Alien species modelling via relational event models

Student: Niccolò Zuppichini
Advisor: Ernst Wit
Co-Advisor: Igor Artico

Università della Svizzera Italiana

Master's thesis defense
June 22, 2022

An **alien species** is a species found outside of their native region.

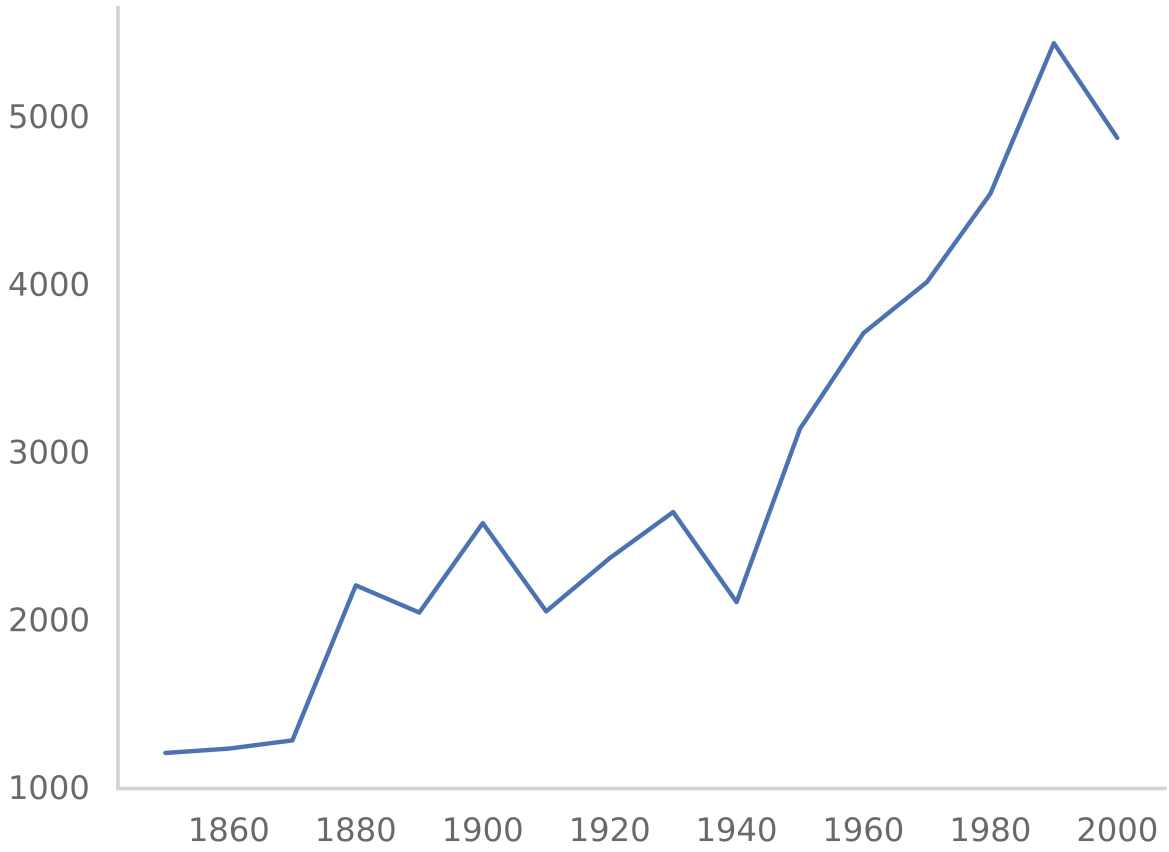


Figure: Number of invasion per year.

- They are a threat to biodiversity.
- The economic impact of alien species in Europe is estimated to be close to 13 billion dollars annually.

We are interested in studying the dynamics of co-invasion of alien species.

The Alien Species First Records database

The dataset contains the years of the first establishment of alien species in regions worldwide. To date is the most exhaustive source of first records of alien species currently available

The current version includes 61,751 invasions of 23,191 species divided into 18 taxonomic families in 276 regions ranging from 7000 BC to 2020 AD.

The most abdundant families are Vascularp Plants and Insects.

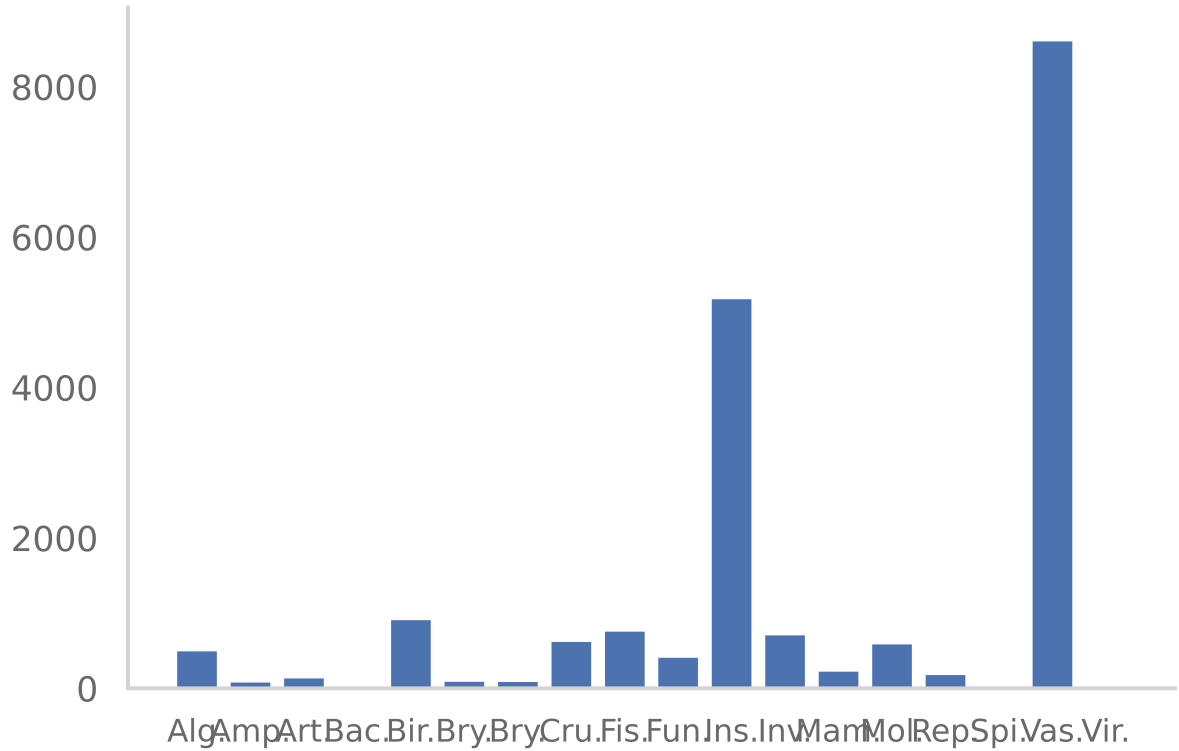


Figure: Number of species per taxonomic family.

We focus our research in the timeframe lasting from 1850 to 2010 to ensure reliability and trustiness of the dataset.

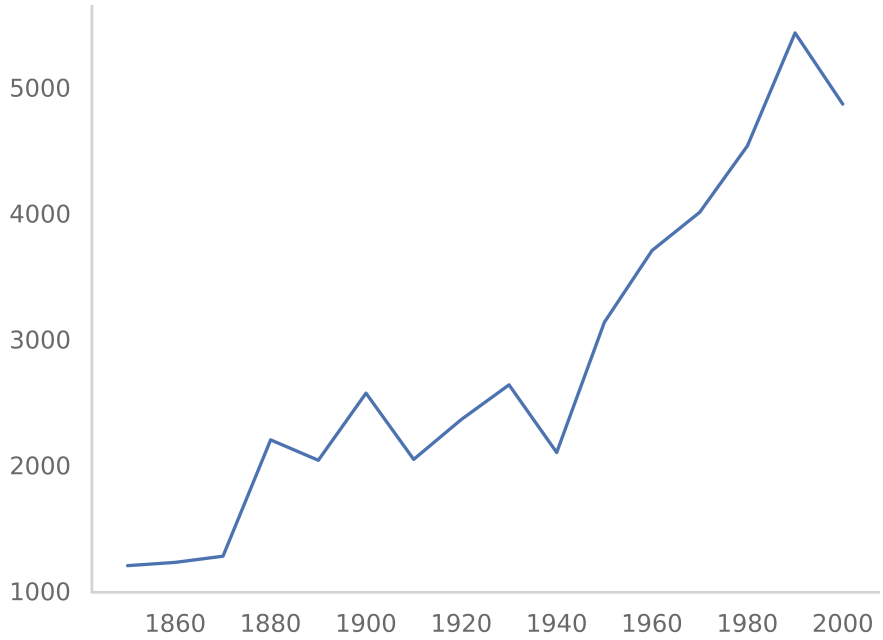


Figure: Number of invasion per year.

Spatial complexity and dataset sparsity

Taking into consideration all taxonomic families, this timeframes covers a total of 19390 alien species. Unfortunately, this number of species is too large for our model requirements.

The peak number of invasions per year is close to 5000 invasions around the year 2000, with an average number of invasions per year during the study of about 2800. Considering this number of average invasions per year and the fact that there are more than 20’000 species the dataset is considered to be sparse.

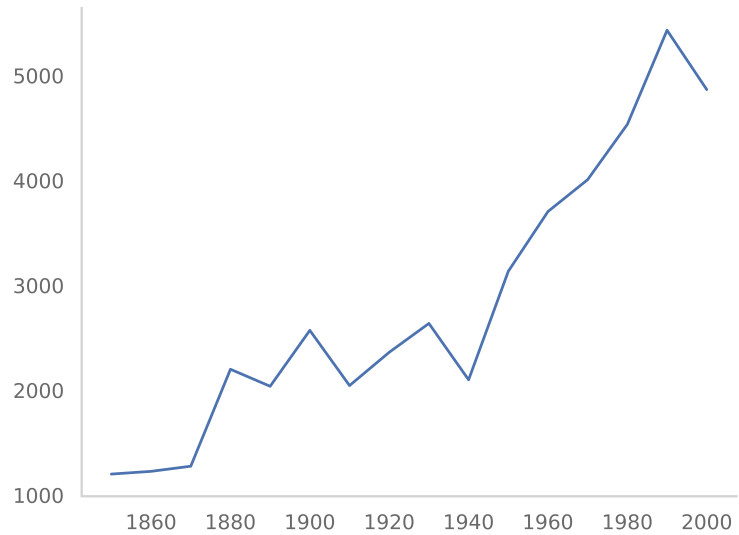


Figure: Number of species per taxonomic family.

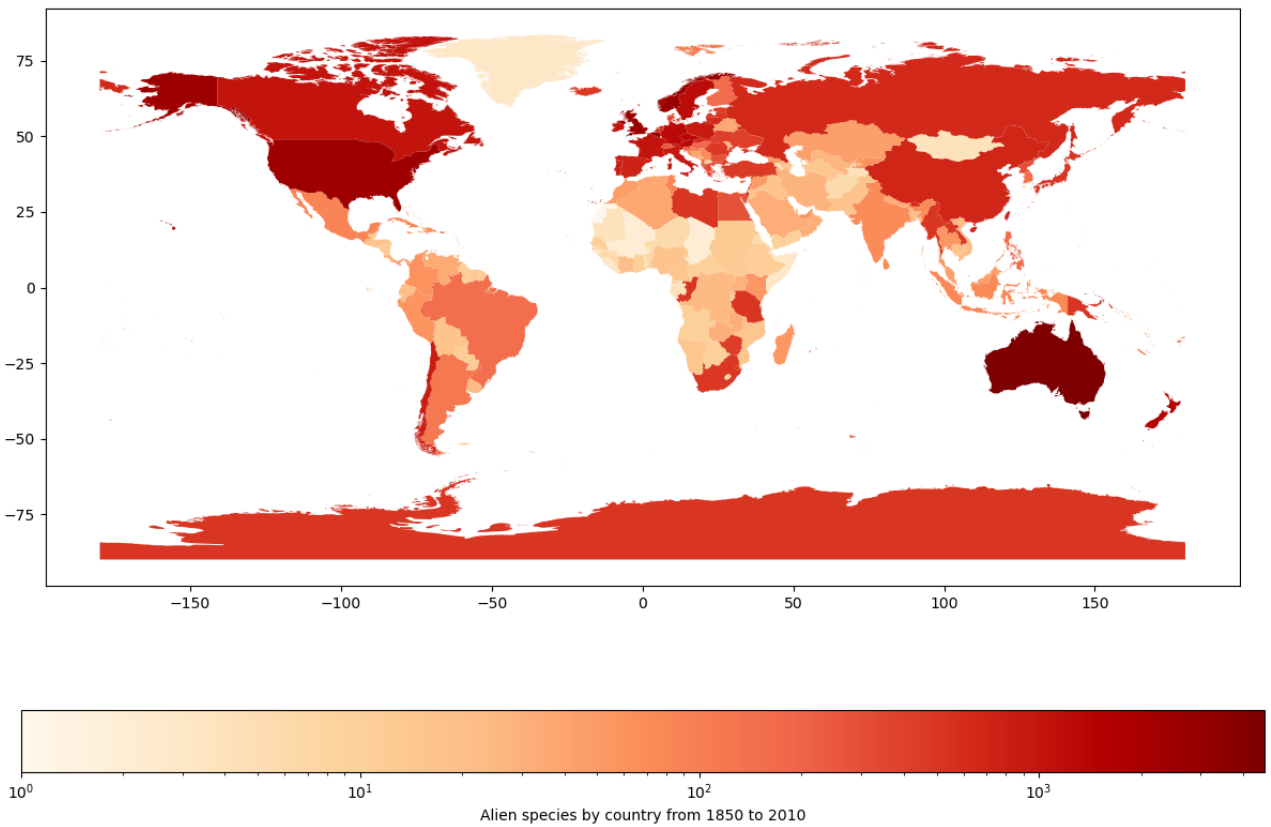
Filtering assumptions

To reduce the spatial complexity of the dataset we made the following two assumptions:

1. Remove all islands
2. Remove all species that did not invade at least 5 regions

Assumption 1: Remove islands

We are mainly interested in the interactions between larger regions. Islands are not large enough to be interesting.



Assumption 2: Remove non-relevant alien species

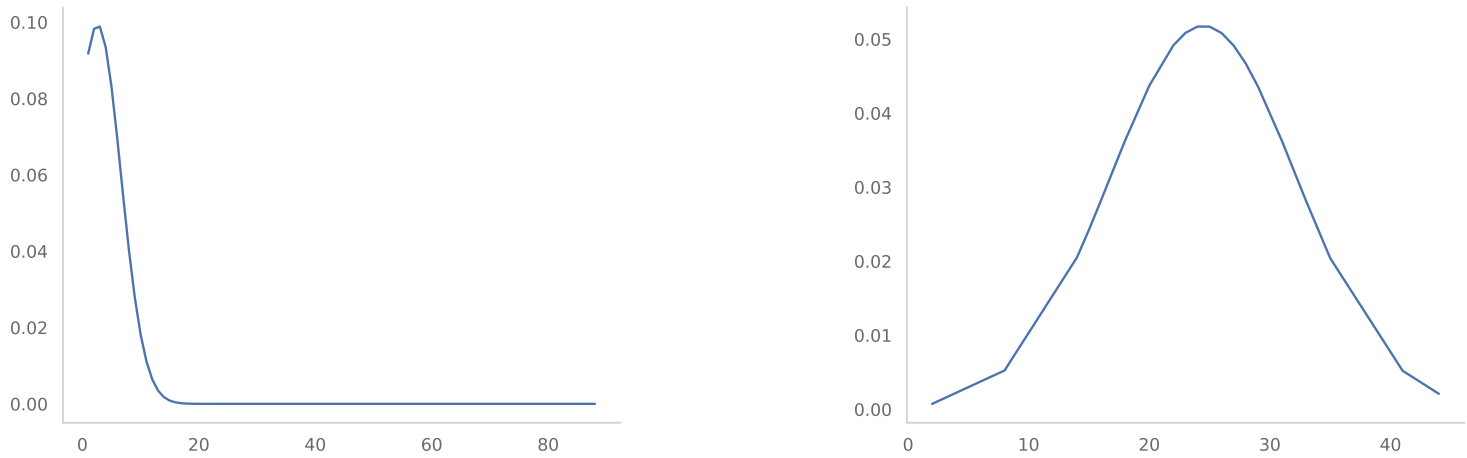


Figure: Probability density function (PDF) of the number of invasions during the entire time frame of the study (1850-2010). The left graph shows that the PDF is strongly left skewed. The second PDF, on the right, is generated after removing all species that did not invade a significant amount of regions.

Under these two assumption, we focused our research on 1724 species and 153 regions in 15947 invasions from 1850 to 2010.

Relational Event Model (REM)

A relational event is an interaction between a sender and a receiver at a specific timestamp. An invasion of a species into a region can be described as a relational event. A relational event model (REM) studies the temporal sequence of such relational events.

$$e_i = (s_i, r_i, t_i)$$

This framework has many advantages in the fact that it is able to study underlying temporal patterns, can efficiently deal with time-varying variables and it is an extremely adaptable hypothesis testing tool.
However, due to the high number of possible local network configurations, it is really hard to keep track of the past configurations to study dynamic networks.

The risk set $\mathcal{R}_t$ is the set of all dyads for which, at a given timestamp t, a relational event can occur,

$$\mathcal{R}_t = \{(s, r) \mid \text{may occurs at } t_i\}$$

We assume the waiting time $T_{s,r}$ for a particular dyad to happen between a species s and a region r is assumed to be exponentially distributed

$$T_{s,r} \sim Exp(\lambda(s, r, t))$$

with the hazard function λ is defined as

$$\lambda(s, r, t) = \lim_{\delta t \lim 0^+} \frac{P(t \leq T_{s,r} \leq t + \delta t | t \leq T_{s,r})}{\delta t}$$

Every invasion of a species $s \in S$ and region $r \in R$, where S, R are respectively the sets of species and regions in the dataset, can be modelled as a multivariate Poisson counting measure N.

$$N_t(s, r) = \text{number of invasions of } s \text{ in } r \text{ in the interval } [0, t]$$

An alien species is not allowed to leave a region after an invasion has happened and, eventually, invade it a second time. Therefore, an invasion can happen only once. The counting process N is therefore bounded by 1.

A latent space relational event model

We dispose species and regions into a latent space $X \in \mathcal{R}^n$. Each actor is assigned a position inside X.

By disposing of the species-region interactions in a latent space we "summarise" their configuration history hence we can effectively approximate the past information and train a REM more efficiently.

$$\begin{cases} x_k = x_{k-1} + \epsilon \ , \quad \text{where } \epsilon \sim N(0, \Sigma) \\ y_k \sim Poi(\lambda_k) \end{cases} .$$

The observed counts y are a result of the dynamics of the actor's positions and interactions in the REM and latent space X. The stochastic intensity of the counting process N is modeled by the function λ. Heuristically, we assume that the rates λ are functions of the distance between species $x_s(t)$ and regions $x_r(t)$ in the latent space.

$$\lambda^{s,r,k} = exp\left(\alpha - dist(x_s(k), x_r(k)) \cdot \lambda_1(k)\right) \cdot I$$

- The position x of each vector in the latent space X is constrained by the distance to all the other vectors inside X.
- The distance in this space represents an affinity of each species s to create a connection to a region r in the REM.
- The structure of the latent space is driven by the similarity of nodes.
- We are interested in studying the structure of X and infer on the co-invasion of species.

We aim to maximise the marginal likelihood:

$$\int_x L(\theta, \Sigma; y, x) dx$$

However a direct maximization is not possible. To infer on the structure of the latent space X, and the parameters θ and Σ we use the Expectation Maximization (EM) algorithm.

1 **while** *not converged* **do**
2 E-step: $E_{x|y}[Q(\beta^n, \Sigma^n]$
3 M-step: $\beta^{(n+1)} = \text{argmax}_\beta \; Q(\beta^n, \Sigma^n)$
4 **end**

By exploiting the natural formulation of our latent space model we can solve the computationally challenging E-step by means of a Kalman filter by approximating the quantity Q in function of its first two conditioned moments E, V.

Initialize:

$\hat{x}_0 = \mu_0 = E[x_0]$

$\hat{V}_0 = V_0 = E[(x_0 - \hat{x}_0)(x_0 - \hat{x}_0)^T]$

for *k=1,...,n* **do**

  Prediction step:

    $\hat{x}_k^- = \hat{x}_{k-1}$

    $\hat{V}_k^- = V_{k-1} + \Sigma$

  Update step:

    $K_k = V_k^- H_k^T (H_k V_t^- H_k^T + R_k)^{-1}$

    $\hat{x}_k = x_k^- + K_k(y_k - h(\hat{x}_k^-))$

    $V_k = (I - K_k H_k) V_k^-$

end

Algorithm 1: Kalman Filter.

The process y is not linear but we can linearise it around the mean, to apply a Kalman Filter

$$\begin{cases} x_k = f(x_{k-1}) + \epsilon \,, \quad \epsilon \, N(0, \Sigma) \\ y_k = h(x_k) \end{cases} \tag{1}$$

Where $f = I$ and h is a function dependent on the latent space x. Providing that $h, y \in C^1$, we can linearise the process around the mean and reconduct the problem to the basic linear case.

Inference framework

1 **Initialize:**

2.1 $\hat{x}_0 = \mu_0 = E[x_0]$

2.2 $\hat{V}_0 = V_0 = E[(x_0 - \hat{x}_0)(x_0 - \hat{x}_0)^T]$

3 **while** *not converged* **do**

4 **E-Step**:

5.1 Kalman Filter

5.2 Smoother

6 **M-Step:**

7.1 $\beta = \text{argmax}_\beta Q$

8 **end**

Algorithm 2: Latent space REM inference.

Results

We now focus on the results of our model applied to the dataset and the dynamics of the species and regions in the latent space.

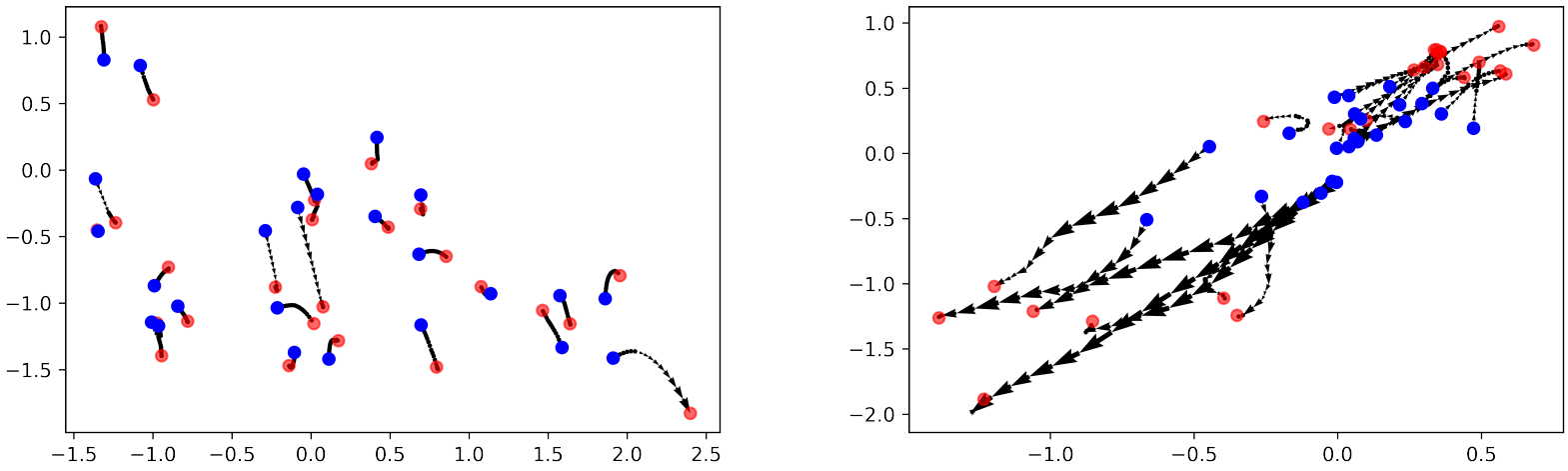


Figure: A set of 25 randomly chosen species (left) and regions (right) and their dynamics in the latent space.

Species that move in a similar way inside the latent space tend to co-invade. Similarly, regions that move analogously will have the tendency to be co-invaded by the same group of species.

Applying *Ordering points to identify the clustering structure* OPTICS to the species and regions in the latent space detected 67 clusters for species and 5 clusters for regions.

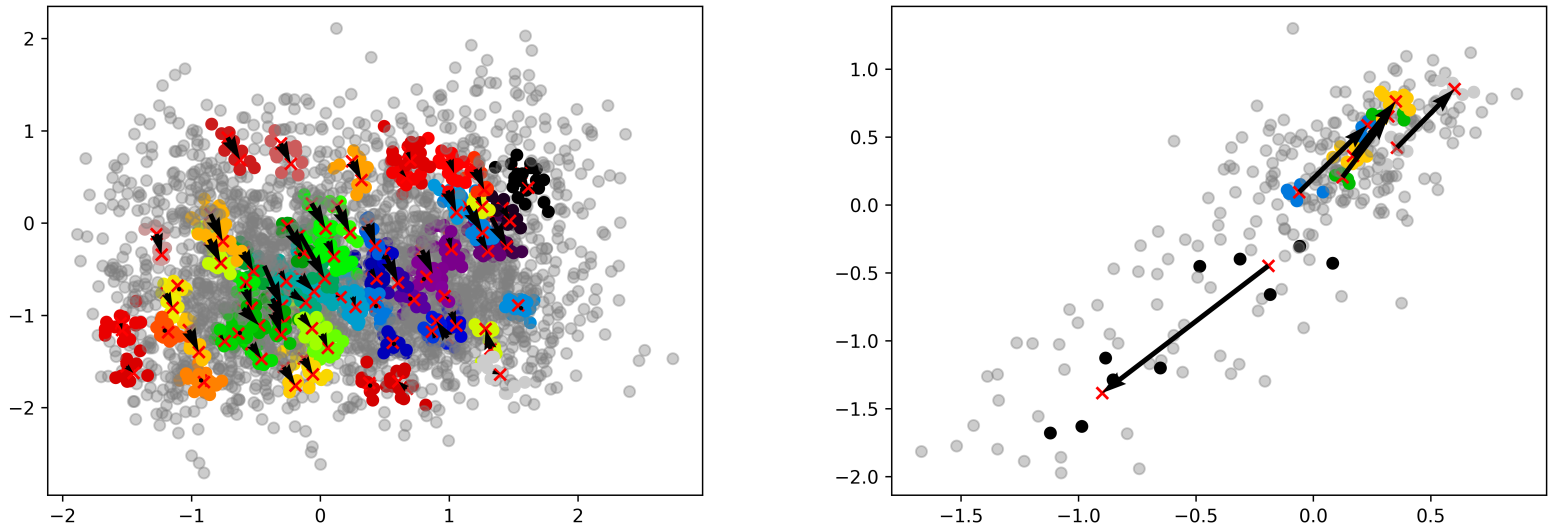


Figure: Latent space initial and final configurations. The black arrows shows the movement in space from the initial state to the final state.

Insights from the species clusters

We select the five largest species and clusters.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Algae			11 %		
Birds	7 %	6 %			7 %
Crustaceans	7 %		16 %	6 %	
Fishes			5 %	6 %	
Fungi				6 %	
Insects	14 %	12 %	16 %		36 %
Mammals	7 %				
Reptiles		6 %			
Vascular plants	64 %	75 %	53 %	81 %	57 %

Table: Species clusters composition.

Insights from the reguon clusters

We select the five largest species and clusters.

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Russia	Armenia	Lesotho	Peru	Iraq
Italy	Chad	Mauritania	Andorra	Nicaragua
Canada	Iran	Gibraltar	Belize	Niger
Estonia	Mongolia	Nepal	Burkina Faso	Vietnam
Slovakia	Somalia	Suriname	Libya	Zambia
			Palestine	CAR
			Senegal	
			Tajikistan	
			Gabon	
			Congo	

Table: Region clusters composition.

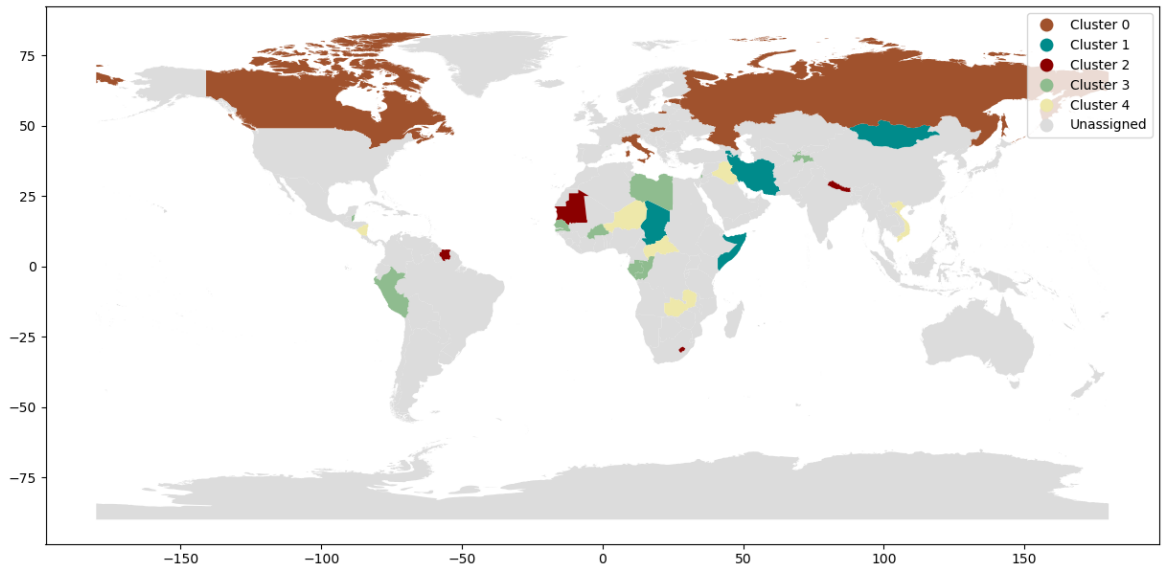
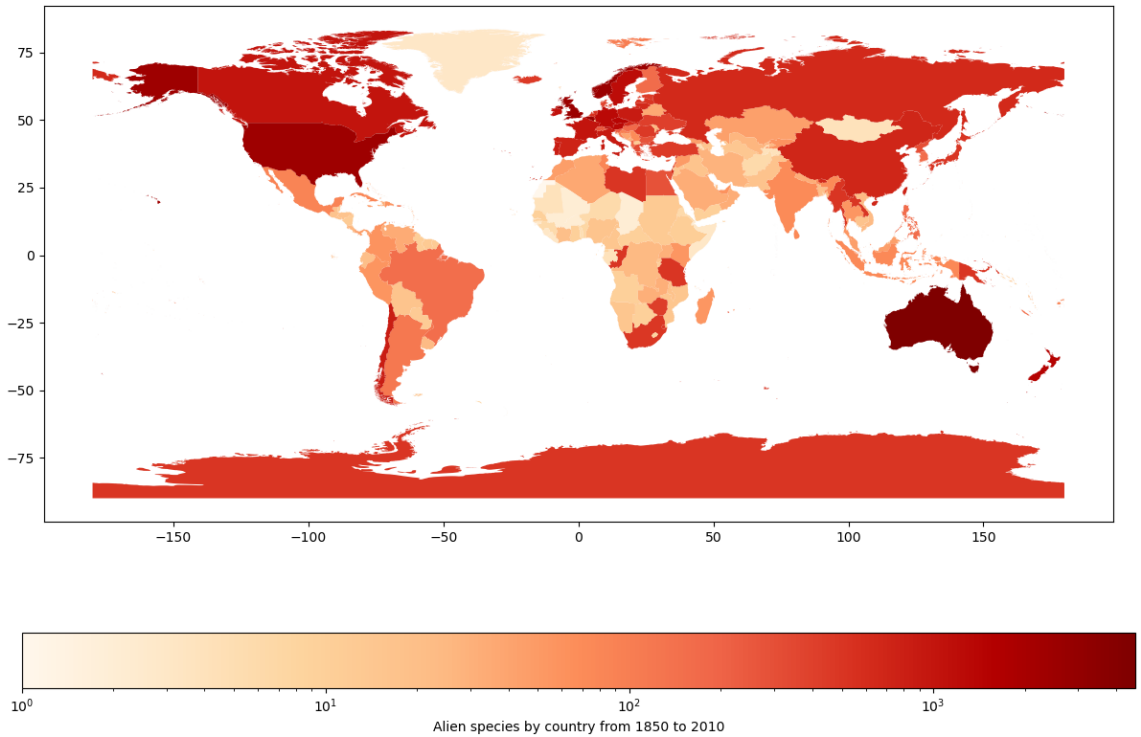


Figure: The five clusters of regions on the world map.

On the biasness of the dataset

The First Records Dataset is currently the largest dataset available for first-time records of alien species, however:

- it is an unevenly populated dataset towards taxonomic groups such as plants and insects
- it is an unevenly populated dataset towards regions such as Europe as a whole
- the information about the year of an alien species introduction in a region is not precise and most of the time does not correspond to the year of actual introduction, which might have happened decades later


75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150
10^0
10^1
10^2
10^3
Alien species by country from 1850 to 2010

Conclusion

- We proposed a relational event model with an underlying latent space formulation.
- We showed a real world application of the model to the on-going problem of alien species.

Thank you for you attention.

Alien species modelling via relational event models

Student: Niccolò Zuppichini
Advisor: Ernst Wit
Co-Advisor: Igor Artico

Università della Svizzera Italiana

Master's thesis defense
June 22, 2022

Please, questions