Spatial Analysis Based on Vulpes vulpes

Group Members
Yuxin Chen
Siyue Gao
Xiaoxiao Fu

Date
April 27, 2023

Dataset
https://www.gbif.org/species/5219243

Contents

1 Introduction

Spatial statistical analysis is a powerful tool for understanding geographic space and studying geographic phenomena. It involves transforming geographic models into mathematical statistical models for quantitative description. Spatial statistical analysis has been applied in many fields, such as population, medicine, economics, and land change. By using spatial information analysis techniques, we can study spatial patterns and trends of geographic phenomena through observation and experimentation of the original data model. This allows us to gain new experience and knowledge and use it as a decision-making basis for spatial behavior. In this study, we will apply spatial statistical analysis techniques and construct a Poisson Process model to study the spatial distribution of Vulpes vulpes (Red Fox) in British Columbia, Canada, in order to explore the factors that influence its distribution.

The Vulpes vulpes is one of the largest and most widely distributed members of the Canidae family, and its distribution covers much of the Northern Hemisphere, including most of Northern America, Europe, and Asia. As an omnivorous animal, the Vulpes vulpes plays a balancing role in the food chain in the natural world, and its food sources are diverse. Monitoring and managing the distribution of this species' populations can help monitor the balance and stability of the ecosystem, and as a widely distributed species, studying the distribution of Vulpes vulpes can also help scientists understand biodiversity and the operation of ecosystems.

There have been a relatively large academic literature on spatial statistical analysis of the Vulpes vulpes. JJA Dekker et al.[1] proposed spatial-statistical methods for the classification of free-ranging foxes in 2001, CD Soulsbury et al.[2] used spatial statistics to analyze the movement of Vulpes vulpes away from their birthplace or home territory in 2011, and PA Fleming et al.[3] studied the dietary composition of Vulpes vulpes in Australia in 2022.

The data used for this study is obtained from the Global Biodiversity Information Network (GBIF). We selected distribution data of Vulpes vulpes in British Columbia, and cleaned the data by removing entries with missing longitude or latitude values, as well as empty variable columns, resulting in a dataset contained 177 entries with 130 variables, including the GBIF identifier, taxonomic information, occurrence date and precision, and identification reference for each species and so on. In this study, we only utilized the longitude and latitude data for our analysis.

In this study, we also used the BC Environmental Variables dataset, which provides a set of environmental variables as covariates to describe the trends of observed point data. This dataset includes five covariates: Window (grid cells covering the entire province), Elevation (topographic height information), Forest (forest coverage), Dist_Water (distance to the nearest water source), and HFI (hazard factor index). These variables can help us understand the impact of altitude, forest distribution, water sources on the habitat and activity of the species, and can combine multiple environmental factors to predict species distribution.

2 Methods

2.1 Explore and inspect data

First we need to convert the latitude and longitude coordinates in the dataset to BC Albers projection for the coherence of covariates information provided. `ppp`function in `spatstat`package was used to convert data to ppp(planar point pattern) object. Since Visualization is always helpful to explore the data initially, we plotted all the covariates with the ppp object in the same image to show if it looks like existing some potential relationship between the vulpes vulpes' distribution and covariates. We also used `persp` function to draw perspective plots in 3 dimensions for more clear visualization. Finally we generated kernel density estimate (KDE) of the distribution of four covariates using `density` function to explore more about vulpes vulpe's distribution and capture its trend.

2.2 First moment descriptive statistics

After some basic exploration about the data, the first summary statistics we want to calculate is the average number of points per unit area ('expectation', or 'first moment'). In point pattern analysis, this quantity is called the 'intensity', denoted λ. For estimating the intensity, we first chose the simplest estimator of λ,the number of points in our window B divided by the area of B under an assumption of homogeneity. Then we used a quadrat test to determine whether the assumption of homogeneity is met. That is, under a null hypothesis that the intensity is homogeneous, it test for significant deviations from complete spatial randomness (CSR). Then with a tiny p-value as return, it suggested that there is a significant deviation from homogeneity. We also visualized both the quadrats and estimated intensity using `quadratcount` function to justify assumption of homogeneity was not appropriate for this dataset.

Thus we can focus on inhomogeneous intensity. A spatially varying intensity can also be estimated non-parametrically by kernel estimation, which places kernels' on each datapoint (often bi-variate Gaussian) and optimizes the bandwidth[6]. We plotted kernel estimated intensity for ppp object as well as the hot spot with the kernel estimate to identify areas of elevated intensity.

2.3 Second moment descriptive statistics

As we already knew that this is a non homogeneous point process, we now were interested in the relationships between points. Patterns existing in the intensity can be caused by external factors underpining the distribution of $\lambda(u)$ or relationships (correlation) between points. As for describing correlation, we started with Ripley's K function instead of Morisita's index due to the homogeneity assumption was broken. Ripley's K-function builds metric directly off of the separation distances between all ordered pairs of distinct points and thus describes the cumulative average number of points falling within distance r of a typical point. We corrected the inhomogeneity using `Kinhom` function and generated bootstrapped estimates using `envelope` function to obtain confidence intervals. Then we plot both empirical K-function and theoretical K-function with CI for identification of deviation.

Then we used the derivative of the K-function with respect to r, the Pair correlation function, which only contains contributions from inter-point distances = r and provides some information on the behaviour of the process. We also corrected for inhomogeneity using `pcfinhom` function and `envelope` function to obtain confidence intervals for comparison.

2.4 Modelling

- Pre-analysis

Since we know much of the correlations between points are due to relationships with covariates, we first need to determine what covairates the intensity depends on. We started to explore the underlying deterministic model form using `rhohat` function, a non-parametric estimate via kernel estimation. Then after getting first impression of the relationship between intensity and covariates, we decided to built log-like polynomial model of all four covariates.Then we checked for the collinearity before the model fitting with a number of potentially correlated variables.

- Model fitting

First we centered and scaled all four covariates since those were on very different scales. Then we used `ppm` function to build a log-like cubic model of all four covariates. Looking at the Ztest value, we only kept the significant item and built another model using `ppm` function. Finally we used `bs` function in `splines` package to realize a GAM model for better fitting result. Then we viusalized them for better communication.

- Model selection/Validation

We first applied AIC to three models we built as well as the intercept model(null model), then with the return result, we decided to focus on full model and spline model due to the after selection model did not varied much from full model in terms of AIC. Then we chose likelihood ratio tests using `anova` function to compare the model we build and the null model. Finally we visualized both the residual plot and partial residual plot for the two model to learn more about the models' capacity.

3 Results

3.1 Descriptive Statistics

- Inhomogeneous Intensity

Under the assumption of homogeneity, we first calculate the intensity of vulpes vulpes/km^2 in BC, which was 0.000187/km^2. However, the assumption has been denied by quadrat test with a tiny p value as a return. Also as shown in Figure 1, the assumption of homogeneity is not appropriate for this dataset clearly. The number and color in each quadrat deviated a lot. The middle-northwestern part in BC as well as southeastern part has more vulpes than other regions, especially the border of BC.

Figure 1: quadrat visualization

In Figure 2, the kernel estimation helps to show the intensity vividly all over the province. From this we can identify areas of elevated intensity such as the northwestern region of BC province.

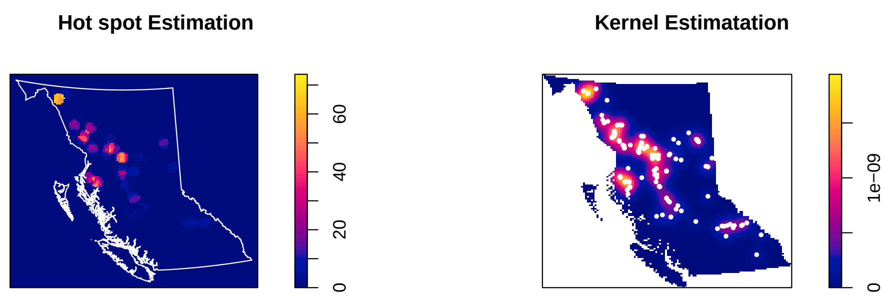

Figure 2: kernel estimate visualization

- Independence

Based on the fact the point process is inhomogeneous, we used Ripley's K function and Pair correlation function(g function) to explore its relationship between points. In Figure 3, after corrected for inhomogeneity, the estimated empirical K-function is within the confidence limits of the sample mean of K from simulations. This suggests that vulpes in BC are not correlated/clustered in space. As for g-function plot, it almost near by the line of one, which also suggested for no clustering and avoidance but independence.

3.2 Model

- Covariates Analysis

Exploring at four covariates influence on intensity, we can find that the intensity of vulpes

vulpes in BC is related to elevation, forest cover, HFI and distance to water, and all the relationship seem to be polynomial. For instance, vulpes vulpes seem to like living in low cover forest area and where is comparatively far to water.

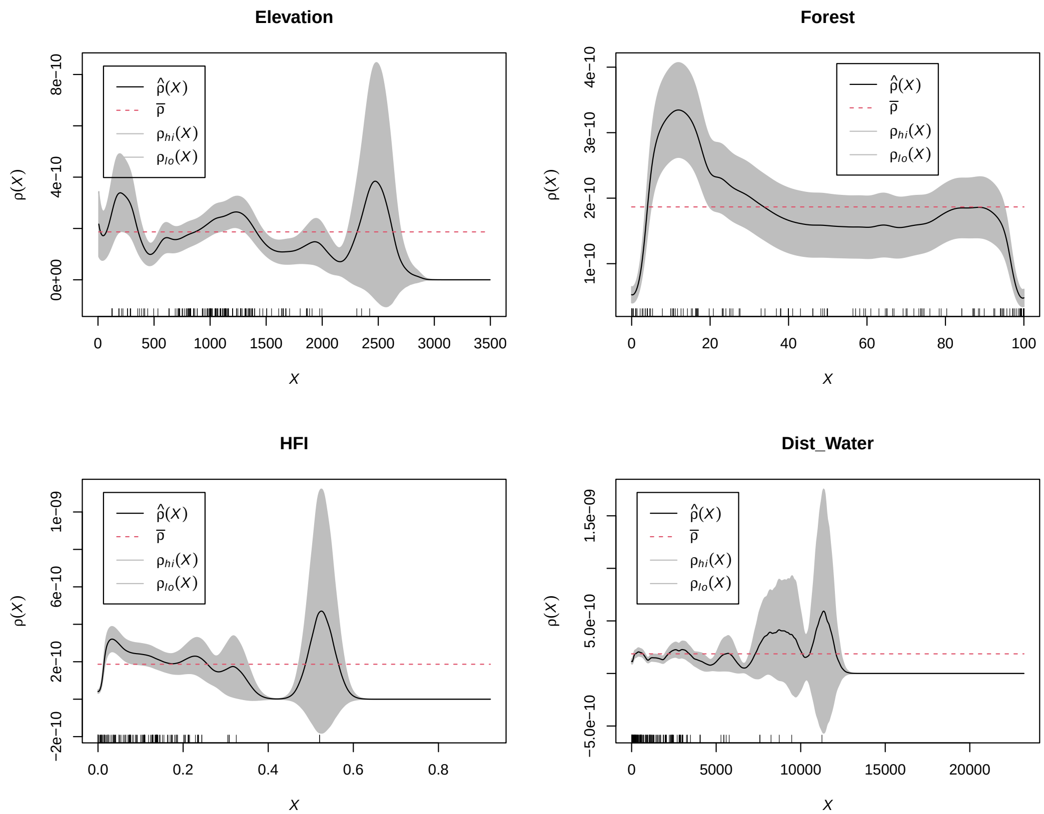

Figure 3: covariates-intensity plot

3.3 Model fitting

As discussed in Methods, we focused on two models in this report. One used cubic form of four covariates. The model formula we obtained is

$$\lambda_{B.vulpes\ vulpes}(u) = e^{\begin{array}{l} -22.85+-0.19\ \mathrm{elevation}(u)-0.11\ \mathrm{elevation}(u)^2-0.01\ \mathrm{elevation}(u)^3 \\ -0.82\ \mathrm{forest}(u)+0.44\ \mathrm{forest}(u)^2+0.40\ \mathrm{forest}(u)^3+0.30\ \mathrm{HFI}(u) \\ -0.32\ \mathrm{HFI}(u)^2+0.03\ \mathrm{HFI}(u)^3-0.48\ \mathrm{Dist\_Water}(u) \\ +0.34\ \mathrm{Dist\_Water}(u)^2-0.05\ \mathrm{Dist\_Water}(u)^3 \end{array}}$$

Notice that here we have already centered and scaled all four covariates for calculation in R. Another model is based on GAM, and we set the degree of freedom for Elevation, Forest, HFI,Dist_Water to 11 from the observation of rho plot. We also visualized these two models. From Figure 4 we can see for first model, it captures the overall trends well except it overestimates the north part intensity. While in second model, the model seems to be overfitting a little bit as it is not very good at showing all trends in the province.

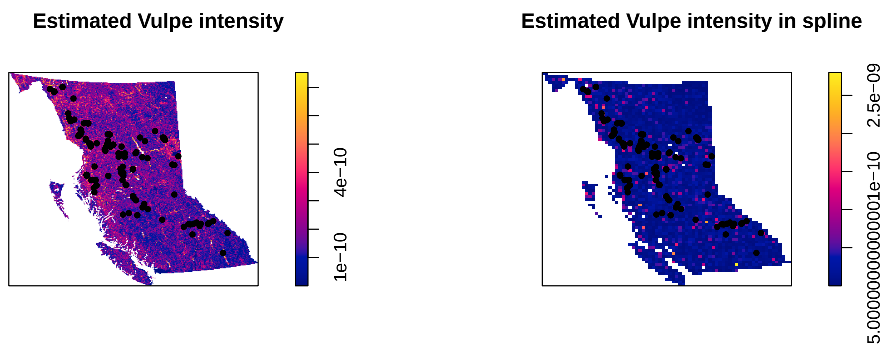

Figure 4: Esitimated Intensity

3.4 Model Elevation

Both models are significant better fit than null model. As for AIC value, while the null model is 8286.2, the first model is 8257.7 with δAIC 28.5, and the second model is 8216.6 with δAIC 69.6.

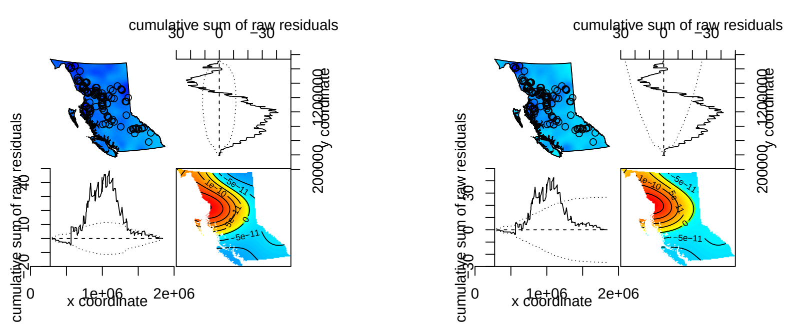

Figure 5: Residual Plot

In Figure 5, we plotted the residual for both model, and it clearly that the residual is less in the second model(the color is lighter). Both model has average-enough residuals value in the whole province.

For the part residual in first model, we can see that it did not capture some obvious trend in terms of elevation, HFI and Dist_Water in Figure 6. In Figure 7, While in the second model,

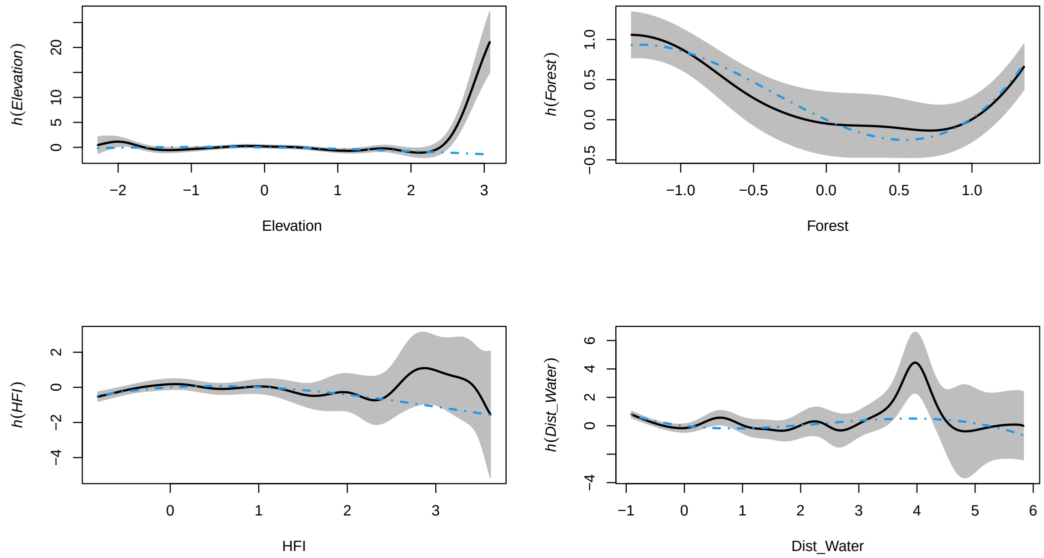

Figure 6: Part Residual Plot

it look like a much better fit due to we have added a lot of complexity. Still there are some improvement space.

4 Discussion

Generally speaking, in terms of research about distribution of vulpes vulpes in BC province, we took a series of step to study on it. First, The spatial distribution of vulpes vulpes is inhomogeneous in BC province. That is, the intensity is not a constant but in relationship with other covariates or in points themselves. The middle-northwestern and southeastern region has higher intensity than other areas in BC. Second, it was found out that there is no clustering or avoidance but independence between points in terms of vulpes vulpes distribtuion in BC. This would suggest the much of the correlations between points are due to relationships with covariates, rather than relationships between the points. Furthermore, the above findings led us to find out the relationship between the intensity and covariates. For the four covariates we had in hand(Elevation, Forest, HFI and Dist_Water), they were all in polynomial relationship with covariates. Thus we built a model which formula is

$$\lambda_{B.vulpes\ vulpes}(u) = e^{\begin{array}{l} -22.85+-0.19\,\text{elevation}(u)-0.11\,\text{elevation}(u)^2-0.01\,\text{elevation}(u)^3 \\ -0.82\,\text{forest}(u)+0.44\,\text{forest}(u)^2+0.40\,\text{forest}(u)^3+0.30\,\text{HFI}(u) \\ -0.32\,\text{HFI}(u)^2+0.03\,\text{HFI}(u)^3-0.48\,\text{Dist\_Water}(u) \\ +0.34\,\text{Dist\_Water}(u)^2-0.05\,\text{Dist\_Water}(u)^3 \end{array}}$$

.

For improving fitting, we build another GAM model which set all degree of freedom to 11

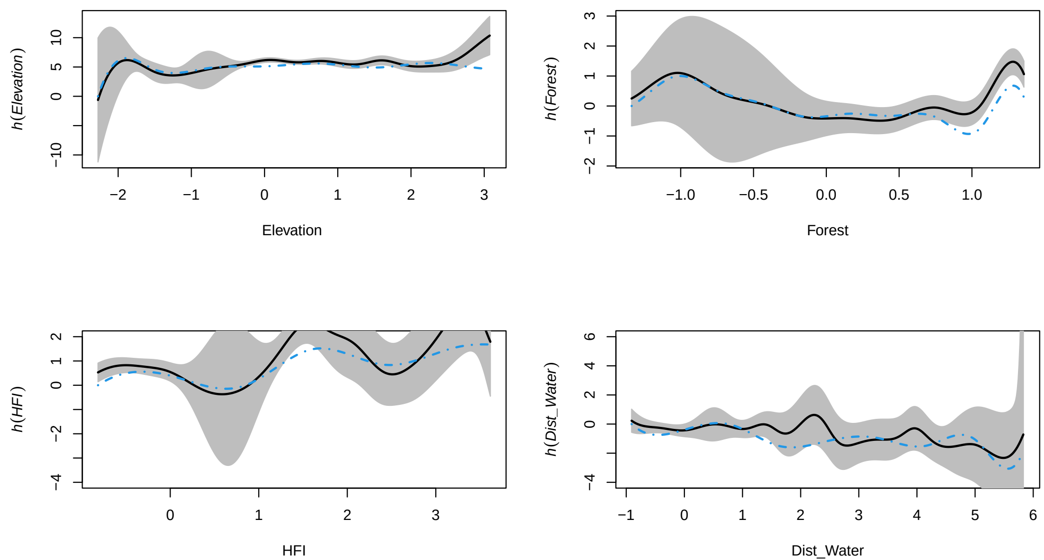

Figure 7: Part Residual Plot for Spline Model

for the complicated relationship. from models we built and some visualization tools, we are aware of that the intensity of vulpes vulpes is higher at lower elevation, lower forest cover, lower HFI value and closer distance to water.

Though those two models performed well, there still are some improvement space. For instance, the first model can not capture some particular patterns well like high elevation or high HFI value, which the second model exists some overfitting issue when looking at the who pattern of model estimation in BC.

For future potential improvement of model, more accurate coefficient-choosing tool and set of degree of freedom can be expected. For the implement of the current research, it can be applied in creating habitats for vulpes vulpes and study about its living habit.

5 Reference

[1] Dekker, J. J. A., Chardonnet, P., & Verboom, J. (2001). Spatial-statistical methods for the classification of free-ranging foxes. Journal of Applied Ecology, 38(5), 1006-1021. doi: 10.1046/j.1365-2664.2001.00642.x

[2] Bowyer, R. T., & Kitchen, A. M. (2011). Demography of a harvested swift fox population. Journal of Mammalogy, 92(1), 190-197. doi: 10.1644/10-MAMM-A-119.1

[3] Schipper, J., Newsome, T. M., Beyer, H. L., & Dickman, C. R. (2021). Diet of the introduced red fox Vulpes vulpes in Australia: analysis of temporal and spatial patterns. Australian Journal of Zoology, 69(1), 1-13. https://doi.org/10.1071/ZO20048

[4] Baddeley, A., Rubak, E. & Turner, R. (2015). Spatial point patterns: methodology and applications with R. CRC press.

[5] Khanmohammadi, M., Waagepetersen, R., Nava, N., Nyengaard, J.R. and Sporring, J. (2014) Analysing the distribution of synaptic vesicles using a spatial point process model. 5th ACM Conference on Bioinformatics, Computational Biology and Health Informatics, Newport Beach, CA, USA, September 2014.

[6]Baddeley, A., Rubak, E. & Turner, R. (2015). Spatial point patterns: methodology and applications with R. CRC press.