both studies exhibit similarities and differences. The extent of suitable habitat modelling for the marmot in the present study is higher compared to that of Koshkina A. et al. This discrepancy is likely due to differences in the number of variables used and the fact that Koshkina et al. immediately included 11 ecologically justified predictors, whereas in the present study, a set of the most influential variables was filtered from an initial pool of 19 variables. Additional differences are observed in cross-validation approaches and bias correction methods.

Overall, both studies demonstrate that climatic and landscape factors play a key role in marmot habitat modelling, specifically temperature, elevation, soil types, snow cover, and NDVI in the present study, as well as temperature, precipitation, NDVI, soil types, and distance to water bodies in the referenced article.

To improve future results, the following activities can be conducted:

1. Conducting field surveys in regions with insufficient species occurrence data. These data will reduce spatial bias in modelling.
2. Incorporation of absence points into the model. Although MaxEnt operates with presence-only data, the inclusion of confirmed absence points can reduce false positive predictions (Phillips, 2021).
3. Include additional justified predictors such as grazing areas, hydrological data, fallow lands, and cultivated fields.
4. Conduct a comparison with other modelling programs and packages (table 5).

<b>Method / Package</b>	<b>Description</b>	<b>Advantages in this analysis</b>
MaxEnt (dismo, terra, raster)	Standard Method for Modelling Species Using Maximum Entropy	Support in R, manual control of background points, high interpretability
ENMeval (R)	Optimization of MaxEnt parameters	Automates parameter selection, reduces overfitting, improves model accuracy
SDMtoolbox (R, ArcGIS)	Tools for working with ecological models	Convenient functions for filtering points, cross-validation, and predictor correlation assessment
MaxEnt GUI (Java)	Official program for MaxEnt with graphical interface	Easy to use, no programming required
MaxLike (R)	Alternative to MaxEnt based on maximum likelihood method	Takes into account data errors, does not require background points
Wallace (R, Shiny)	Interactive platform for species modelling	MaxEnt support, user-friendly interface, automation of some processes
Random Forest (RF, R, Python)	Machine learning for species modelling	High accuracy, takes into account nonlinear dependencies

Generalized Additive Models (GAM, R, Python)	Flexible models with nonlinear effects	Allows to take into account complex ecological dependencies
--	--	---

Table 5. Alternative methods and packages for spatial modelling of species

As a result, the model quality will increase, and predictive performance will improve. With the growing use of modelling software in conservation, the results of MaxEnt analysis can make a significant contribution to the identification of key habitats for species, the monitoring of populations across large areas, and serve as an additional tool in the study of burrowing rodent distributions. Moreover, these results may inform adjustments in rangeland management practices, taking into account the needs of wildlife dependent on these ecosystems.

## Conclusion

This study assessed the application of the MaxEnt software for modelling the distribution range of the *Marmota bobak* in Kazakhstan, identified the variables contributing most significantly to the species' distribution according to the analysis, and subsequently validated the model using remote sensing data.

The results indicate that the accuracy of the MaxEnt model is approximately 60-70%, with an overall model error of about 30-40%. The omission error accounts for approximately 10-20%, while the commission error ranges from 15-25%. All five models effectively predict the core range, while the range boundaries and the density of predicted areas respond to variations in the set of predictors. The use of non-local variables (climatic, topographic, and soil-related) leads to an expansion of the predicted areas beyond the limits of confirmed occurrence data. The inclusion of local predictors (land use, vegetation data) produces more detailed maps of suitable habitats, with a greater focus around the known distribution range of the species. Thus, achieving optimal results requires a combination of local and non-local variables to better understand the influence of specific factors on the distribution of the *Marmota bobak* across large territories such as Kazakhstan. The most significant factors identified include mean annual temperature, soil type, snow cover, elevation, arable land, NDVI, and grassland areas, which are ecologically justified. However, further refinement of the MaxEnt model is necessary, as it tends to predict a larger extent of suitable habitats than observed in reality. Comparisons with remote sensing maps reveal that the model overestimates the range, extending it towards high-altitude regions in the east and more arid areas in the west, where no *Marmota bobak* occurrences have been recorded. And it omitted a large part of the northern range where the species is present.

The obtained results contribute to the advancement of spatial modelling of burrowing rodents across large territories and demonstrate the potential of MaxEnt and satellite imagery in situations where field surveys are not feasible.

Thus, the study confirms the relevance of modelling large rodents at broad spatial scales, highlights the importance of using a combination of local and non-local predictors in the analysis, and emphasizes the significance of an integrated approach in enhancing the effectiveness of the results.

## **Acknowledgments**

I would like to express my sincere gratitude to my supervisors, Alyona Koshkina and Sebastiaan van der Linden, for their invaluable guidance, support, and constructive feedback in this research.

I would also like to thank Maksim Shashkov for his consultations on MaxEnt and Aidar Aitkulov for his assistance in collecting maps for the analysis.

Finally, I would like to express my gratitude to my family, friends, and colleagues for their moral support and patience throughout this journey.

## References

1. Abaturov B.D. Mammals as a component of ecosystems (on the example of herbivorous mammals in the semi-desert). Moscow: Nauka, 1984, 286 p. (in Russian)
2. Branch, L.C., Hierro, J.L., Villarreal, D., 1999. Patterns of plant species diversity following local extinction of the plains vizcacha in semi-arid scrub. *Journal of Arid Environments* 41, 173–182.
3. Ceballos, G., Pacheco, J., List, R., 1999. Influence of prairie dogs (*Cynomys ludovicianus*) on habitat heterogeneity and mammalian diversity in Mexico. *Journal of Arid Environments* 41, 161–172.
4. Davidson, A.D., Lightfoot, D.C., 2006. Keystone rodent interactions: prairie dogs and kangaroo rats structure the biotic composition of a desertified grassland. *Ecography* 29, 755–756.
5. Davidson, A.D., Lightfoot, D.C., 2007. Interactive effects of keystone rodents on the structure of desert grassland arthropod communities. *Ecography* 30, 515–525.
6. Davidson, A.D., Lightfoot, D.C., 2008. Burrowing rodents increase landscape heterogeneity in a Desert Grassland. *Journal of Arid Environments* 72, 1133–1145.
7. Zoltán Rádai, Tatyana M. Bragina, Yevgeny A. Bragin, Balázs Deák. 2020. steppe marmota (*Marmota bobak*) as ecosystem engineer in arid steppes.
8. Kirikov S.V. "Person and nature of the steppe zone." Science Publishing House 1983. 84-87 pp (in Russian)
9. Bibikov D. I. / Marmots / D. I. Bibikov. - Moscow: Agropromizdat, 1989. page 254, (in Russian)
10. Formozov A. N. Theriology: Textbook for state universities / Edited by A. N. Formozov. - Moscow: Higher School, 1963. - 396 c. (in Russian)
11. Zimina V.N., Isakov Yu.A. (Ed.). (1980) /Marmots. Biocenotic and practical significance. Moscow: Nauka, (in Russian)
12. Koshkina, A., Grigoryeva, I., Tokarsky, V., Urazaliyev, R., Kuemmerle, T., Hölzel, N.& Kamp, J. 2020. Marmots from space: assessing population size and habitat use of a burrowing mammal using publicly available sat-ellite images. *Remote Sensing in Ecology and Conservation*, 6 (2), 153–167. <https://doi.org/10.1002/rse2.138>
13. R.K. Bangert, C.N. Slobodchikoff. Conservation of prairie dog ecosystem engineering may support arthropod beta and gamma diversity. *Journal of Arid Environments* 67 (2006) 100–115
14. Leyequien, E., Verrelst, J., Slot, M., Schaepman-Strub, G., Heitkonig, " I.M.A., Skidmore, A., 2007. Capturing the fugitive: applying remote sensing to terrestrial animal distribution and diversity. *Int. J. Appl. Earth Obs. Geoinf.* 9, 1–20. <https://doi.org/10.1016/j.jag.2006.08.002>.
15. Hollings, T., Burgman, M., van Andel, M., Gilbert, M., Robinson, T., Robinson, A., 2018. How do you find the green sheep? A critical review of the use of remotely sensed imagery to detect and count animals. *Methods Ecol. Evol.* 9, 881–892. <https://doi.org/10.1111/2041-210X.12973>.
16. Skidmore, A.K., Coops, N.C., Neinavaz, E., Ali, A., Schaepman, M.E., Paganini, M., Kissling, W.D., Vihervaara, P., Darvishzadeh, R., Feilhauer, H., Fernandez, M.,

- Fernandez, ´ N., Gorelick, N., Geijzendorffer, I., Heiden, U., Heurich, M., Hobern, D., Holzwarth, S., Muller-Karger, F.E., Van De Kerchove, R., Lausch, A., Leitao, P.J., Lock, M.C., Múcher, C.A., O'Connor, B., Rocchini, D., Roeoesli, C., Turner, W., Vis, J. K., Wang, T., Wegmann, M., Wingate, V., 2021. Priority list of biodiversity metrics to observe from space. *Nat. Ecol. Evol.* 5, 896–906. <https://doi.org/10.1038/s41559-021-01451-x>.
17. Fleming PA, Anderson H, Prendergast AS, Bretz MR, Valentine LE, Hardy GES (2014) Is the loss of Australian digging mammals contributing to a deterioration in ecosystem function? *Mammal Review* 44: 94–108
  18. Gabrielle BECA, Leonie E. VALENTINE, Mauro GALETTI, Richard J. HOBBS (2021) Ecosystem roles and conservation status of bioturbator mammals. *Mammal Review* ISSN 0305-1838: 1-16
  19. Valiente-Banuet A, Aizen MA, Alcántara JM, Arroyo J, Cocucci A, Galetti M et al. (2015) Beyond species loss: the extinction of ecological interactions in a changing world. *Functional Ecology* 29: 299–307.
  20. Miranda V, Rothen C, Yela N, Aranda-Rickert A, Barros J, Calcagno J, Fracchia S (2019) Subterranean desert rodents (genus *Ctenomys*) create soil patches enriched in root endophytic fungal propagules. *Microbial Ecology* 77: 451–459.
  21. Villarreal D, Clark KL, Branch LC, et al. 2008. Alteration of ecosystem structure by a burrowing herbivore, the plains vizcacha (*Lagostomus maximus*). *J Mammal* 89: 700–11.
  22. Valentine LE, Bretz M, Ruthrof KX, Fisher R, Hardy GESJ, Fleming PA (2017) Scratching beneath the surface: bandicoot bioturbation contributes to ecosystem processes. *Austral Ecology* 42: 265–276
  23. Eldridge DJ, Koen TB (2021) Temporal changes in soil function in a wooded dryland following simulated disturbance by a vertebrate engineer. *Catena* 200: 105166.
  24. White RP, Murray S, Rohweder M, Prince S, Thompson K (2000) *Grassland Ecosystems*. World Resources Institute Washington, DC, USA
  25. Martínez-Estévez L, Balvanera P, Pacheco J, Ceballos G (2013) Prairie dog decline reduces the supply of ecosystem services and leads to desertification of semiarid grasslands. *PLoS One* 8
  26. Yoshihara Y, Ohkuro T, Buuveibaatar B, et al. 2009. Spatial pattern of grazing affects influence of herbivores on spatial heterogeneity of plants and soils. *Oecologia* 162: 427–34
  27. Hogan BW. 2010. The plateau pika: a keystone engineer on the Tibetan Plateau (doctoral thesis). Tempe, AZ: Arizona State University
  28. Dudnikov A. A., Kurochkin A. S., Fokina M. E., Sharonova I. V./2021/ Current state of settlements of the Marmota bobak (*Marmota bobak* Müll.) in the conditions of the mixed-grass-tyrchak-sooty steppe of the Pestravsky district of Samara region. (in Russian)
  29. Zimina RP. 1978. Marmots: distribution and ecology. Moscow, Russia: Nauka Publishing House (in Russian).
  30. Kolesnikov VV. 2011. Ecology and current state of the population of the Marmota bobak (*Marmota bobak* Müller, 1776) in the Middle Volga region (candidate thesis). Kirov, Russia (in Russian).

31. Zarubin BE. 1997. Resources and management of populations of the steppe (Marmota bobak), gray (M. baibacina), and Mongolian (M. sibirica) marmots in Russia (candidate thesis). Moscow, Russia (in Russian).
32. Shubin IG, Abelentsev VI, Semikhatova SN. 1978. Marmots: distribution and ecology. Moscow, Russia: Nauka Publishing House (in Russian).
33. Rumyantsev VY. 1991. The Marmota bobak on arable lands of Kazakhstan. Vestnik Moskovskogo Universiteta. Seriya 16. Biologiya. 96(4):15–28 (in Russian).
34. Anke Hoffmann, Jan Decher, Francesco Rovero, Juliane Schaer Field Methods and Techniques for Monitoring Mammals. 2010. Chapter 19, 482-529 In book: Manual on field recording techniques and protocols for All Taxa Biodiversity Inventories and Monitoring.
35. Daniel J. Leedy, Biologist, U. S. Fish and Wildlife Service/ AERIAL PHOTO USE AND INTERPRETATION IN THE FIELDS OF WILDLIFE AND RECREATION/ International Photogrammetry Congress, Commission VII/ 1953
36. Buechner, H.K., Craighead Jr., F.C., Craighead, J.J., Cote, C.E., 1971. Satellites for research on free-roaming animals. Bioscience 21, 1201–1205
37. Fancy, S.G., Pank, L.F., Douglas, D.C., Curby, C.H., Garner, G.W., 1988. Satellite Telemetry: A New Tool for Wildlife Research and Management. Fish and Wildlife Service, Washington DC
38. Saxon, E., 1983. Mapping the habitats of rare animals in the Tanami wildlife sanctuary (Central Australia): an application of satellite imagery. Biol. Conserv. 27, 243–257
39. Lciffler E. and Margules C., Wombats Detected from Space REMOTE SENSING OF ENVIRONMENT 9:47-56 (1980) 47 Division of Land Use Re.search, CSIRO, Ganb6nn~ A.C.T., Australia
40. Dalsted, K. J., S. Sather-Blair, B. K. Worcester, and R. Klukas. 1981. Application of remote sensing to prairie dog management. J. Range Manag. 34, 218–223.
41. Olofsson, J., Tommervik, H. & Callaghan, T. V. 2012. Vole and lemming activity observed from space. *Nat. Clim. Change* 2, 880. <https://doi.org/10.1038/nclimate1537>.
42. Tape, K. D., Jones, B. M., Arp, C. D., Nitze, I. & Grosse, G. 2018 Tundra be dammed: Beaver colonization of the Arctic. *Glob. Change Biol.* 24, 4478 4488. <https://doi.org/10.1111/gcb.14332>.
43. Swinbourne, M. J., Taggart, D. A., Swinbourne, A. M., Lewis, M. & Ostendorf, B. 2018. Using satellite imagery to assess the distribution and abundance of southern hairy-nosed wombats (*Lasiorhinus latifrons*). *Remote Sens. Environ.* 211, 196-203. <https://doi.org/10.1016/j.rse.2018.04.017>
44. Sidle, J. G., D. H. Johnson, B. R. Euliss, and M. Tooze. 2002. Monitoring black-tailed prairie dog colonies with high-resolution satellite imagery. *Wildl. Soc. Bull.* 30, 405–411
45. Bibikov DI, Chekalin VB. 1959. Experience in applying the mapping method to study certain features of gray marmots. In: Geography of the distribution of terrestrial animals and methods of its study. Moscow, Russia: USSR Academy of Sciences Publishing House; p. 95–107 (in Russian).
46. Vinogradov BV, Leontyeva EV. 1957. The use of aerial methods in the study of vegetation in Northern Kazakhstan. In: Materials on the use of aerial methods in

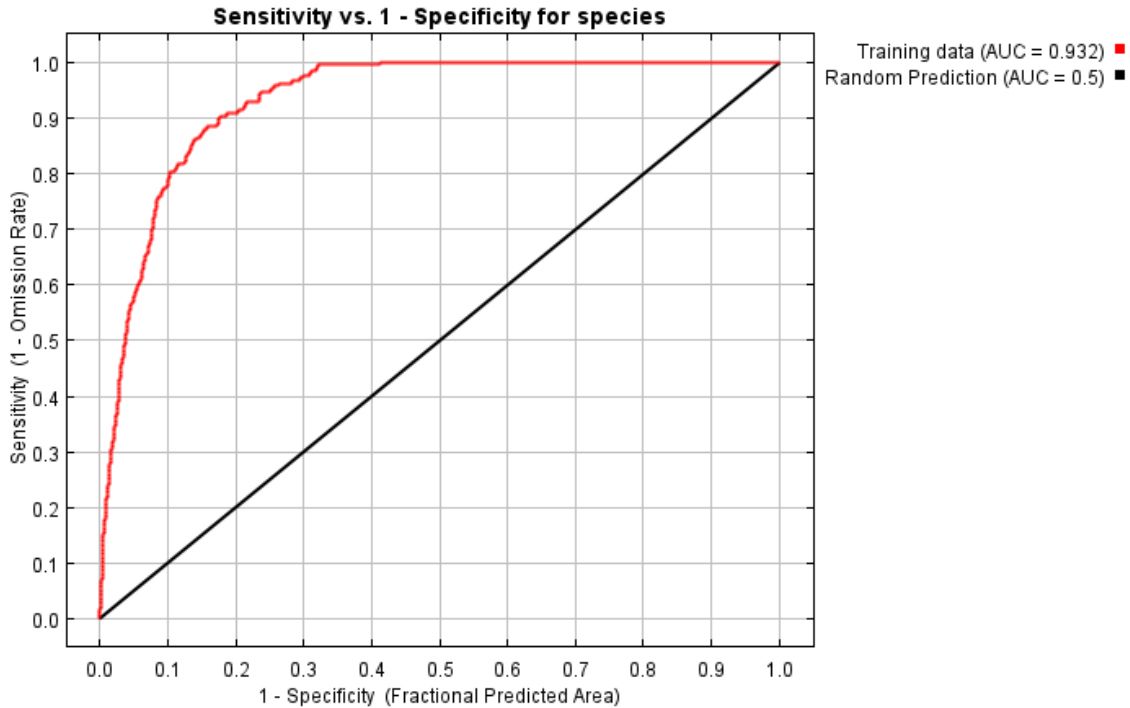
- the study of soils and vegetation of Northern Kazakhstan. Moscow–Leningrad: Publishing House of the USSR Academy of Sciences; p. 47–75 (in Russian).
47. Rumyantsev VY. 1993. Application of aerial photographs for mapping the distribution of the Marmota bobak (*Marmota bobak*). *Zoologichesky Zhurnal*. 72(9):137–148 (in Russian).
  48. Dubinin MY, Kostikova AA. 2008. Introduction to geographic information systems: vector and raster data. Available from: <http://gis-lab.info/docs/giscourse/11-vector-raster.html> (in Russian).
  49. J. Elith and J. Franklin, *Species Distribution Modelling*, In Reference Module in Life Sciences, Elsevier, 2017, ISBN: 978-0-12-809633-8, <http://dx.doi.org/10.1016/B978-0-12-809633-8.02390-6>
  50. Srivastava, V., Lafond, V. and Griess, V. C. 2019. Species distribution models (SDM): applications, benefits and challenges in invasive species management. – *CABI Reviews*, pp. 1–13, [10.1079/PAVSNNR201914020](https://doi.org/10.1079/PAVSNNR201914020)
  51. Niklaus E. Zimmermann, Thomas C. Edwards Jr, Catherine H. Graham, Peter B. Pearman, Jens-Christian Svenning. 2010. New trends in species distribution modelling. *Ecography* 33: 985–989 <https://doi.org/10.1111/j.1600-0587.2010.06953.x>
  52. Guillera-Arroita G, Lahoz-Monfort JJ, Elith J, Gordon A, Kujala H, Lentini PE et al (2015) Is my species distribution model fit for purpose? Matching data and models to applications. *Glob Ecol Biogeogr* 24(3):276–292. <https://doi.org/10.1111/geb.12268>
  53. Ehrlén, J. and Morris, W. F. 2015. Predicting changes in the distribution and abundance of species under environmental change. – *Ecol. Lett.* 18: 303–314.
  54. Zurell D, Franklin J, König C, Bouchet PJ, Dormann CF, Elith J et al (2020) A standard protocol for reporting species distribution models. *Dent Echo* 43(9):1261–1277. <https://doi.org/10.1111/ecog.04960>
  55. Ferrier, S. et al. 2016. IPBES – the methodological assessment report on scenarios and models of biodiversity and ecosystem services. – Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, Bonn, Germany
  56. Franklin, J. et al. 2017. Big data for forecasting the impacts of global change on plant communities. – *Global Ecol. Biogeogr.* 26: 6–17.
  57. Wüest, R. O. et al. 2020. Macroecology in the age of big data – where to go from here? – *J. Biogeogr.* 47: 1–12
  58. Golding, N. et al. 2018. The zoon R package for reproducible and shareable species distribution modelling. – *Methods Ecol. Evol.* 9: 260–268.
  59. Merow, C. et al. 2013. A practical guide to maxent for modelling species' distributions: what it does, and why inputs and settings matter. – *Ecography* 36: 1058–1069
  60. Miller J (2010) Species distribution modelling. *Geography. Compass* 4(6):490–509. <https://doi.org/10.1111/j.1749-8198.2010.00351.x>
  61. Lissovsky, A.A. and Dudov, S.V., Advantages and limitations of application of the species distribution modelling methods. 2. *MaxEnt, Zh. Obsch. Biol.*, 2020, vol. 81, no. 2, pp. 135–146

62. Thekke Thumbath Shameer, Raveendranathanpillai Sanil (2023) Machine Learning-Based Predictive Modelling Approaches for Effective Understanding of Evolutionary History, Distribution, and Niche Occupancy: Western Ghats as a Model, DOI: 10.1007/978-981-99-0131-9\_3 In book: Ecosystem and species habitat modelling for conservation and restoration.
63. Mitchell TM (2006) The discipline of machine learning, vol 9. Carnegie Mellon University, School of Computer Science, Machine Learning Department, Pittsburgh
64. Dhyani S, Kadaverugu R, Pujari P (2020) Predicting impacts of climate variability on Banj oak (*Quercus leucotrichophora* A. Camus) forests: understanding future implications for Central Himalayas. *Reg Environ Chang* 20:113. <https://doi.org/10.1007/s10113-020-01696-5>
65. Shalini Dhyani, Dibyendu Adhikari, Rajarshi Dasgupta, Rakesh Kadaverugu., 2023., Ecosystem and Species Habitat Modelling for Conservation and Restoration, [https://doi.org/10.1007/978-981-99-0131-9\\_7](https://doi.org/10.1007/978-981-99-0131-9_7). Pp 121-141.
66. VanDerWal, J., Shoo, L.P., Graham, C., Williams, S.E., 2009. Selecting pseudo-absence data for presence-only distribution modelling: How far should you stray from what you know? *Ecol. Modell.* 220, 589–594.
67. Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. *Ecol. Modell.* 148, 1–13.
68. Longhui Lu <sup>1</sup>, Zhongxiang Sun <sup>2</sup>, Eerdeng Qimuge <sup>3</sup>, Huichun Ye <sup>1,4</sup>, Wenjiang Huang <sup>1,4,5</sup>, Chaojia Nie <sup>1</sup>, Kun Wang <sup>1</sup> and Yantao Zhou <sup>6</sup> 2022. Using Remote Sensing Data and Species–Environmental Matching Model to Predict the Potential Distribution of Grassland Rodents in the Northern China. <https://doi.org/10.3390/rs14092168>
69. Lydolph, P. (1965). *Geography of the USSR*. NY, USA: Wiley.
70. Kirsten M. de Beurs, Geoffrey M. Henebry, Land surface phenology, climatic variation, and institutional change: Analyzing agricultural land cover change in Kazakhstan, February 2004, *Remote Sensing of Environment* 89(4):497-509 DOI: 10.1016/j.rse.2003.11.006
71. Brinkert, A., N. Holzel, T. V. Sidorova, and J. Kamp. 2016. Spontaneous steppe restoration on abandoned cropland in Kazakhstan: grazing affects successional pathways. *Biodiversity and Conservation* 25(12), DOI: 10.1007/s10531-015-1020-7
72. Kapitonov VI. 1966. Distribution of marmots in Central Kazakhstan and prospects for their harvesting. *Proceedings of the Institute of Zoology of the Academy of Sciences of the Kazakh SSR*. 26:94–134 (in Russian).
73. Sludskii AA, et al. 1969. *Mammals of Kazakhstan: Rodents (marmots and ground squirrels)*. Vol. 1, Part 1. Alma-Ata, Kazakhstan: Nauka of the Kazakh SSR. 454 p. (in Russian).
74. Sludsky A.A. 1980a. Marmot hunting in Kazakhstan. In: *Marmots: biocenotic and practical significance*. Moscow, Russia: Nauka; p. 181–190. (in Russian).
75. Bibikov D.I. 1989. *Marmots*. Moscow, Russia: Agropromizdat. 250 p. (in Russian).
76. Zimina R.P, Bibikov DI. 1967. The state and main objectives of scientific research on the geography of marmot resources in the USSR. In: *Marmot fauna resources in the USSR*. Moscow, Russia: Science. pp. 3–5. (in Russian).

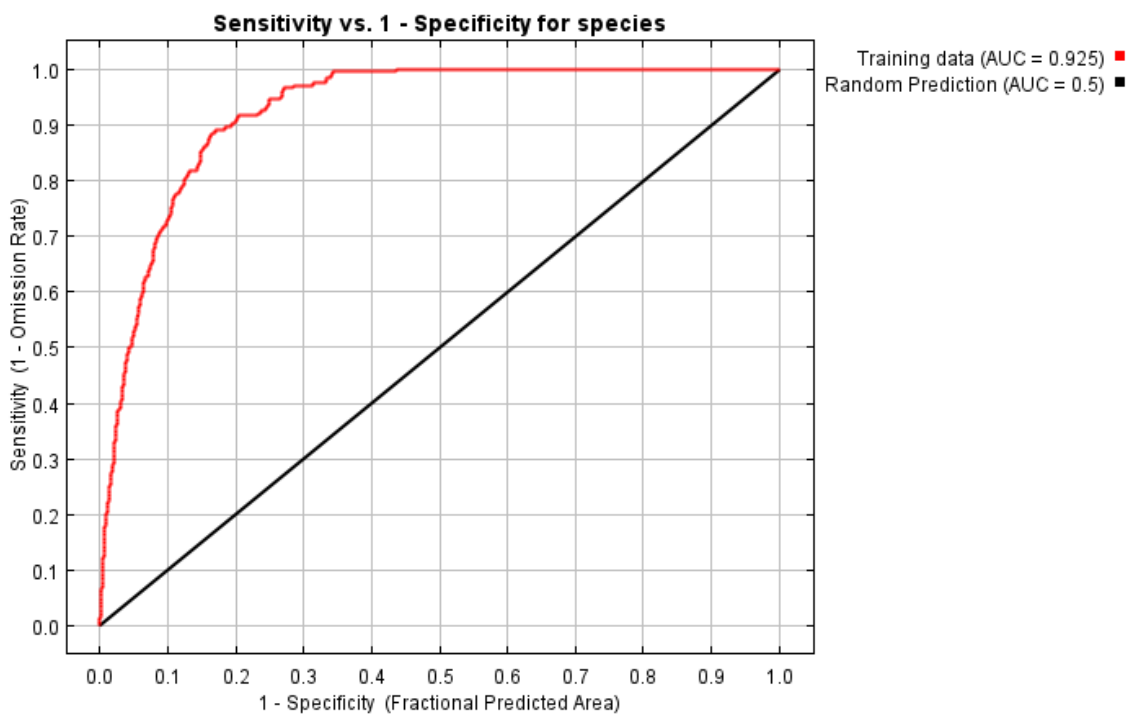
77. Zimina R.P. 1980. Game animals of the USSR and their habitats: Marmots. Biocenotic and practical significance. Moscow, Russia: Nauka Publishing House. 92 p. (in Russian).
78. Vinogradov B.V. 1985. Zoogenic spatial complexes in terrestrial ecosystems. A.N. Severtsov Institute of Evolutionary Morphology and Ecology of Animals, Academy of Sciences of the USSR. Moscow, Russia: Nauka. 5 p. (in Russian).
79. John Wiens, Diana Stralberg, Dennis Jongsomjit, Mark A Snyder. 2009. Niches, models, and climate change: Assessing the assumptions and uncertainties, Proceedings of the National Academy of Sciences 106, 19729-36. <http://dx.doi.org/10.1073/pnas.0901639106>
80. Natalia Telnova 2017. Revealing and mapping long-term NDVI trends for the analysis of climate change contribution to agroecosystems' productivity dynamics in the Northern Eurasian forest-steppe and steppe January 2017 Modern problems of remote sensing of the Earth from space 14(6):97-107, DOI:[10.21046/2070-7401-2017-14-6-97-107](https://doi.org/10.21046/2070-7401-2017-14-6-97-107) (in Russian)
81. Ya Liu, Yan Li, Shuangcheng Li and Safa Motesharrei, 2015. Spatial and Temporal Patterns of Global NDVI Trends: Correlations with Climate and Human Factors. Remote sensing ISSN 2072-4292 [www.mdpi.com/journal/remotesensing](http://www.mdpi.com/journal/remotesensing). DOI:10.3390/rs71013233
82. Steven J. Phillips, AT&T Research, 2021. A Brief Tutorial on Maxent. [https://biodiversityinformatics.amnh.org/open\\_source/maxent/Maxent\\_tutorial\\_2021.pdf](https://biodiversityinformatics.amnh.org/open_source/maxent/Maxent_tutorial_2021.pdf)
83. Robert J. Hijmans 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model/ Volume 93, Issue 3/March 2012/Pages 679-688/ <https://doi.org/10.1890/11-0826.1>
84. Melo, F. (2013). Area under the ROC Curve. In: Dubitzky, W., Wolkenhauer, O., Cho, KH., Yokota, H. (eds) Encyclopedia of Systems Biology. Springer, New York, NY. [https://doi.org/10.1007/978-1-4419-9863-7\\_209](https://doi.org/10.1007/978-1-4419-9863-7_209)
85. Cory Merow, Matthew J. Smith, John A. Silander Jr, 2013/A practical guide to MaxEnt for modelling species' distributions: what it does, and why inputs and settings matter/Volume 36, Issue 10/October 2013/Pages 1058-1069/ <https://doi.org/10.1111/j.1600-0587.2013.07872.x>
86. Mohsen Ahmad, Mahmoud-Reza Hemami, Mohammad Kaboli, Farzin Shabani, 2023/MaxEnt brings comparable results when the input data are being completed; Model parameterization of four species distribution models/Ecol Evol. 2023 Feb 17;13(2):e9827. doi: [10.1002/ece3.9827](https://doi.org/10.1002/ece3.9827)

# Appendices A (graphs)

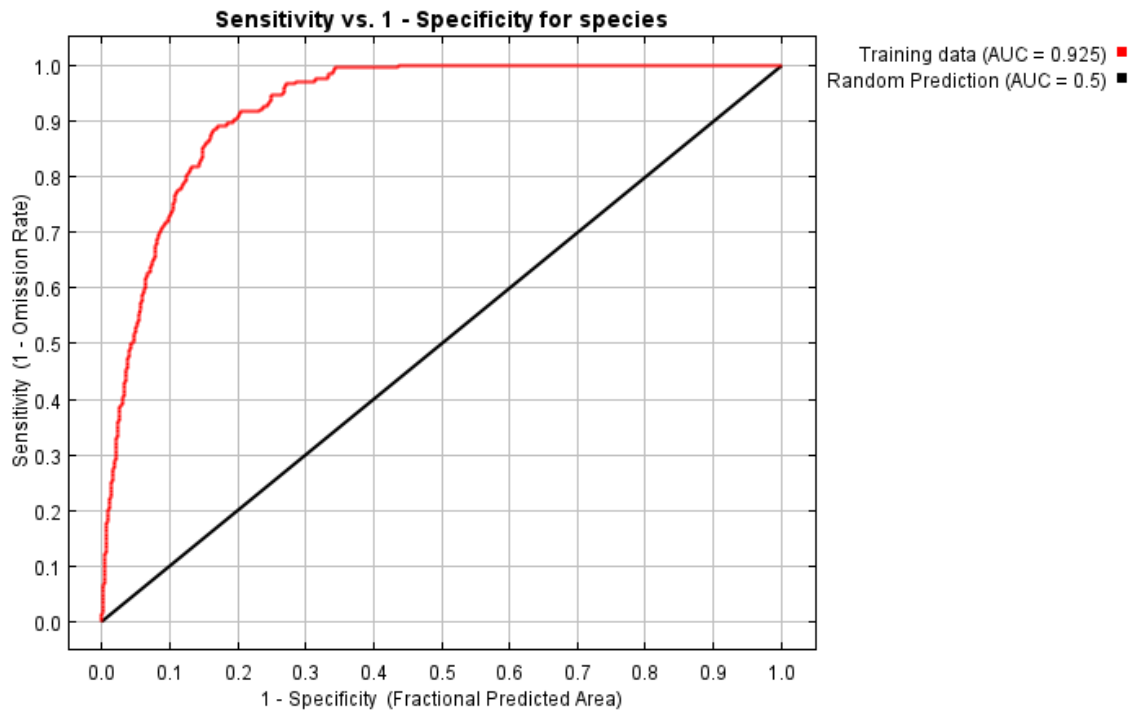
## MaxEnt – 1, ROC AUC



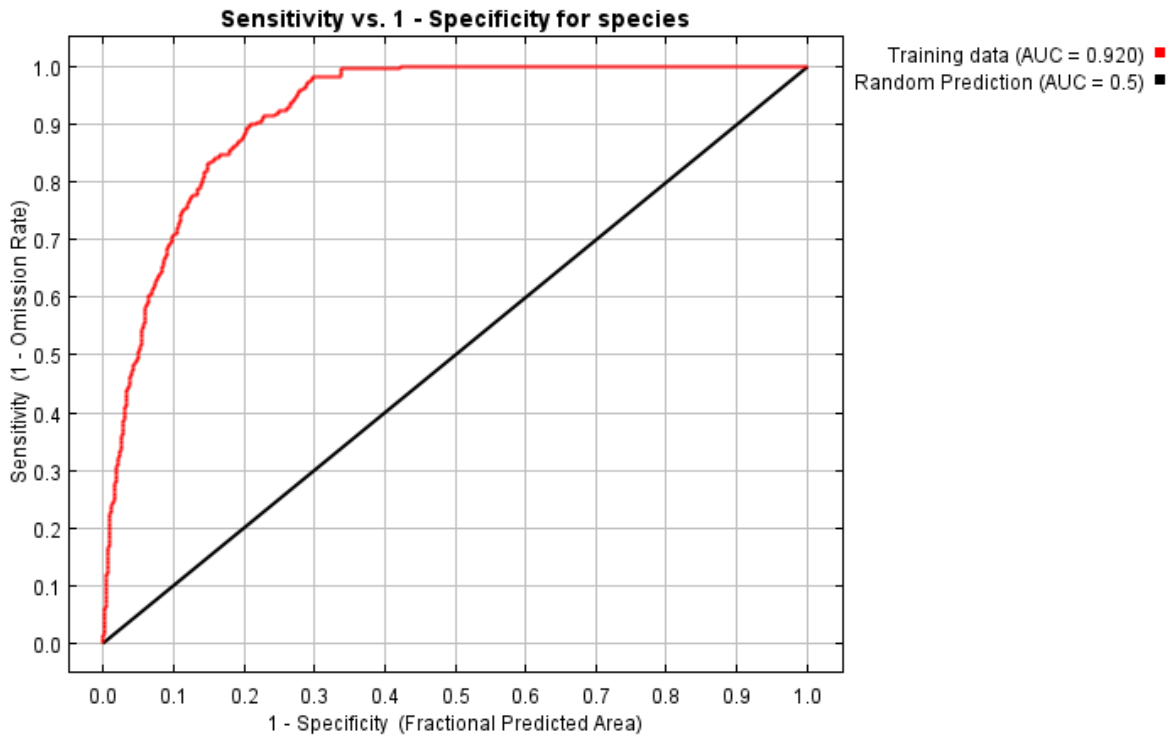
## MaxEnt – 2, ROC AUC



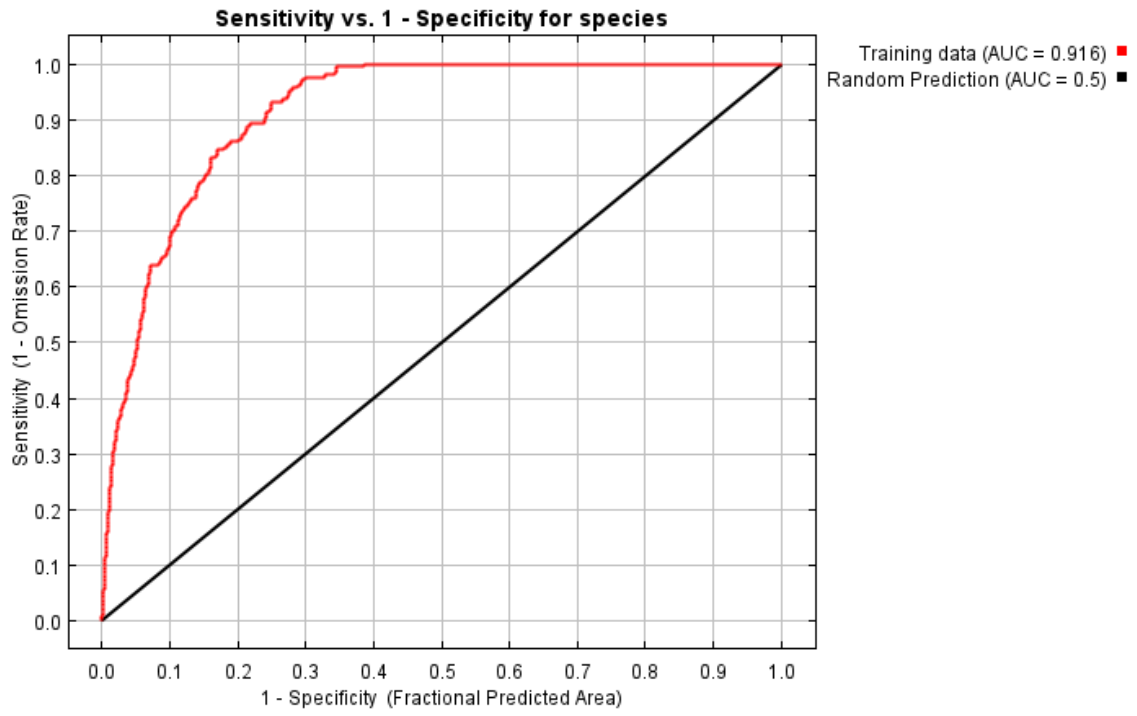
### MaxEnt – 3, ROC AUC



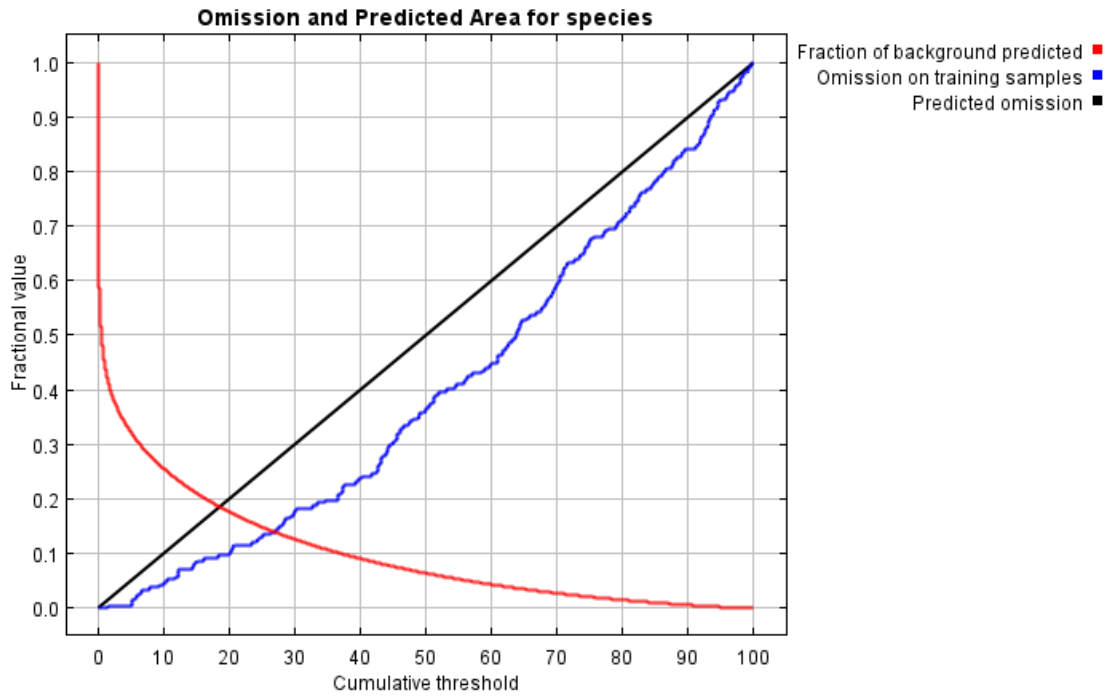
### MaxEnt – 4, ROC AUC



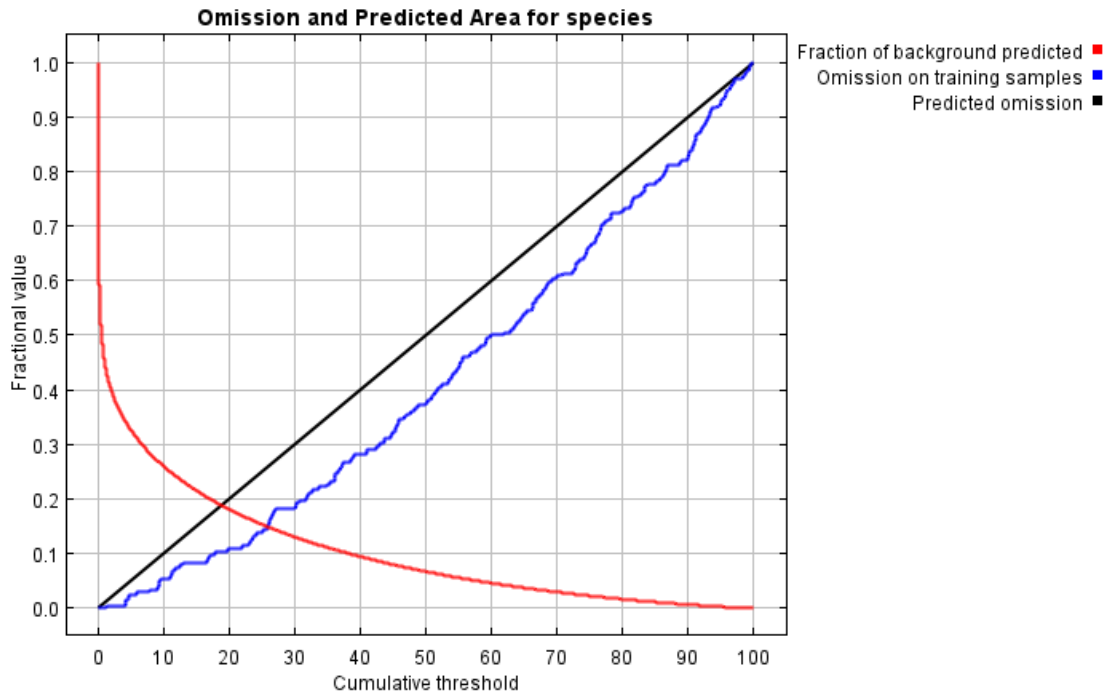
## MaxEnt – 5, ROC AUC



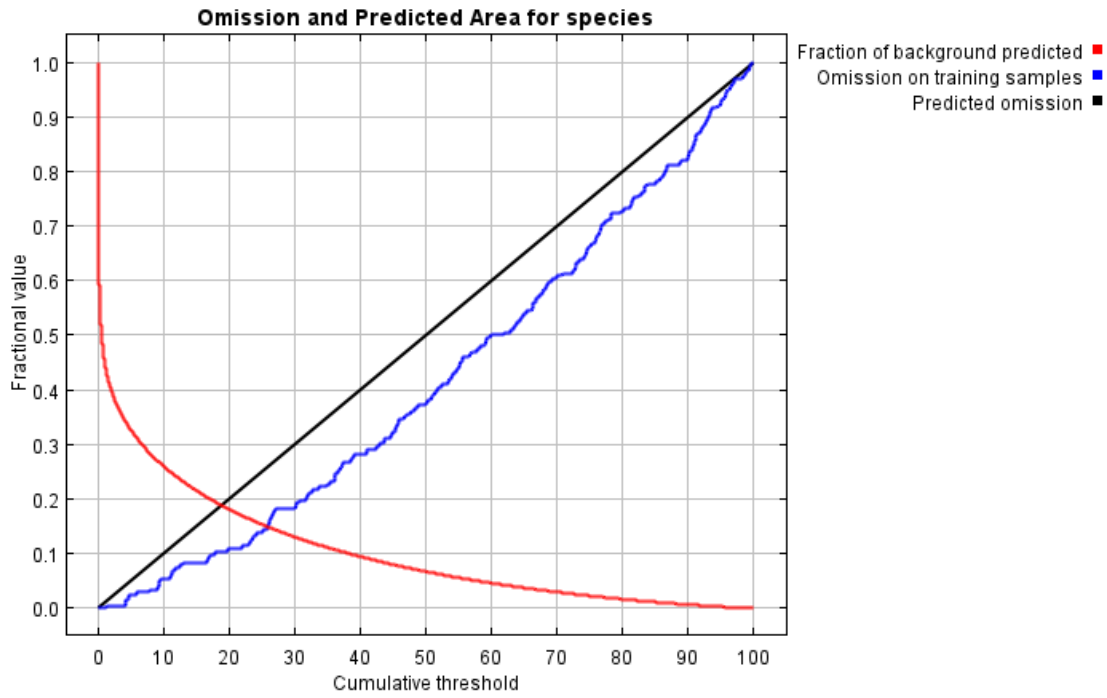
## Omission and Predicted Area graph/MaxEnt Model 1



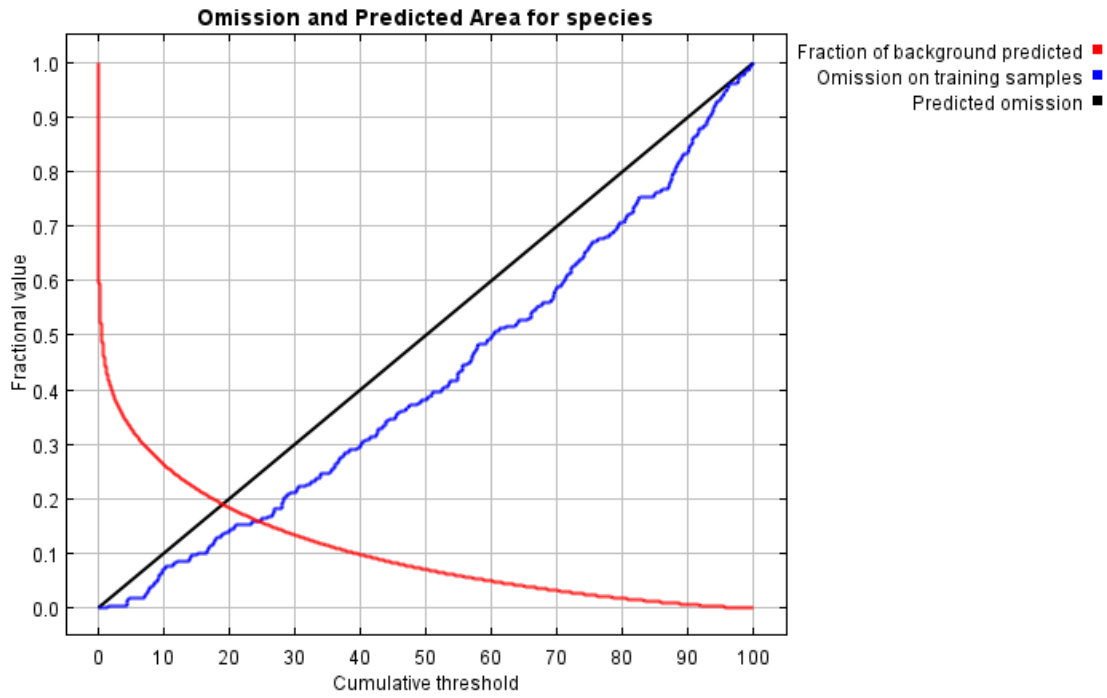
## Omission and Predicted Area graph/MaxEnt Model 2



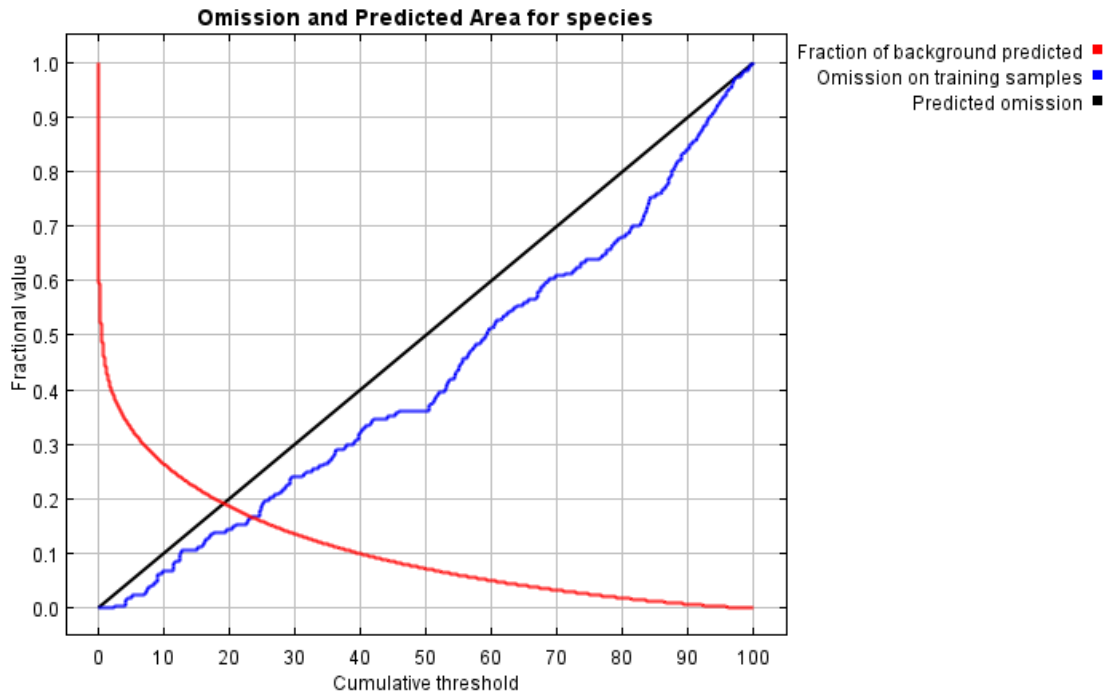
### Omission and Predicted Area graph/MaxEnt Model 3



### Omission and Predicted Area graph/MaxEnt Model 4

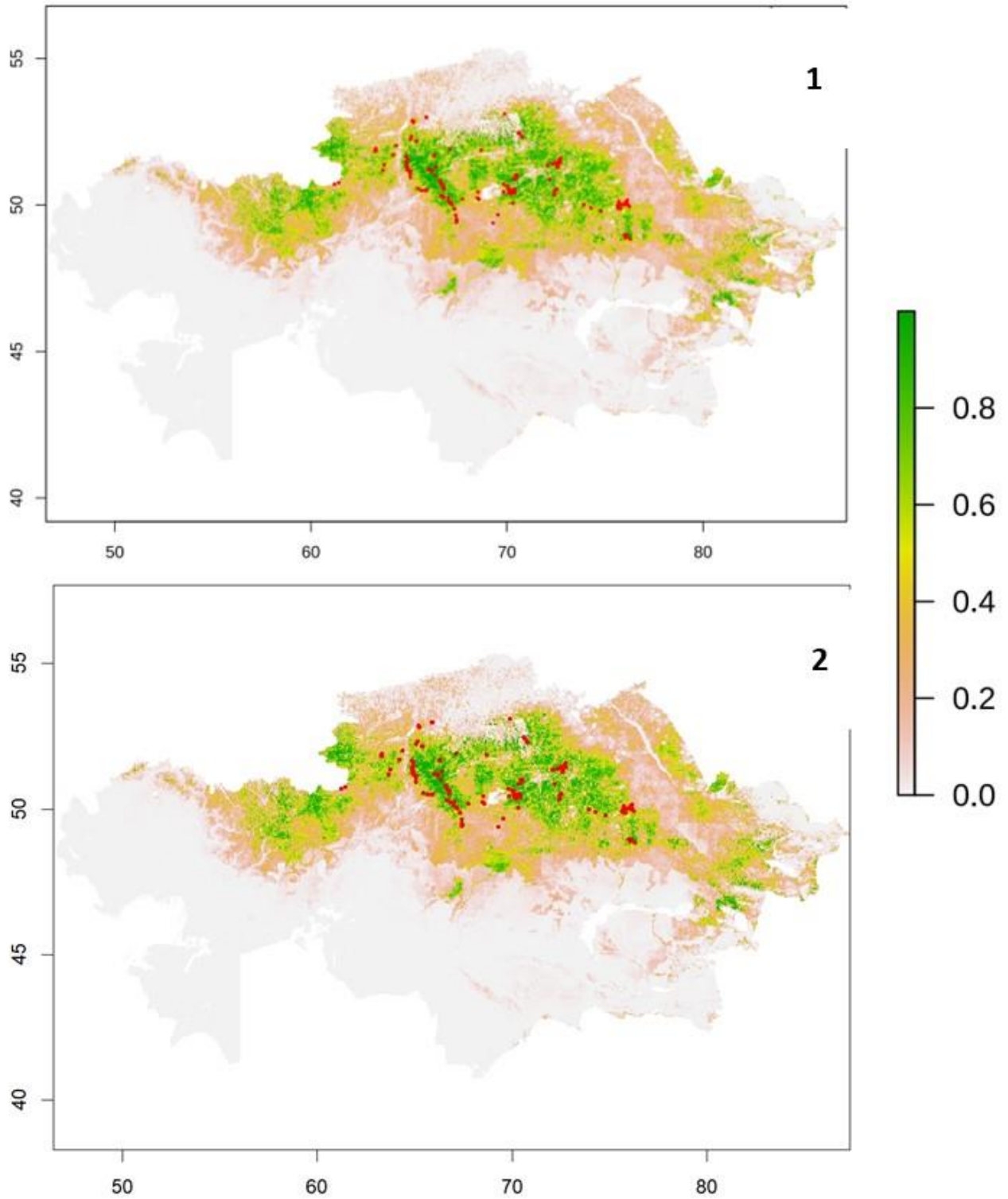


## Omission and Predicted Area graph/MaxEnt Model 5

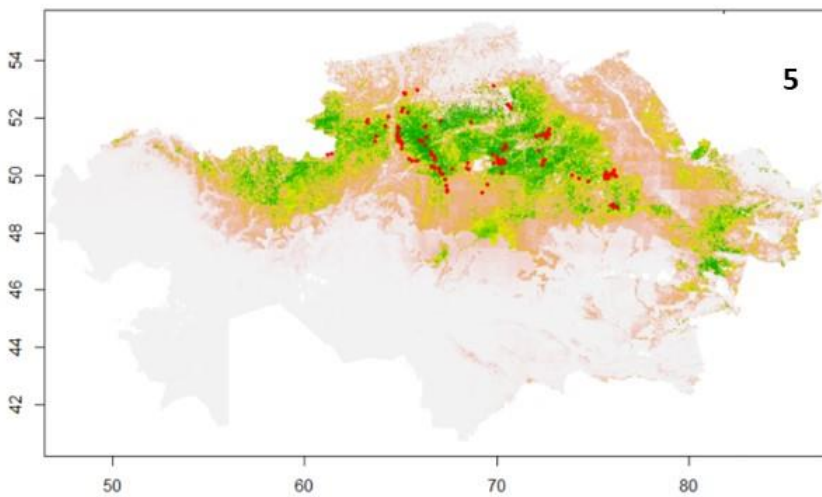
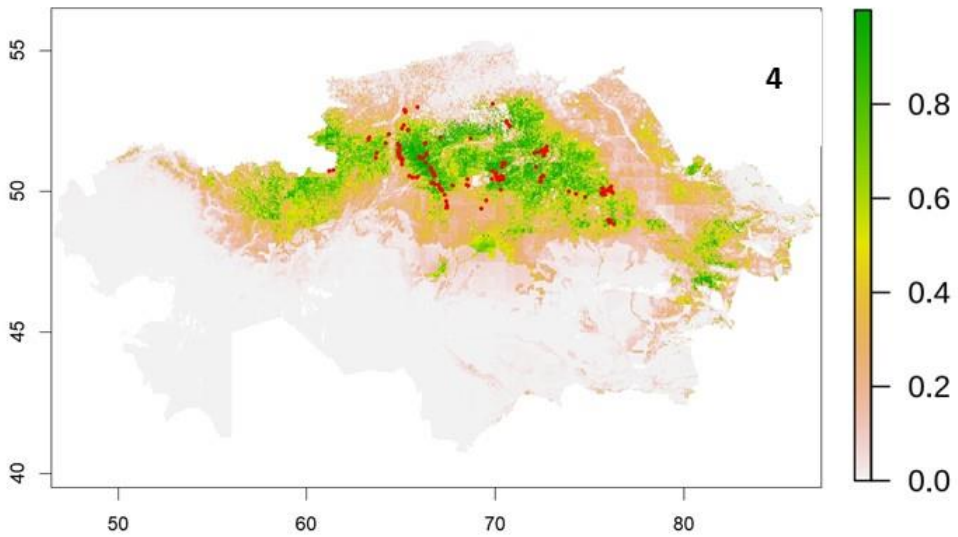
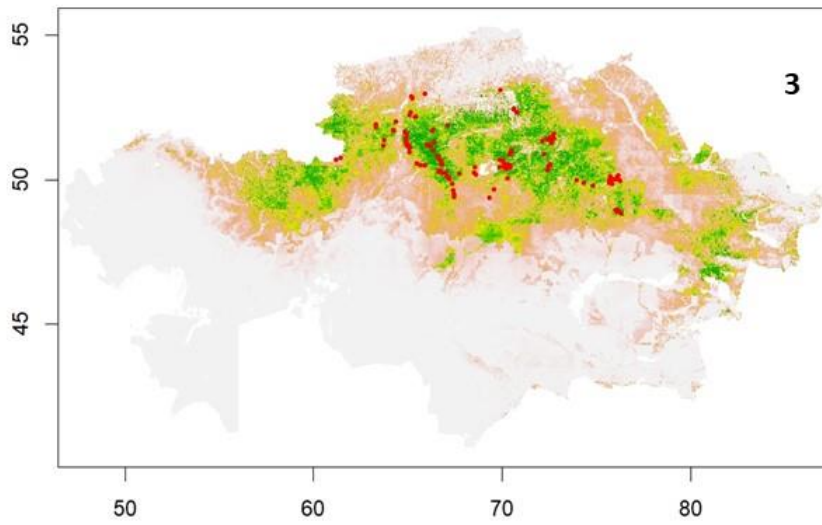


## Appendices C (maps)

### MaxEnt habitat suitability maps/ Models 1 and 2

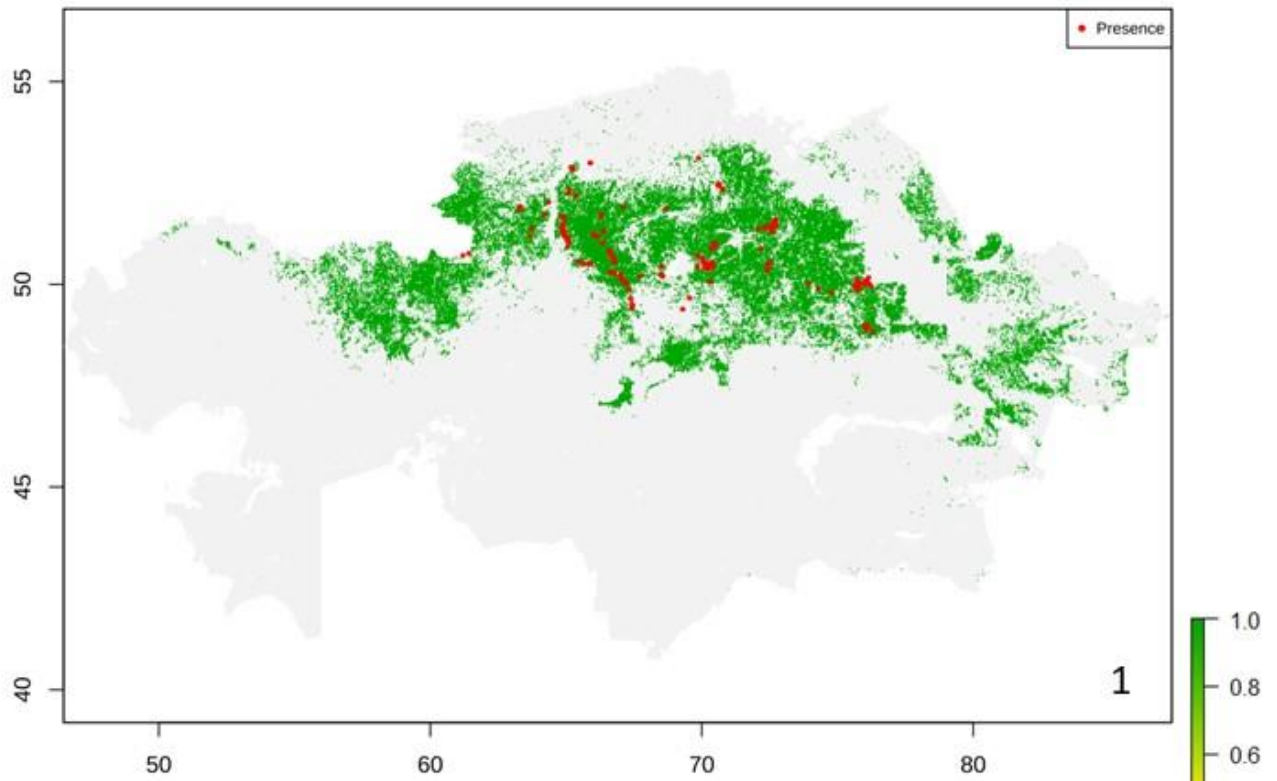


### MaxEnt habitat suitability maps /Models 3 and 5

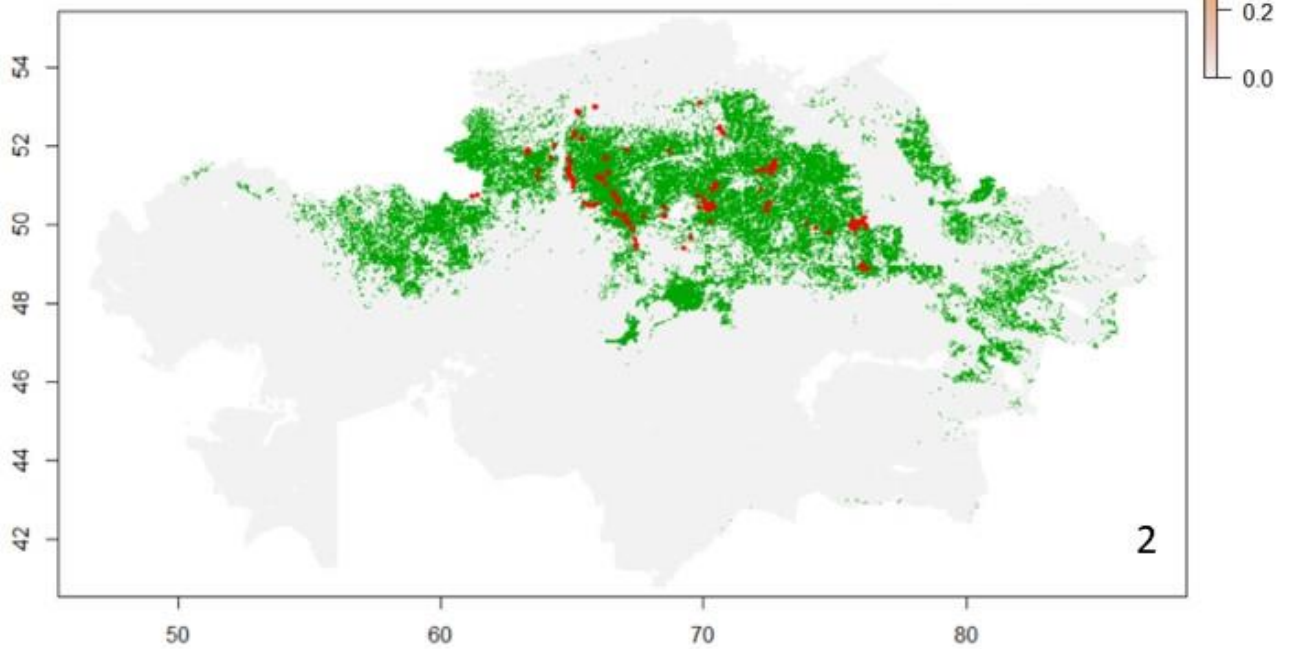


## MaxEnt Binary maps 1-2

Binary Map (Threshold = 0.395 )

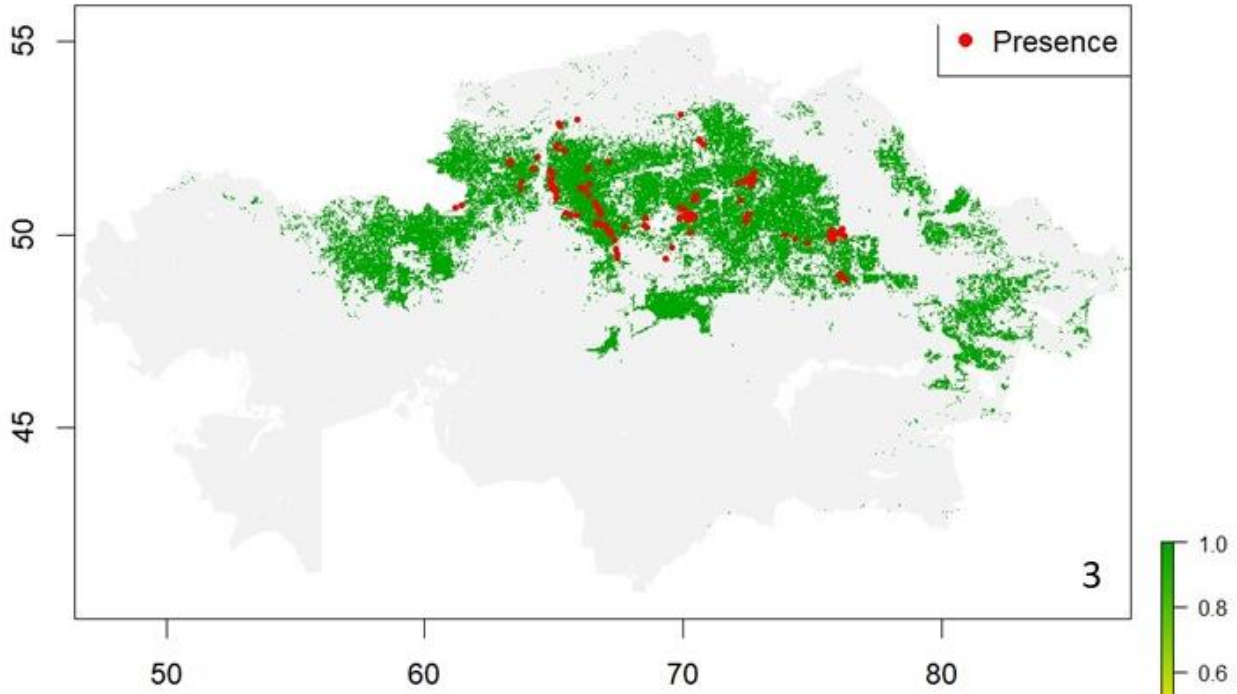


Binary map (Threshold = 0.399 )

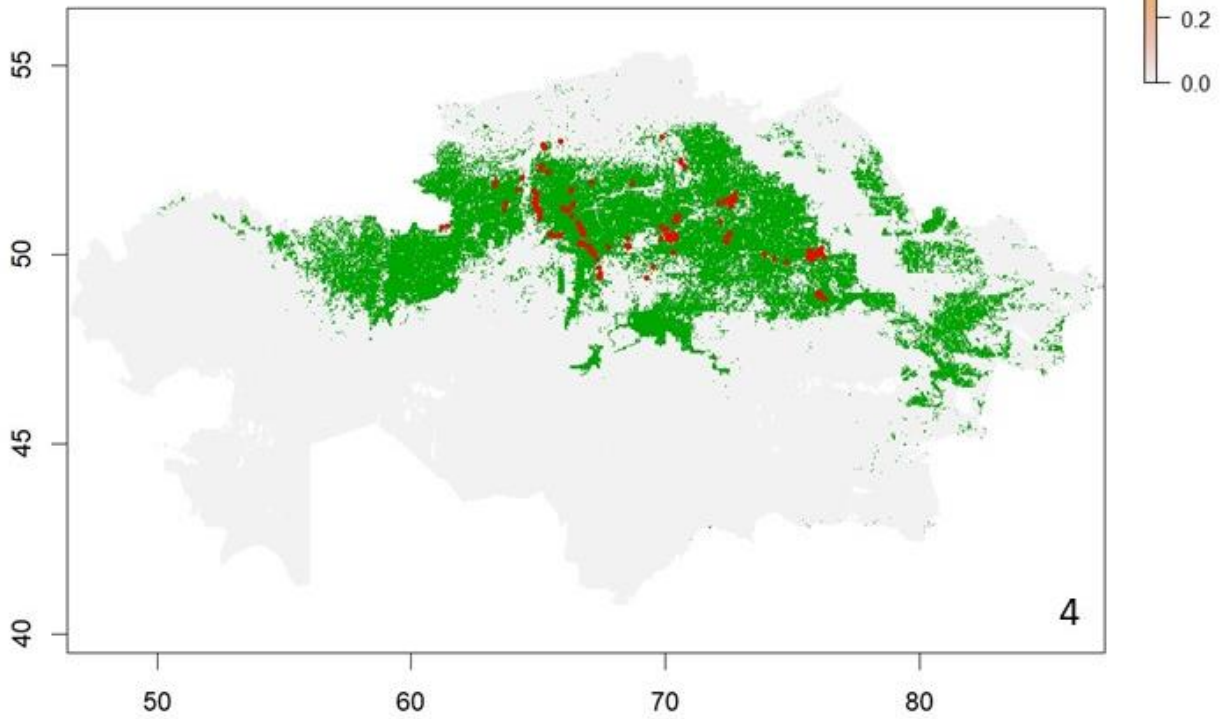


### MaxEnt Binary maps 3-4

Binary map (Threshold = 0.401 )



Binary Map (Threshold = 0.328 )



**MaxEnt Binary map 5**  
**Binary map (Threshold = 0.265 )**

