Alien species modelling via relational event models

Master's Thesis submitted to the
Faculty of Informatics of the *Università della Svizzera Italiana*
in partial fulfillment of the requirements for the degree of
Master of Science in Informatics

presented by
Niccolò Zuppichini

under the supervision of
Prof. Ernst-Jan Camiel Wit
co-supervised by
Igor Artico

June 2022

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Niccolò Zuppichini
Lugano, 23 June 2022

Abstract

During the last centuries, human research on the diffusion of species substantially intensified. However, we know little to nothing about the underlying process that shapes the behaviour of alien species invasion across regions, countries, and ecosystems, where alien species is a species found outside of its native region.

In this thesis, we present a novel relational event model approach with an underlying latent space framework. We show how this model can be applied to a bipartite unidirectional graph by studying the real-world case of alien species invasions. We fit our model to the most exhaustive dataset of first alien species records currently available.

Using years of first records of 15947 invasions of alien species from 18 taxonomic groups spanning from 1850 to 2010 we show how a custom relational event model can be used to model the dynamics of species co-invasion by embedding the species-region interactions into a latent space. We then show how to infer the structure of this space and the relationship between species and regions.

The goal of this thesis is twofold: to show a novel approach to REM by using an underlying latent space and to show a real-world application of the model.

Chapter 1

Introduction

The rate at which alien species invade countries have increased over the last century and it is becoming an important issue. The economic impact of alien species in Europe is estimated to be close to 13 billion dollars annually [6]. The interest in the study of the dynamics of alien species has been increasing over the years for this reason. Unfortunately, due to the complexity of the problem, it is hard to come up with real actions. Intuitively, the invasion rate of species differs for species belonging to different taxonomic groups. However, a comprehensive global invasion dynamics study of the last centuries subdivided by taxonomic groups is still lacking. The dynamics of invasive species are driven by a multitude of factors simultaneously. These factors can be complex involving exogenous, endogenous, sociological, ecological, and socio-economic factors [5]. It is important to develop a framework capable of studying and analyzing all these underlying effects simultaneously.

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. REMs were first developed in the field of social network analysis [1]. 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. Due to its versatility, in the years it is been applied in a large number of different fields such as animal-animal interaction [12], the evolution of team networks [9] and national hospitals collaborations in transfering critical patients [10].

In the years, there have been a large number of different statistical approaches with the goal of predicting and describing the spread of alien species. However, traditional statistical tools, such as multidimensional scaling, fail to capture the similarity of nodes on this kind of dataset due to their bipartite nature (i.e. no interaction of the kind species-species and region-region). A REM approach is a good approach to this problem, but, 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.

In this thesis, we contribute to the practical applicability of a REM with a latent space underlying framework modeled on a bipartite social network graph with a real-world case study of the dynamics of co-invasion of species. Our model is capable of taking multiple species simultaneously into consideration and to study the underlying relationship between different species co-invading regions by training on the most exhaustive source of first records of alien species

in multiple regions in the world. We propose a different approach from the traditional literature by disposing the species and regions (which sometimes I simply refer to as nodes) into a latent space and by fitting a custom REM on it. By studying the proximity of the nodes in the latent space, our model is capable of determining the group of nodes that tend to co-invade a region and, in an analogous way, the group of regions that are more likely to be invaded by the same group of species. 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. In this space, nodes are allowed to move by attracting or repulsing other nodes. The closeness between nodes describes they tend to interact (which is to co-invade a region). Within the limits of this latent space configuration, a node location at a certain time describes its history evolution and by studying this evolution we can infer the group of nodes that have similar tendencies and behaviours.

We structure the rest of this thesis as follows. In chapter 2 we make a preliminary study of the dataset of the alien species and discuss previous research. In chapter 3 we present an exhaustive statistical background of the methods used for the reader. In chapter 4 we provide a formulation of our latent space REM and show how to make inferences on it. In chapter 5 we report our results. In chapter 6, we conclude with a final discussion.

Chapter 2

Materials and methods

Our study of the co-invasion of species is based on the Alien Species First Records database [3] which to date is the most exhaustive source of first records of alien species currently available. The dataset contains the years of the first establishment of alien species in regions worldwide. This dataset has been updated many times to include a more in-depth amount of data on the alien species [4] and nowadays still in continuously update. The current version includes 61,751 invasions of 23,191 species divided into 18 taxonomic families, listed in Table 2.1, in 276 regions[1] ranging from 7000 BC to 2020 AD.

Algae	Birds	Fishes	Mammals	Vascular plants
Amphibians	Bryophytes	Fungi	Molluscs	Viruses
Arthropods	Bryozoa	Insects	Reptiles	
Bacteria	Crustaceans	Invertebrates	Spiders	

Table 2.1. The 18 taxonomic families.

We focus our research on the time interval spanning from 1850 to 2010. Taking into consideration all taxonomic families, this timeframes covers a total of 19390 alien species. As Fig. 2.3 shows, the distribution of alien species is not uniform between taxonomic families, with the largest two taxonomic families in the dataset being the "Vascular Plants" and "Insects". This is important to take into consideration when analysing the results because a non-uniform distribution of species into taxonomic families will result in a higher probability of a species interacting, in some way, with a species belonging to one of the most abundant taxonomic families.

[1] Regions generally correspond to countries but in the dataset available almost half regions are islands.

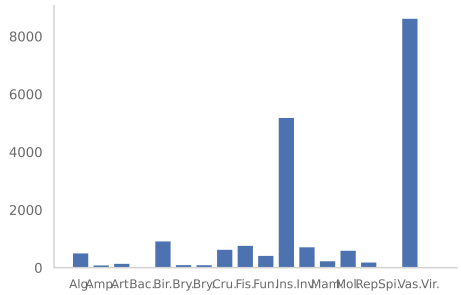


Figure 2.1. Histogram of Taxonomic Families and their respective number of alien species in the interval from 1850 to 2010.

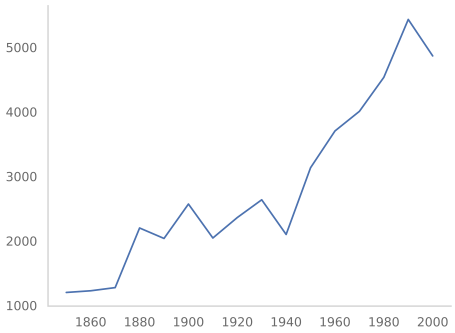
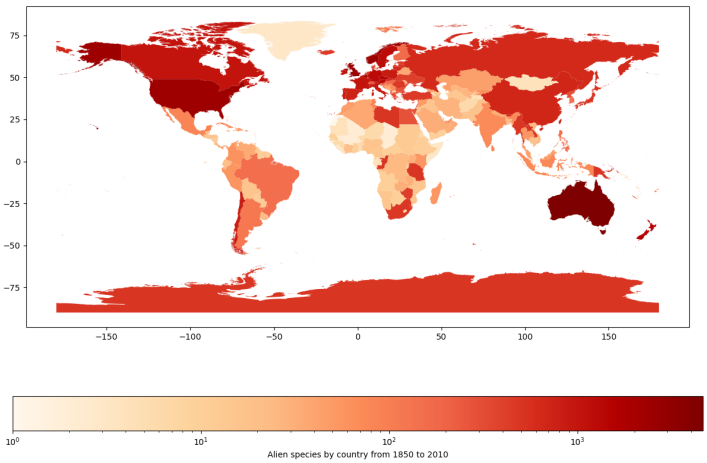


Figure 2.2. Number of invasion per year from 1800 to 2020.

Intuitively, the amount of invasion per year increases over time mainly due to factors such as worldwide effects such as the growth of intercontinental trading and an increased worldwide connectivity [7]. This claim is indeed supported by the dataset. The number of invasions per year is monotonically increasing (Fig. 2.2). The high downward trend around the year 2000 of invasions should not be trusted because the dataset is being constantly updated and most recent data may still be missing. With the same argument, data before 1850 may not be accurate and reliable enough for an in-depth study. For these reasons, we decided to trim the dataset from 1850 to 2010 to ensure the trustiness of the data.



Contrary to what was expected, some regions have a significantly larger number of invasions during this timeframe. Australia has had more than 4000 invasions. Belgium, the UK, USA, and Norway more than 2000. In contrast, most of the regions in the dataset have less than 3 invasions. Switzerland has been invaded 258 times. Considering that the average number of

invasions per region is 200, the distribution of invasion species per region is largely skewed.

Figure 2.3. Symmetric interactions between insects in co-invading events. Each node represents a species and the edge represents the strength of the co-invasion factor. A thick edge indicates that a spread of one of the two species in the relationship is typically followed by the invasion of the second. Image gently taken from Juozaitiene et al. [7]

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. To effectively show the high sparsity of the dataset conveniently, I reported two probability density functions (PDF). The PDF shown in Fig. 2.4 is left-skewed with a mean of 4 and a standard deviation of 5, this means that most species only invaded 4 regions out of the 276 in the entire timeframe of our study. It is worth mentioning that the maximum number of invasions is 97 with one outsider that invaded 176 regions.

2.1 Data preprocessing

Due to the high sparsity of the dataset and the large computational cost of our method, we made a couple of assumptions to reduce the data complexity and intrinsic computational requirements.

The first assumption is that any species that did not invade at least 5 regions during the entire timeframe is not relevant to our study. We hence removed more than 14'000 species resulting in a more computationally manageable dataset of around 1700 species. In Fig 2.4

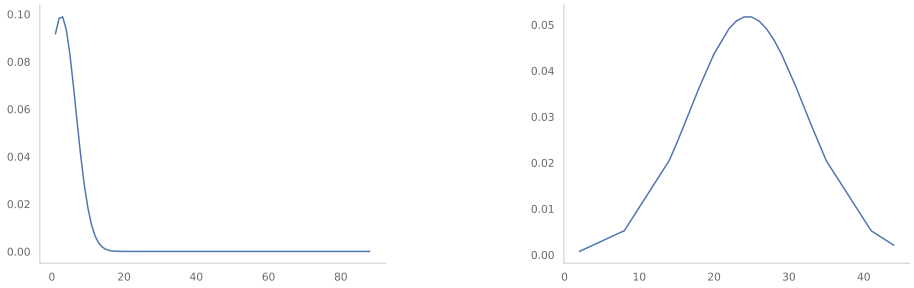


Figure 2.4. Probability density function (PDF) of the number of invasions during the entire time frame of the study (1850-2010). The left graph shows that the PDF is strongly left skewed. The second PDF, on the right, is generated after removing all species that did not invade a significant amount of regions.

I reported the resulting PDF resulted by removing all species that did not invade at least 5 regions.

The second assumption is that islands do not significantly contribute to the diffusion of species. We hence removed 123 regions from the dataset corresponding to islands. The reason to remove islands is two-fold. First, we are mostly interested in studying the interactions of alien species between larger regions and islands are typically not large enough to be significant. Furthermore, long-distance invasions [2] are typically rare [13] and islands are generally more remote. This assumption is supported by the dataset, the average number of invasions that happened in islands is 80 compared to the 200 average invasions of all regions combined. However, there are a couple of islands worth mentioning that contrast this assumption. The Hawaiian Islands and the Azores islands have more than 1000 invasions. We are interested in studying large region interactions and islands generally are "intermediary steps" in long-distance interactions between large regions such as continents. Removing these intermediary steps does not affect the interaction between larger regions, which we are interested in.

Under these two assumptions, we filtered all non-irrelevant species and regions out of the dataset to 1724 species and 153 regions and 15947 invasions from 1850 to 2010.

[2]The distance between regions is considered as the distance between the closest borders.

Chapter 3

Literature background

In this section, we cover all the necessary background tools to develop our latent space relational event model. First, we introduce the concept of a latent space model employing a state space model. Then we cover in-depth how to infer this model. We introduce then the reader to the Expectation-Maximization and Kalman Filter algorithms.

3.1 State space model

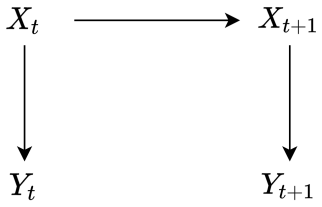


Figure 3.1. State space model diagram representation.

A state-space model is a class of temporal probabilistic models that describe the statistical dependency of latent variables and observed measurement, respectively in the state space $X \in R^d$ and the observed space $Y \in R^m$. In the simplest state-space model, the dynamics are assumed to be a random walk where the jumps between two-time intervals are affected by noise. The first example of such models is the linear Gaussian state-space model, given as follows for $k = [0, ..., n]$

$$\begin{cases} x_k = A_k x_{k-1} + \epsilon\,, \quad \epsilon \sim N(0, \Sigma) \\ y_k = B_k x_k + \delta\,, \quad \delta \sim N(0, \Delta) \end{cases} . \tag{3.1}$$

where the matrix $A \in R^{dxd}$ is the state-transition matrix where d is the dimensionality of the state space X. Similarly, $B \in R^{mxm}$ is the observation-transition matrix where m is the dimensionality

of the observation space Y. The linear Gaussian model can be extended in various ways but in this section, to keep things simple, we focus on the linear Gaussian model.

The state and observation equations are generally or partially unknown and the parameter $\theta = (A, B, \Sigma, \Delta)$ and $x(k)$ need to be jointly estimated. Maximum likelihood is a well established method in the statistics field to make such parameter estimation. One popular method to estimate state space parameters is the Expectation Maximization algorithm.

3.2 Expectation Maximization

The expectation Maximization (EM), introduced in 1977 by Dempster et al. [2], is a family of algorithms that iteratively cycles between two states until convergence is satisfied. In doing so, it first computes the conditional expectation of the complete likelihood and then maximizes it. This approach is really powerful when applied to problems in which a direct calculation of the likelihood $p(y|\theta)$ is computationally unfeasible. Furthermore, we are in particular interested in the application of this family of algorithms to problems where the full data is not available or some variables are latent (see Sec. 3.1).

The first step called, the expectation procedure (E-step), computes the expectation of the logarithm of the complete data likelihood Q

$$Q(\theta^{(k)}) = E_x\left[log(p(x, y|\theta^{(k)})\right]$$

then, with the updated parameter $\theta^{(k)}$ we maximizise Q on the second step, called Maximization step (M-step) and obtain $\theta^{(k+1)}$. The procedure is illustrated in Fig. **??**.

while *not converged* **do**
    E-step: $E_x[Q(\theta^{(k)})]$
    M-step: $\theta^{(k+1)} = \text{argmax}_\theta\, Q(\theta)$
end

Algorithm 1: A general Expectation Maximization framework.

The computation of the expectation $E[Q(\theta^n)]$ is typically computationally untractable. To make things computationally worse, its complexity increases with the amount of data fed to the model. However, there are many approaches available to estimate this quantity [14]. In this thesis, we are interested in an approach to estimate Q called the Kalman Filter.

3.3 Kalman Filter

In 1960 R.E. Kalman published his paper describing a recursive solution to the discrete-data linear filtering problem Kalman [8]. Since then, the Kalman filter gained extreme popularity in the machine learning and robotics research field, particularly in the area of autonomous navigation systems. The Kalman filter estimates the evolution of a process by using feedback control. The filter first computes an estimation of the current process state at some time, then, tries to "filter out" the measurement noise in an iterative process. As such, the Kalman filter equations can be subdivided into two categories: the prediction step and the update step. The prediction step is responsible for projecting forward in time the current state of the system. The update step

takes the a-prior estimate given from the prediction step and by taking new measurements of the error tries to obtain an improved a-posterior estimate of the state. Indeed, the general form of the algorithm resembles a predictor-corrector algorithm for solving numerical problems. In this chapter, a brief non-technical overview of the Kalman filter is provided. A reader interested in a more in-depth discussion of the formulation of the Kalman filter is advised to read Maybeck [11].

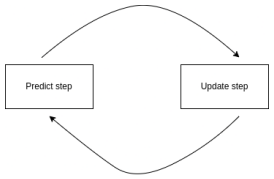


Figure 3.2. The Kalman filter cycle. The time update forwards the current state estimate in time. The measurement update filters the noise out of the projected estimate.

We are interested in estimating the configuration of the state $x \in \mathscr{R}^d$ given in equation 3.1. We consider the initial state x_0 as a random vector with a known mean and variance.

$$\begin{aligned} \hat{x}_0 &= \mu_0 = E[x_0] \\ \hat{V}_0 &= V_0 = E[(x_0 - \hat{x}_0)(x_0 - \hat{x}_0)^T] \end{aligned} \tag{3.2}$$

We define $\hat{x}_k^-$ to be our a-priori estimate of the state x at time k, given the knowledge of the previous state. Similarly, we define $\hat{x}_k$ to be our a-posteriori estimate at time k given a measurement y_k.

We forecast an update of the estimation of $\hat{x} \approx x$ in time

$$\begin{aligned} \hat{x}_k^- &= A_k \hat{x}_k \\ \hat{V}_k^- &= A_k \hat{V}_k^- A_k^T + \Sigma_k \end{aligned} \tag{3.3}$$

where Σ_t the process noise covariance matrix.

The goal of the Kalman filter is to express the a-posteriori state $\hat{x}_k^-$ in terms of a linear combination of the a-priori state and the difference between an actual measurement y and a measurement prediction $B\hat{x}_k^-$:

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

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

For a statistical motivation, I redirect the reader to Maybeck [11]. The difference $y_t - h_t(\hat{x}_t^-)$ is a residual of the estimation against the measurement, reflecting the error between the predicted measurement and the real measurement. The matrix K is called the Kalman gain matrix and is defined as

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

The Kalman gain matrix goal is to minimize the a-posteriori covariance error matrix V_t at time t. With some ease, a way of thinking K is that as the measurement error covariance R is close to zero the measurement y is trusted more and the update is made with high precision. On contrary, if the a-priori estimate error covariance P^- is close to zero, the actual measurement y is less trusted and the predicted measurement $H\hat{x}_t^-$ trust is increased. The "trust" is then reflected in the weights of the Kalman gain matrix K.

The choice of Σ is not much deterministic. In a practical way, the noise source Σ is often used to capture uncertainty in the process model. Hence, a poor model choice can suffice by simply choosing a large enough uncertainty, by selecting a large covariance Σ.

Chapter 4

Model

In the previous section, we introduced to the reader how to make inference on a linear gaussian state space model. In this section, we introduce the REM based on the underlying Alien Species First Records database [3]. We first define what the relational events are in our study, how the species and regions are disposed in the latent space, and how to infer it using the Expectation Maximization algorithm and the Kalman filter similarly to as done in the previous section.

4.1 Relational event models (REM)

Butts [1] proposed a statistical modeling framework capable of analyzing temporal information and the evolution of a sequence of relational events. Given a set of receiver nodes, a set of sender nodes, and a timeframe, a relational event e is a triplet of a sender, a receiver, and a timestamp. Depending on the underlying data, the sender set and receiver set may be the same set or be disjoint sets, depending on the kind of relationship between senders and receivers. In our study of the co-invasion of alien species, the species are the sender and the regions the receivers. Senders and receivers are disjoint sets. This relational event is unidirectional and bipartite. In fact, only a species can invade a region, a region cannot invade any other region (or species), and no species-species (or region-region) relational event is allowed. In general, a REM specifies varying distributions for all dyads as a function of past events $E = (e_1, ..., e_N)$ where each event e_i is defined as

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

In the above notation, $s_i \in S_t$ is a sender node (i.e. a species), $r_i \in R_t$ a receiver node (i.e. a region) and t_i is the time of the interaction. The sets S_i, R_i respectively contain the receiver and senders nodes.

The risk set $\mathscr{R}_t$ [1] is the set of all dyads for which, at a given timestamp t, a relational event can occur,

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

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

[1] To avoid confusion, note that Butts originally called this set the *support set*.

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

In the above equation, the hazard function λ is defined as

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

The hazard function can be interpreted as the expected number of relational events in a time interval of length one, conditionally on the previous network events. The most commonly used hazard function is the Cox proportional hazard model.

$$\lambda(s,r,t) = \lambda_0(t) \cdot exp(X_{sr}^{(t)}\beta)$$

In this equation, λ_0 represents a baseline hazard for all dyads in the risk set $\mathscr{R}$. Instead, the variable $X_{sr}^{(t)}$ is some dyadic covariate at time t. To estimate the parameters β we maximize the partial likelihood of the Cox proportional hazard model

$$L(\beta) = \prod_{i=1}^{n} \frac{exp(X_{s_i r_i}^{(t_i)}\beta)}{\sum_{(s,r)\in\mathscr{R}_{t_i}} exp(X_{sr}^{(t_i)}\beta)}$$

if the time is considered to be continuous. In a discrete time scenario, the parameter β can be estimated in a more straightforward way by maximising the Poisson likelihood by a linear regression:

$$L(\beta) = \prod_{i=1}^{n} P(Y_k | \lambda_k)$$

Typically the research interest is to estimate the factors that increase, or decrease, the likelihood L, more specifically, the parameter β.

4.2 Latent space REM

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. We define $T(s,r)$ to be a function that returns the time of invasion for specie s an region r. The function N can either take a value of one or zero depending on whether an invasion happened previously to t for a determined region and species.

$$\begin{aligned} N_t(s,r) &= 0 \; if \; t < T(s,r) \\ N_t(s,r) &= 1 \; if \; t \geq T(s,r) \end{aligned} \tag{4.1}$$

We dispose our actors, which are species and regions, into a latent space $X \in \mathscr{R}^n$. Each actor is assigned a position inside X. It is worth mentioning, to avoid confusion, that these positions

in X are disconnected from the real world. The closeness of two actors in the latent space X does not imply their closeness on the world map. We then define Y, the observed state. In our study, the observed state is the observed count of invasions modelled as Poisson. Similarly to the state space model in Eq. 3.1, we define

$$\begin{cases} x_k = x_{k-1} + \epsilon\,, & \text{where } \epsilon \sim N(0,\Sigma) \\ y_k \sim Poi(\lambda_k)\,, & \text{where } \lambda^{s,r,k} = exp\,(\alpha - dist(x_s(k), x_r(k)) \cdot \lambda_1(k)) \cdot I \end{cases} . \tag{4.2}$$

for $k = [0, ..., n]$. The rates λ are censored by multiplying them by the matrix I after an invasion occurs. The matrix I entries are either one or zero depending on if an invasion has occurred.

$$I_{s,r} = T(s,r)$$

Without loss of generality, we assume that relational events happen at discrete time intervals. The dynamics of the latent space can be modeled as a random walk where the length of the jumps is dependent on the variance Σ.

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 (i.e. to invade) a region r in the REM. The structure of the latent space X is therefore driven by the similarity between the nodes s, r. We expect species that tend to co-invade the same regions to be close in the latent space. Analogous, regions invaded by the same group of species should be near. We are interested in studying the structure of the latent space X to infer which group of species has the tendency to co-invade regions.

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 (Eq. 4.2).

Distances between vectors can be defined in many different ways. Two main popular choices are euclidean distance and dot product. These two distances give a different interpretation to the model.

$$\begin{aligned} dist_{euc}(u,v) &= \textstyle\sum_{i=0}^{d} |u_i - v_i|^2 \\ dist_{dot}(u,v) &= u \cdot v \end{aligned} \tag{4.3}$$

We chose the distance in the vector space X to be the dot product. The Euclidean distance imposes a geometric interpretation. In contrast, the dot product yields a spherical structure to the data. We expect the species-region similarities to be better described by the latter and has the additional advantage to be piecewise linear in the latent locations X.

$$\lambda^{s,r,k} = exp\,(\alpha - (x_s(k) \cdot x_r(k)) \cdot \lambda_1(k)) \tag{4.4}$$

4.3 Inference

To infer on the structure of the latent space X, and the parameters θ and Σ we use the Expectation Maximization (EM) algorithm (3.2). Since the latent state X is not observed, direct opti-

mization of the maximum likelihood is not possible. We aim to maximise the $\int_x L(\theta, \Sigma; y, x)dx$. As explained in the background section 3.2, the EM instead of directly maximizing this integral, it iteratively computes and maximizes an estimate Q of the parameters θ, Σ conditioned on the previous estimate.

$$Q(\beta^n, \Sigma^n) = E_{x|y}\left(log(p(x, y|\beta^{n-1}, \Sigma^{n-1})\right) \tag{4.5}$$

where β are the remainder parameters related to the covariates and θ estimates the volatility of the latent process.

4.3.1 Extended Kalman Filter

In section 3.3 we showed how to apply a Kalman filter to solve the E-step of an Expectaction Maximizaton problem of a Linear Gaussian model, but our process y is not linear. We can rewrite the latent space formulation in Eq. 4.2 slightly differently:

$$\begin{cases} x_k = f(x_{k-1}) + \epsilon\,, \quad \epsilon\ N(0, \Sigma) \\ y_k = h(x_k) \end{cases} \tag{4.6}$$

Where $f = I$ and h is a function dependent on the latent space x. The function h is not linear because it is a Poison distribution, however, providing that $h, y \in C^1$, we can linearise the process around the mean and reconduct the problem to the basic linear case. By assuming that a step in time in the latent state can be approximated by the conditional distribution

$$x_k \sim N(\hat{x}_k, \hat{V}_k)$$

the predict step of the Kalman filter estimates the first two moments of x_k.

$$\hat{x}_k = E[\hat{x}_k^-]$$
$$\hat{V}_k = V[\hat{x}_k^-]$$

Where $\hat{x}_k^-$ is an a-priori distribution of the state x at time k, given the knowledge of the previous state. Hence, by exploiting the natural formulation of our latent space model (Eq. 4.2) 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.
The initial state x_0 is a random vector with a known mean and variance.

$$\begin{aligned} \hat{x}_0 &= \mu_0 = E[x_0] \\ \hat{V}_0 &= V_0 = E[(x_0 - \hat{x}_0)(x_0 - \hat{x}_0)^T] \end{aligned} \tag{4.7}$$

We forecast an update of the estimation of $\hat{x} \approx x$

$$\begin{aligned} \hat{x}_k^- &= \hat{x}_{k-1} \\ \hat{V}_k^- &= V_{k-1} + \Sigma \end{aligned} \tag{4.8}$$

Equation 4.8

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 2: Extended Kalmann Filter

We now express the a-posteriori estimation $\hat{x}_k^-$ in terms of a linear combination of the a-priori state and the difference between an actual measurement y and a measurement prediction $h(\hat{x}_k^-)$ state.

$$\begin{aligned} \hat{x}_k &= \hat{x}_k^- + K(y_k - h(\hat{x}_k^-)) \\ V_k &= (I - K_k H_k) V_k^- \end{aligned} \tag{4.9}$$

Equation 4.9

We define the Kalman gain matrix K as

$$K_k = V_k^- H_k^T (H_k V_k^- H_k^T + R_k)^{-1} \quad , H_k = \frac{d}{dx} h(x) \tag{4.10}$$

Equation 4.10

We know updated the a-priori estimates of the mean and variance in Eq. 4.8 with the observation of state y. By combining equations 4.8, 4.9 and 4.10 we get an algorithm for the Extended Kalman, described in Algorithm 2.

After n iterations of the Extended Kalman Filter are performed, we move backward from the last prediction to the starting point x_0 with the goal to update and correct the predictions given by the Extended Kalman Filter. A popular smoother choice is the Rauch-Tung-Striebel smother, described in Algorithm 3.

for *k=n...1* **do**
  $B_k = V_{k-1} V_{k-1}^{-1}$
  $\hat{x}_{k-1} = x_{k-1}^- + B_k(\hat{x}_k - x_k^-)$
  $V_{k-1} = \hat{V}_{k-1}^- + B_K(V_k - \hat{V}_k^-) B_k^T$
end

Algorithm 3: Smoother

4.4 Closed form Algorithm

We now summarise into a close form algorithm based on the Expectation Maximization framework to make inference on our latent space REM. We approximate the E-step through an Extended Kalman Filter proposed in the previous section, then we maximize Q. For a number s of steps, we report our model in a closed form

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 (Alg. 2)
5.2 Smoother (Alg. 3)
6 **M-Step:**
7.1 $\beta = \text{argmax}_\beta Q$
8 **end**

Algorithm 4: Latent space REM inference.

Chapter 5

Results

In this section, we analyze the results from fitting our model on 15947 alien species invasions from 1850 to 2010, as described in Section 2. In the dataset, there are 1724 species and 153 regions. To report the results in the clearest way, we decided to distinguish the region and spaces nodes in the latent space. In Fig. 5.1 the initial and final configuration of the latent space is reported for species and regions. Here, the trajectories are linear for the sake of visualization.

Unfortunately, due to the high number of nodes, it is hard to see the evolution of the latent space. In Fig. 5.2 we focus on a set of 25 randomly chosen nodes to highlight the dynamics of nodes in the latent space. The nodes that tend to have similar behavior (i.e. have the tendency to co-invade) will be close to each other. In contrast, nodes that express a contrary behavior will push each other further. The black arrows show the trajectory in time. There are 17 timesteps in total (the years from 1850 to 2010).

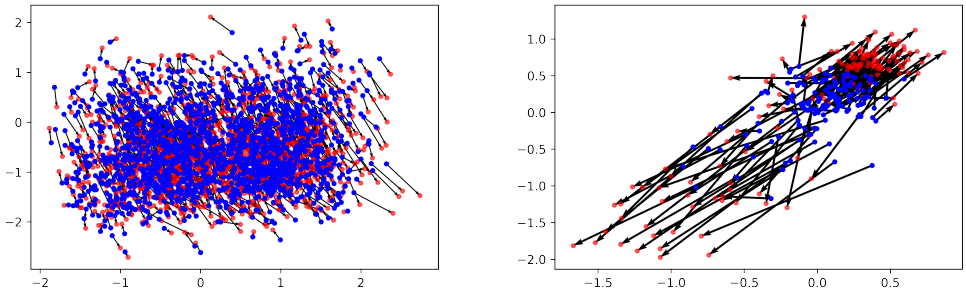


Figure 5.1. Latent space initial and final configurations. The black arrows shows the movement in space from the initial state (blue) to the final state (red).

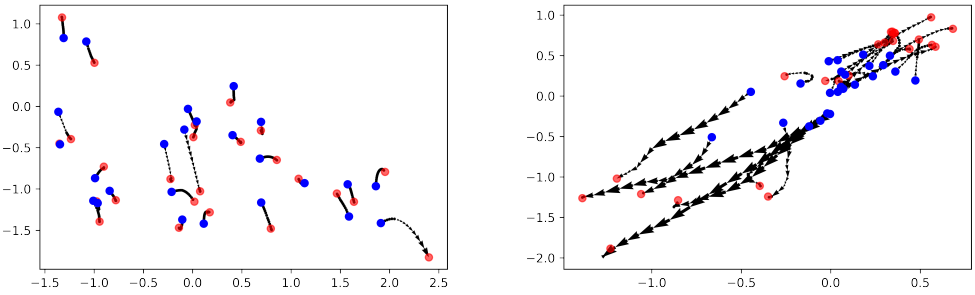


Figure 5.2. A set of 25 randomly chosen species (left) and regions (right) and their dynamics in the latent space.

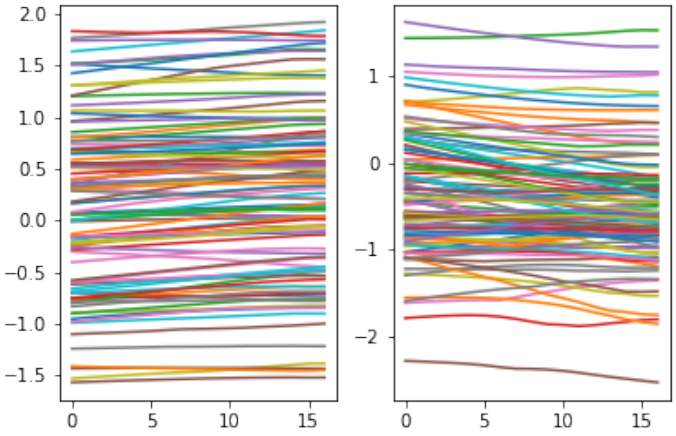


Figure 5.3. Plot of the evolution of trajectories in time for 100 randomly selected species. The left image is for the X coordinate, the right image for the Y coordinate.

In Fig. 5.3 and 5.4 the trajectories in the latent space by 100 randomly selected species and regions are reported. An in-depth look at these trajectories suggests that there are patterns in the movements of regions and species in the latent space. We now focus our interest on analyzing these patterns in the trajectories.

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. To detect nodes that move in a homogeneous way we decided to use a clustering algorithm over the estimated trajectories. We settled on using a *Ordering points to identify the clustering structure* (**OPTICS**) algorithm to perform the clustering. OPTICS is a variation of the DBScan clustering algorithm, which an in-depth cover is out of the scope of this thesis. This

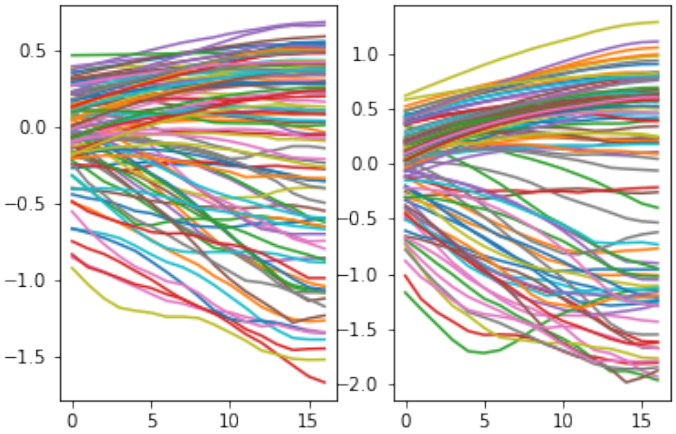


Figure 5.4. Plot of the evolution of trajectories in time for 100 randomly selected regions. The left image is for the horizontal coordinate, the right image for the vertical coordinate.

algorithm is really similar to DBSCan but addresses one of the major flaws of the algorithm. DBSCan does not perform very well on data with a non-uniform density.

Applying OPTICS to the species and regions in the latent space detected 67 clusters for species and 5 clusters for regions 5.6. The dynamics of the clusters are reported in Fig. 5.5 with all the nodes not belonging in any cluster greyed out.

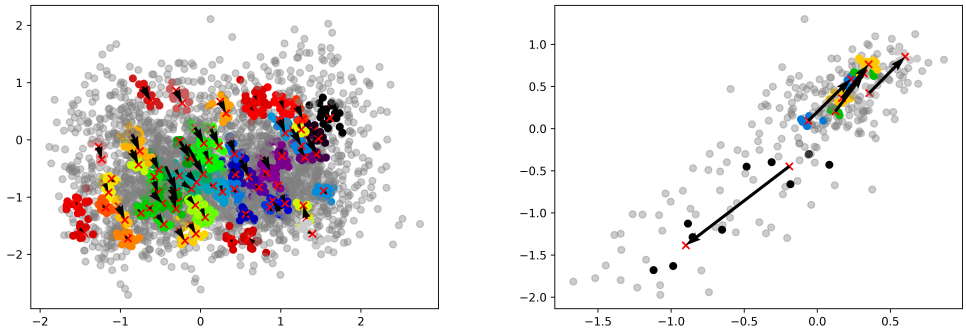


Figure 5.5. Latent space initial and final configurations. The black arrows shows the movement in space from the initial state to the final state.

5.1 Insights from the species clusters

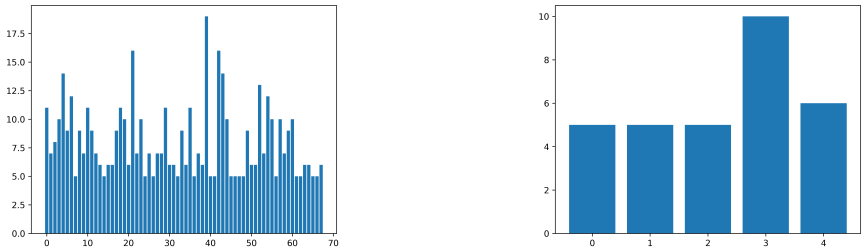


Figure 5.6. Size distribution of the cluster of species (left) and region (right).

We now focus on analysing the taxonomic families of the species belonging to the five largest clusters. Out of the 18 Taxonomic Families, only 9 are present in the five largest clusters. This shows that some taxonomic families are more prominent to co-invade than others. In Table 5.1 we reported the composition of the five largest clusters. If a taxonomic family is not found in a cluster, the entry is left empty. The plants family seems to have a significant relevance, it is present in all the clusters. Birds and Insects seem to have the tendency to belong in the same clusters, meaning that they are likely to co-invade a region.

The dimension of the Taxonomic Families is non-homogeneous. As shown in Fig. 2.3, Vascular Plants and the Insects are the largest Taxonomic Families. This should be taken into account when analysing the composition of clusters: a species belonging to one of the most abundant families will have an higher probability of (indirectly) interacting with other species in the latent space. In fact, plants and insects are generally the largest components of the clusters we detected.

	Cluster 4	Cluster 21	Cluster 39	Cluster 42	Cluster 43
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 5.1. Species clusters composition.

5.2 Insights from the regions clusters

The clusters for the regions in the latent space are also interesting. Applying the OPTICS clustering algorithm to the trajectories of the regions in the latent space detected 5 clusters. They are reported in Table 5.2. For better visualization, the regions are reported respectively with different colors for the cluster they belong to in Fig. 5.7. It's interesting to note how some clusters in the latent space are close in terms of distance such as some of the regions belonging to the clusters number 3 and 4. Regions that are geographically close are expected to be invaded by the same group of species. However, this effect does not apply to all regions. Cluster 1, in particular, contains regions that are geographically not adjacent. The co-invasion relation might be justified due to tight relationships between these states such as trade routes.

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	Central African Republic
			Senegal	
			Tajikistan	
			Gabon	
			Congo	

Table 5.2. Region clusters composition.

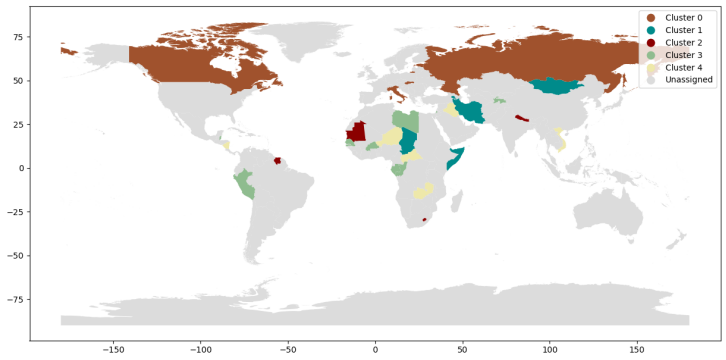


Figure 5.7. The five clusters of regions on the world map.

Chapter 6

Conclusion

The problem of alien species requires real action but without first understanding the in-depth drivers of invasion it's hard to come up with any action at all. In the last decades, relational event models have been successfully used to study the underneath drivers of dynamic social networks. Originally, relational event models were designed for small scale social networks. However, a traditional REM approach might not be always feasible due to the large size and complexity of a dataset and its high computationally cost.

In this thesis, we showed how to dispose a relational event model actors in a 2-dimensional latent space and how to make inferences on their latent coordinates. Our inference approach combined many different tools such as the Expectation-Maximization algorithm, Kalman filters, and smoothers. Our method formulation can be applied to many different use cases and has many advantages over traditional statistical tools such as dimensionality reduction, a compact local configuration history of nodes, and large flexibility. Our study on the alien species co-invasion shows that the method is feasible for a real-world study. However, with the correct amount of modifications, our model can be applied to different problems and is not only limited to the study of a bipartite unidirectional graph as presented in this thesis.

We showed the applicability of our model by analyzing 15947 alien species invasions given by First Records Dataset. The analysis of the results of our model is based on studying clusters of trajectories in the latent space for both species and regions. We detected 67 clusters for the species and 5 for the regions in the latent space. Due to the high number of clusters found in the species, we focused on the 5 largest ones resulting in a non-exhaustive analysis. This approach is acceptable for a first analysis and to show the capabilities of our custom latent space REM, but, for an in-depth study of the co-invasion of species, this is not sufficient. A more exhaustive approach would have been to study the probability of, given a species, or a taxonomic family, finding another species inside all the clusters found in the analysis.

Our approach requires the user to select the amount of noise Σ that our model will use in the latent space process. A large Σ will grant the model more dynamicity. In contrast, a small Σ will make the model more conservative. The choice of this parameter is somehow arbitrary. A user might want to try different values of Σ and analyze the dynamics of the latent space. Then, based on this feedback, update the value Σ by either increasing or decreasing it.

The First Records Dataset is currently the most comprehensive and largest dataset available for first-time records of alien species. However, it is an unevenly populated dataset towards regions such as Europe as a whole and taxonomic groups such as plants and insects. Further-

more, 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. These biases and the unreliability of the dataset affected our results.

Concluding, we showed how a relational event model can be applied to a bipartite dynamic network and extended by including a latent space configuration of the nodes. We then presented to the reader a framework to analyze the results based on a clustering algorithm. The improvements and extension of already existing tools such as relational event models into more advanced and efficient methods capable of performing time-dependent analysis such as the study of alien species' first records will help ecologists to investigate this ecological phenomenon in-depth and to come up with real-world actions.

Bibliography

[1] Carter Butts. A relational event framework for social action. *Sociological Methodology*, 38: 155 – 200, 07 2008. doi: 10.1111/j.1467-9531.2008.00203.x.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977. doi: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.

[3] Hanno Seebens et al. No saturation in the accumulation of alien species worldwide. *Nature Communications*, 2017. doi: DOI:10.1038/ncomms14435.

[4] Hanno Seebens et al. Global rise in emerging alien species results from increased accessibility of new source pools. *Proceedings of the National Academy of Sciences*, 115: 201719429, 02 2018. doi: 10.1073/pnas.1719429115.

[5] O. Floerl, G. J. Inglis, K. Dey, and A. Smith. The importance of transport hubs in stepping-stone invasions. *Journal of Applied Ecology*, 46(1):37–45, 2009. doi: https://doi.org/10.1111/j.1365-2664.2008.01540.x.

[6] P. E. Hulme, P. Pysek, W. Nentwig, and M. Vila. Will threat of biological invasions unite the european union. *Scienc*, 32:40–4, 200.

[7] Ruta Juozaitiene, Hanno Seebens, Guillaume Latombe, Franz Essl, and Ernst C. Wit. Analysing ecological dynamics with relational event models: the case of invasion events. Technical report, 2022.

[8] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960. ISSN 0021-9223. doi: 10.1115/1.3662552. URL https://doi.org/10.1115/1.3662552.

[9] Roger Th. A. J. Leenders, Noshir S. Contractor, and Leslie A. DeChurch. Once upon a time: Understanding team processes as relational event networks. *Organizational Psychology Review*, 6(1):92–115, 2016. doi: 10.1177/2041386615578312.

[10] Alessandro Lomi, Daniele Mascia, Duy Vu, Francesca Pallotti, Guido Conaldi, and Theodore Iwashyna. Quality of care and interhospital collaboration a study of patient transfers in italy. *Medical care*, 52:407–14, 05 2014. doi: 10.1097/MLR.0000000000000107.

[11] Peter S. Maybeck. *Stochastic models, estimation, and control*, volume 141 of *Mathematics in Science and Engineering*. IET, 1979.

[12] Kym Patison, Eric Quintane, Dave Swain, Garry Robins, and Pip Pattison. Time is of the essence: An application of a relational event model for animal social networks. *Behavioral Ecology and Sociobiology*, 69, 05 2015. doi: 10.1007/s00265-015-1883-3.

[13] Hanno Seebens, Tim Blackburn, Ellie Dyer, Piero Genovesi, Philip Hulme, Jonathan Jeschke, Shyama Pagad, Petr Pyšek, Mark van Kleunen, Marten Winter, Michael Ansong, Margarita Arianoutsou, Sven Bacher, Bernd Blasius, Eckehard Brockerhoff, Giuseppe Brundu, César Capinha, Charlotte Causton, Laura Celesti-Grapow, and Franz Essl. Global rise irmn emerging alien species results from increased accessibility of new source pools. *Proceedings of the National Academy of Sciences*, 115:201719429, 02 2018. doi: 10.1073/pnas.1719429115.

[14] Simo Särkkä. *Bayesian filtering and smoothing*. Cambridge University Press. Cambridge University Press, 2013.