POLITECNICO
MILANO 1863

Mathematical Engineering A.Y. 2022-23
Nonparametric Statistics

# HONEY BEE HEALTH

## Evaluation of Factors and Strategies to Mitigate Colony Loss in High-Risk Beekeeping Environments

Final Report
16th February 2023

M. Cerri, L. Mainini, L. Marsigli, E. Varetti

## Contents

# 1 Introduction

> If the bees disappear, humans would have 4 years to live
>
> *Albert Einstein*

**Bees are one of the most ecologically and commercially important insects in the world** (Brown and Paxton, 2009). Bees are critical for human survival, pollinating one-third of the food we eat, including fruits, vegetables, oils, seeds, and nuts. It is estimated that some seed and nut crops would decrease by more than 90% without these pollinators.[1] Because pollination aids in tree regeneration - and consequently in preserving forest biodiversity - bees also help to maintain native ecosystems.

Despite the fact that bees are extremely important, an **urgent worry has been raised in the last 20 years due to the global decline in honey bee numbers**, which has an impact on the quality of life for all human populations that depend (directly or indirectly) on them. **There are significant gaps in our understanding of normal bee behavior and causes of colony collapse**. Intensified agriculture, habitat loss, mites like the Varroa, bee pathogens, pesticides, and climate change are all complex interacting factors in the significant decline in bee populations.

During the winter of 2006-2007, beekeepers began to report unusually high losses of 30-90 percent of their hives. As many as 50 percent of all affected colonies demonstrated symptoms inconsistent with any known causes of honey bee death. In particular, the majority of worker bees left the colonies, leaving behind a queen, plenty of food and a few nurse bees to care for the remaining immature bees. This unprecedented phenomenon was called "Colony Collapse Disorder". Numerous causes for CCD have been proposed, but no single stressor alone seems to be responsible for the malady.[2] **The severity of CCD has brought greater attention to bees, fostering scientific research and leading U.S. public authorities to start collecting data**.

This data allows us to draw some general conditions on bee health. Our project addresses public opinion and government authorities with the aim of raising awareness of the issue. **Our first objective is to statistically test whether certain factors correlate with colony losses** and describe this relation with **a model**. This analysis will be conducted in the section 3 and 4.

In our opinion, **the problem is still largely ignored despite the scale of the phenomenon**. That's why, analysis will be conducted in section 5 to show the extent of the phenomenon, both in terms of affordability for beekeepers and in terms of economic impact.

Considering risks related to climate change and the scale of the problem, governments and beekeepers should adopt strategies aimed at building resilience and ensure a good risk management. That's why we provide some **data insights**

**to better understand which environments are the most favorable** that allow for the greatest profiling / least hive decline. This analysis considers atmospheric (pressure, temperature...) and pollen-related phenomena.

# 2 Data description

To provide more insightful analysis, we merged different data set:

- factors which stress directly bee colonies as Varroa Mite or pesticides (source: USDA)
- other possible influential factors, in particular Temperature, Drought, Precipitation (source: National Centers for Environmental Information)
- annual production and price of the honey for each state (USDA)

The first one was collected by the United States Department of Agriculture through a survey. For each quarter, they provided data on honey bee colonies in terms of total colonies, maximum colonies during the period, colonies lost, percent lost, colonies added, renovated, and renovated as a percentage, and colonies lost due to symptoms of Colony Collapse Disorder. In addition, the report lists stressors to colony health (when there are five or more colonies affected). If no colonies are moved out of state the following formula should hold in theory:

$$n_0(t,s) \; - \; \text{Lost n}(t,s) + \; \text{Added n}(t,s) \; = n_0(t+1,s) \tag{1}$$

At first glance, the data seemed to have some inconsistencies in the sense that even taking into account the colonies lost and those added, the number of colonies at the previous semester was not consistent with that of the following. For more information, look at the Notebook *Data_Consistency.ipynb* As Seth Riggins, Bee & Honey Statistician at the National Agricultural Statistics Service center, confirmed to us by email, this is due to two factors:

- If no colonies are moved out of state, then the calculation above should be within 2% of the starting inventory of the following quarter. These small differences are mainly due to recall errors or lack of precision in filling out the survey.
- Pollinating colonies are moved from state to state throughout the year following the crops that need to be pollinated. Therefore, the maximum number of colonies in a state can vary widely during a quarter and throughout the year.

At national level, it is possible to check that

$$\text{Max n}(t,s) - \; \text{Lost n}(t,s) + \; \text{Added n}(t,s) \; = \; \pm 2\% \; n_0(t+1,s) \tag{2}$$

# 3 Analysis of influencing factors

## 3.1 General trends: higher losses in winter

Losses recorded in the United States vary from year to period according to the data we have. In general, losses seem to have decreased slightly in recent years, but they still remain at worrisome levels. With respect to seasonality, the loss is generally higher in the winter period. We performed a local permutational test on difference in colony lost percentage between summer and winter and all years but 2020 (and 2019 which has missing values in Q2) presented a significant difference (at the level of 5%). A Global Permutational test for functional data

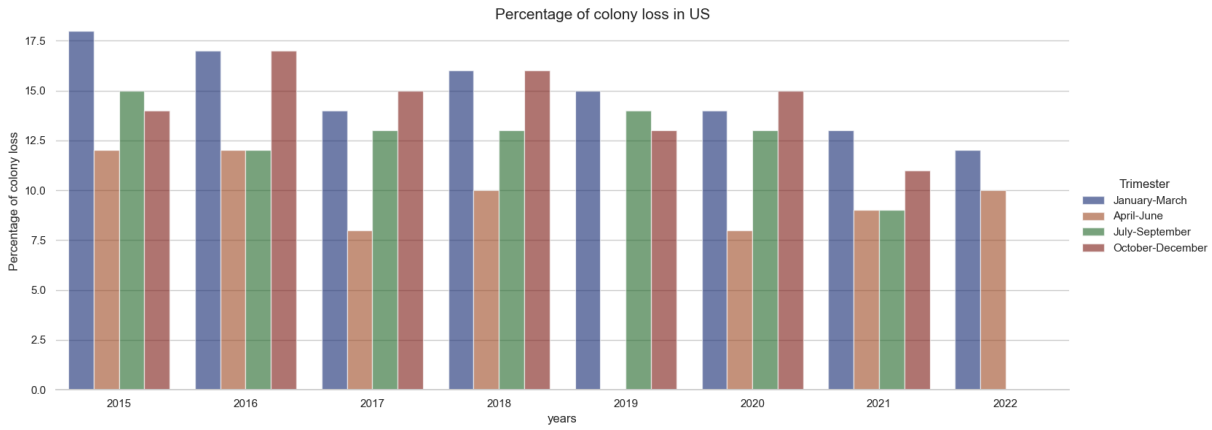



Figure 1: Losses are generally higher during winter than summer

confirms this result.

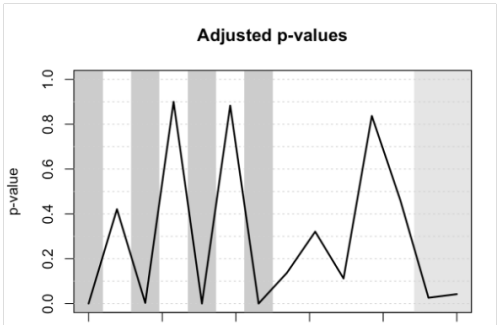



Figure 2: Permutational test for difference in colony lost % between summer (functional data Q2-Q3) and winter (funtional data Q4-Q1)

According to our first analysis, winter is the season where bees are most at risk. However, bees are known to be able to withstand the cold. Indeed, as soon as the temperature drops to below 10°C, bees come together to form what is known as the "winter cluster" (allowing them to stay at 25-30°C). Thus, the reason why the bees are at risk cannot be only the cold weather. Nevertheless, it is true that winter is the time when the colony is most at risk: the fact that they

have to live very close together without being able to get out too much creates a particularly favorable situation for the spread of viruses and parasites. Some of these virus are related to Varroosis. If the infestation by Varroa Mites is high at the beginning of winter and nothing is done to decrease it, bee chances of survival can dramatically drop. Also, bees must rely only on supplies gathered in the previous months. If, for any reasons, these are not enough, the colony might be at serious risk.

## 3.2 Influence of the weather

In this section, we will investigate possible relationships between temperature and colony dispersion. In particular, since we previously saw a greater loss during winter, we divide the states according to minimum temperatures. A

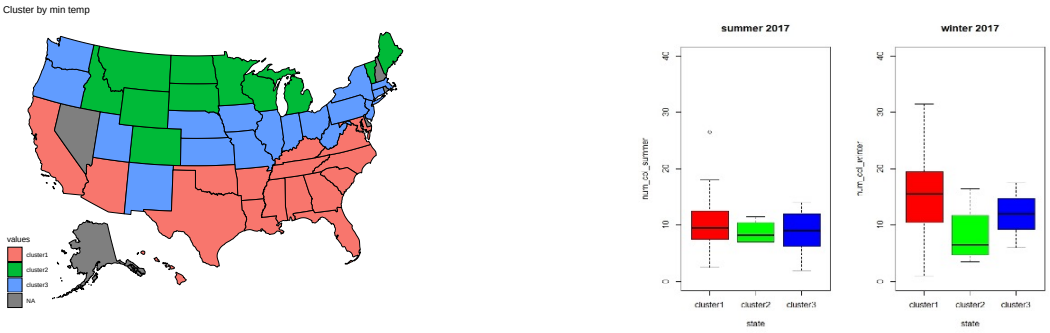



Figure 3: average losses in the 3 groups (obtained with a functional clustering by minimum temperature) during 2017's summer and winter

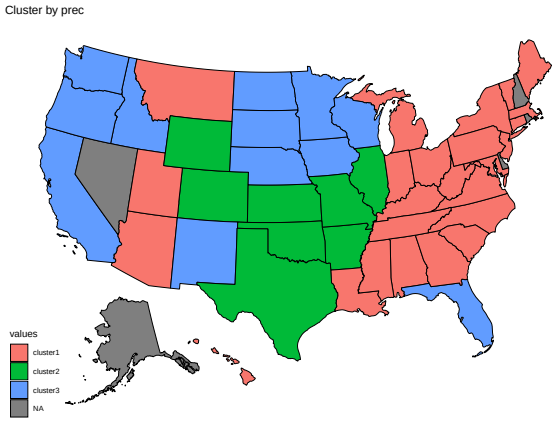



Figure 4: Functional clustering by precipitation

permutational one-way ANOVA shows that colony losses are different in these groups obtained clustering by minimum temperature for all the fall/winters. In particular, Subtropical/Mediterranean and Continental regions show higher bee diminution than Alpine [Figure 3].

This may be related to different phenomena or correlated factors. In particular, Mattila showed that during late autumn, cold spells with adequate duration and intensity trigger colony wintering at the right time, reducing the likelihood of colony loss over wintern [3]. Finally, using the same methodology, we can show that losses are higher in states with lower precipitation.

## 3.3 Main colony stressors: pesticides, diseases and Varroa Mites

This session will investigate possible relationships between some stressors and colony disappearance. At national level, colony loss during summer is strongly correlated with Pesticides and Diseases (look at Figure 1).

| | Varroa Mites | diseases | pesticides | other pests, paras. | other | unknown |
|---|---|---|---|---|---|---|
| loss | 0.3 | 0.55 | 0.71 | 0.17 | -0.0088 | 0.14 |
| loss % | 0.41 | 0.54 | 0.66 | 0.27 | 0.18 | 0.21 |

Table 1: Spearman correlation on national data - Summer US

The adoption of pesticides thus seems to lead to a paradox: when used properly, they have an obvious economic advantage, but they may also put bee health at risk, reducing then indirectly yields of fields. A recent study by Stuligross and colleagues suggest the harm of pesticides can accumulate over multiple generations, which could exacerbate the loss of species that provide valuable pollination for farms and ecosystems. [4]

During winter, all stressors are weakly correlated with losses (similar Spearman's coefficients in the range 11%-17%). It is also interesting to point out that in this season, stressors are strongly correlated with each other.

| | Varroa Mites | diseases | pesticides | other pests, paras. | other | unknown |
|---|---|---|---|---|---|---|
| loss % | 0.17 | 0.17 | 0.14 | 0.13 | 0.19 | 0.11 |

Table 2: Spearman correlation on national data - Winter US

Considering Spearman correlations on functional data over all years by state, we can see the impact of Varroa and Pesticides. These coefficients also show how still in so many cases the stressors are not really known.

## 3.4 On Varroa Mites

Varroa Mites is, according to our data, the stressor which affects the higher number of colonies. This is consistent with the literature, according to which Varroa mite is the world's most devastating honey bee pest. It is a highly efficient vector of honey bee viruses and drives changes in virus distribution,

| | Loss % | Varroa | Other P. | Disease | Pestic. | Other | Unknown |
|---|---|---|---|---|---|---|---|
| **Loss %** | 1.00 | 0.53 | 0.41 | 0.37 | 0.54 | 0.66 | 0.74 |
| **Varroa** | 0.53 | 1.00 | 0.68 | 0.50 | 0.57 | 0.46 | 0.37 |
| **Other P.** | 0.41 | 0.68 | 1.00 | 0.26 | 0.45 | 0.51 | 0.60 |
| **Disease** | 0.37 | 0.50 | 0.26 | 1.00 | 0.79 | 0.68 | 0.46 |
| **Pesticides** | 0.54 | 0.57 | 0.45 | 0.79 | 1.00 | 0.70 | 0.67 |
| **Other** | 0.66 | 0.46 | 0.51 | 0.68 | 0.70 | 1.00 | 0.81 |
| **Unknown** | 0.74 | 0.37 | 0.60 | 0.46 | 0.67 | 0.81 | 1.00 |

Table 3: Spearman correlation matrix based on functional data of states (Ordering with Modified Hypograph Index)

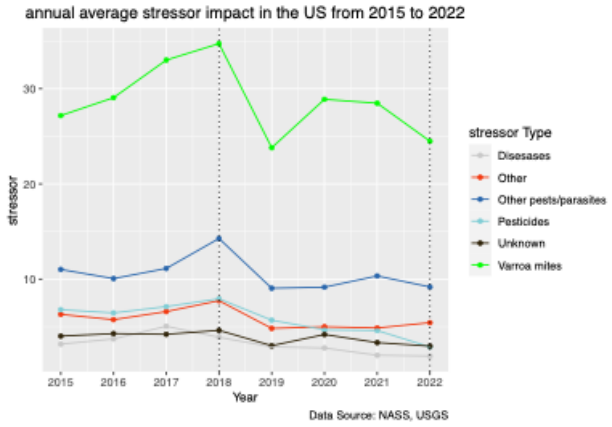



Figure 5: Stressors' prevalence

prevalence, and virulence. Heavy Varroa mite infestations can build up in 3–4 years and cause scattered brood, crippled and crawling honey bees, impaired flight performance, a lower rate of return to the colony after foraging, a reduced lifespan and a significantly reduced weight of worker bees.

According to researchers, the highest infestation is in late fall, as confirmed by our data. The life cycle of Varroa is split into two distinct phases: the reproductive phase that takes place inside honey bee brood cells and dispersal phase in which mature female mites travel and feed on adult bee. By early spring the colony commences rearing drone brood (dark brown), preferentially invaded by varroa (red mites). After swarm season, bees cease rearing drones, forcing varroa to reproduce in worker brood. Pick of varroa is in fall. As the colony stops rearing brood, varroa has no place to reproduce and their population sinks. [5]

We noticed that some states show similar patterns in overwintering reductions of bee colonies and number of stressed colonies by Varroa mites. This relation is often clearer shifting the Varroa's time series. An exception is the New Mexico (which however is a shape outlier) where this pattern is clear even without shifting [Figure 7].

We conclude this section showing that Varroa's diffusion seems influenced by temperature. If we consider 3 groups of states based on average temperature

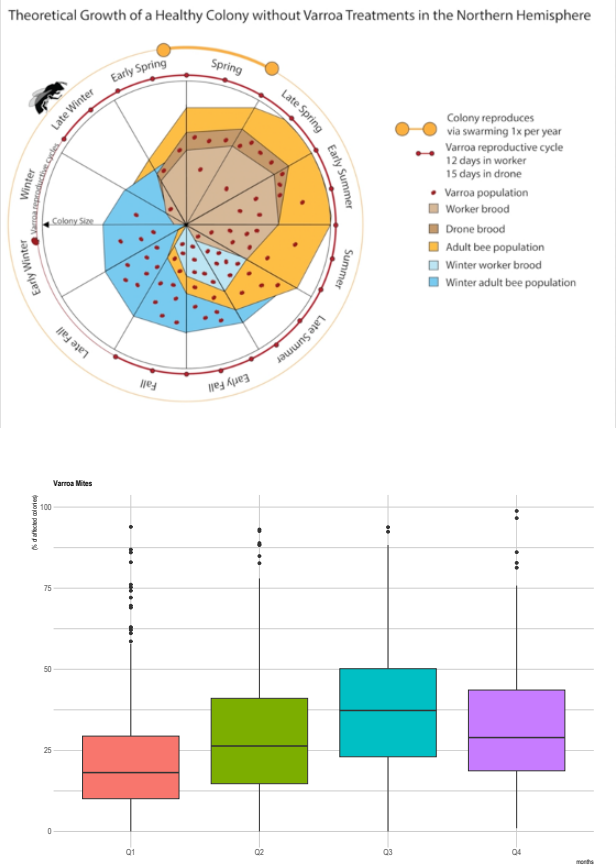



Figure 6: Number of colonies impacted by Varroa Mites

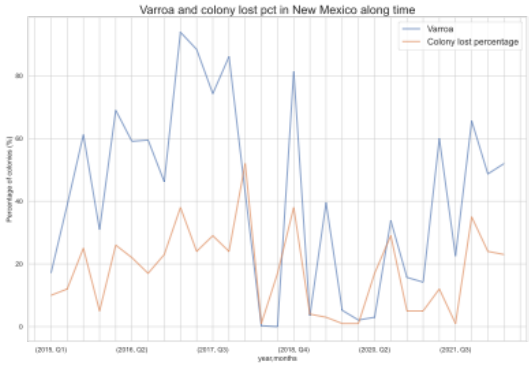



Figure 7: New Mexico

[Figure 8], Varroa mites strike the most warmer countries and the phenomenon seems to worsen in the last 4 years under study. This could be due to rising temperatures that we are experiencing in last years.

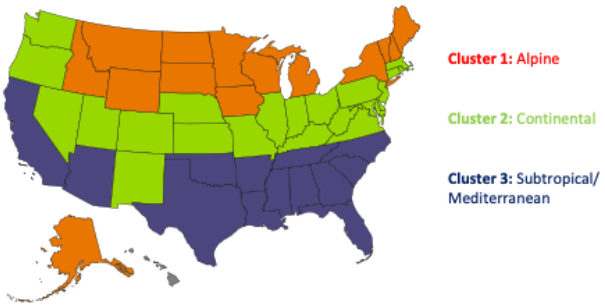



Figure 8: Functional clustering by average temperature

# 4 A Model of loss-stressors relation

This section will introduce model capable of accurately representing the relationship between stressors and loss of colonies. The statistical analysis of our data is complicated for several reasons:

1. There may be both temporal and spatial correlation
2. The data contains missing values in the time series (difficult to assess the right correlation structure)
3. The data may be incorrect, because were counted by different apiarists collected in different states.
4. there may be heterogeneity over time (more variation in winter than in summer)
5. trends over time and in space may be non-linear

Given that we are tallying the number of colonies lost from the maximum number present in a semester within each particular state, we can utilize a Binomial distribution. Additionally, our observations are not independent due to the inclusion of repeated observations both over time and in space within the dataset. Therefore, to effectively model our data, we have opted to employ a generalized additive model (GAM) with a binomial link.

$$n_i \sim Bin(\text{Max } n_i(t), \pi_i) \tag{3}$$

where

$$\begin{aligned} E[n_i] &= \text{Max } n_i(t) \times \pi_i \\ Var(n_i) &= \text{Max } n_i(t) \times \pi_i \times (1 - \pi_i) \end{aligned} \tag{4}$$

$$logit(\pi_{i,t}) = f(Stressors_{i,t}) + f(MinTemp_{i,t}) + f(space, time) + \epsilon_{i,t} \tag{5}$$

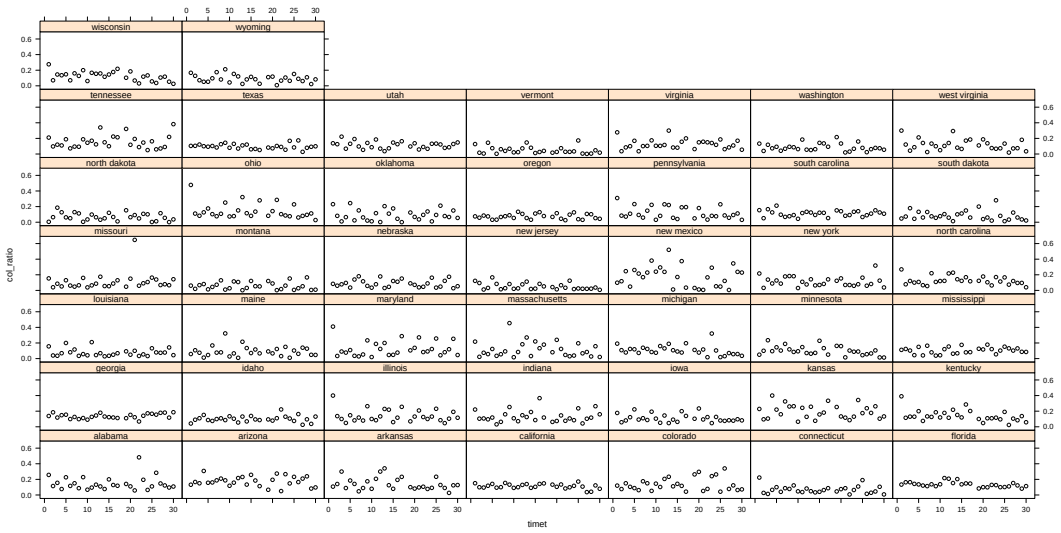



Figure 9: US states show different range of variability

## Generalize Additive Model

We started from the reasonable assumption (looking at the previous results) that colony losses depend on stressors and on time:

$$\pi_i(t) = \alpha_i + f\left(\text{stressors}_i(t)\right) + f_i(t) + \epsilon_i(t) \tag{6}$$

where $\pi_{i,t}$ is the value of time series relative to the state $i = 1, ..., 40$ in time $t = 1, ..., 30$, corresponding to our 4 trimesters repeated over the years (from 2015 to 2022) we are considering, where $f_i(t_i)$ is a smoother for time in each state. It is to say that our hope were to remove these index in order to have a global interpretation of the effect of Varroa on the colonies. Indeed, if we remove the index $i$ we assume the same smoother for each state, hence the time series are assumed to follow the same trend.

As we can notice in Figure 9, states have different range of variability, hence it could means that our model should account for heterogeneity that we model through a smoothing on the space. We firstly tried to use GAMM (Generalized Additive Mixed Models) as suggested in [6]; unfortunately, for numeric reasons, we were not able to follow this path and finalize the discussion. For this reason we decide to include a tensor product of space and time, choosing respectively a thin-plate spline basis for the space location and a ciclic-cubic spline for the time (we did this choice because the fourier-basis is not allowed in mgcv).

Looking at the estimated smoothing curves, we can conferm that colony lost increases when Varroa increases. On the other hand, we observe that colony loss also increases with the increase in minimum temperature. As confirmed in the literature, bees are not affected by low temperatures, but rather by high minimum temperatures. Other stressors, like Diseases and Pesticides are statistically relevant in explaining colony loss, but unfortunately their interpretation is not clear.

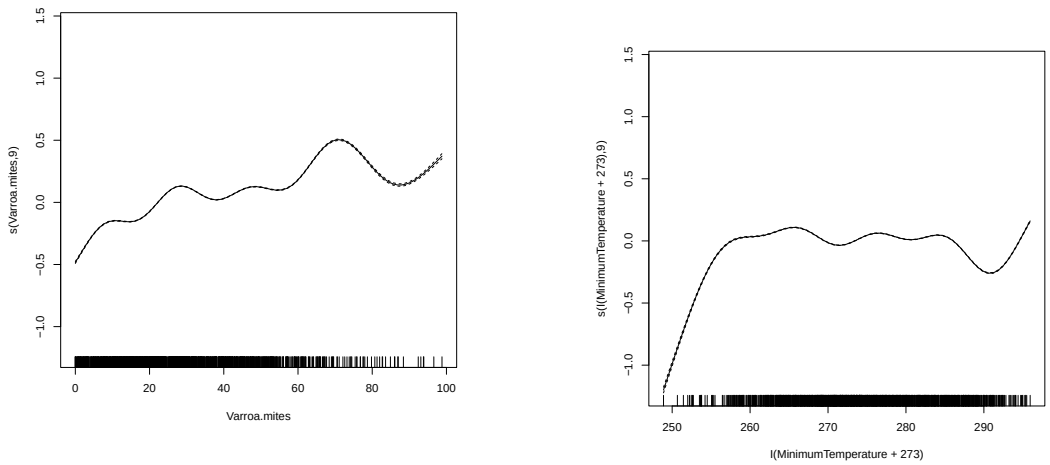



Figure 10: Estimated smoothing curves for Varroa Mites and Minimum Temperature. The solid line is the smoother and the dotted lines are 95% point-wise confidence bands

We then decided to include a non-linear smoother on space

# 5 Study of the impact of colony losses

## 5.1 Impact on beekeepers under a survival approach

The purpose of this section is to understand whether the current losses are sustainable for beekeepers, in other words if they are able to cope with this problem on their own in time or if a governmental action is needed.

To do this, we introduced a new metric that we called cumulative "after action" loss. It is a relative loss that takes into consideration the colony added by beekeepers to cope with colony losses. We will use AAL to keep track of the actual decreasing of the population, despite the intervention of beekeepers. We believe that the situation may be considered worrying if this value exceeds 20%. This metric is defined as:

$$AAL_t = AAL_{t-1} + \frac{\text{Lost n}_t - \text{Added n}_t}{\text{Max n}_t} \tag{7}$$

This metric allows us to define a time to death event as the first time $t$ when $AAL_t > 20\%$. To have a more robust information we check if also in the following quarter the population is still decreasing. Now we can study the problem by performing a survival analysis.

We notice that already after the first winter a lot of states reaches the threshold and we want to understand if they have a characteristic in common. We realized that most of these states belong to the cluster characterized by harsh winters

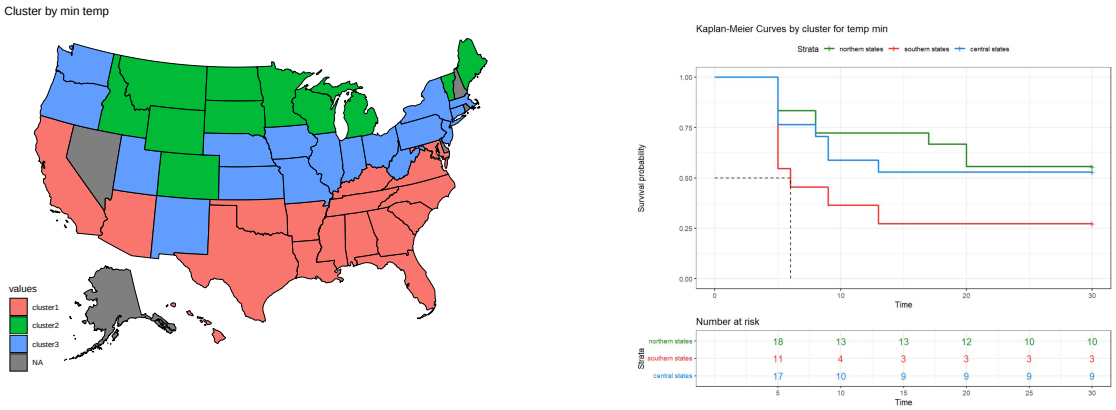



Figure 11: Kaplan-Maier Curves by minimum-temperature based groups

and indeed there is some differences in the curves between northern states and southern states.

The Kaplan-Meier curves describe the survival according just to the minimum temperature and ignoring the impact of any other predictors. So, to conclude our analysis, we perform a cox regression model to describe how different factors jointly impact on survival. For each state i we have the following information:

$$\boldsymbol{x}_i = [\text{avg temp}_i, \text{prec}_i, \text{pdsi}_i, c_i]$$

where $c_i$ indicates the group they belong (minimum-temperature-based cluster).

We consider the following Cox PH model:

$$h_i(t|\boldsymbol{x}_i) = h_0(t) \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right)$$

where $\boldsymbol{h_0(t)}$ is an unspecified non-negative function of time called baseline hazard and $\boldsymbol{\beta}$ is the vector of coefficients that we want to estimate.

Table 4: Summary of COX Model

| Variable | $\beta$ | Exp($\beta$) | SE($\beta$ ) | P-value | lower | upper |
|---|---|---|---|---|---|---|
| Average Temp | -0.27095 | 0.76265 | 0.06068 | 7.99e-06 | 0.67 | 0.85 |
| Precipitation | -0.52042 | 0.59427 | 0.18169 | 0.00418 | 0.41 | 0.84 |
| PDSI | 0.21544 | 1.24040 | 0.12151 | 0.07624 | 0.97 | 1.57 |
| cluster2 (min temp) | -3.51243 | 0.02982 | 1.10189 | 0.00143 | 0.003 | 0.25 |
| cluster3 (min temp) | -2.23167 | 0.10735 | 0.78771 | 0.00461 | 0.022 | 0.50 |

**Note.** SE denotes Standard Error

The p-values for the three overall tests (likelihood, Wald, and score) are extremely small, indicating that the model is significant. Moreover, all the covariates are significant with a confidence level of 90%. The Hazard Ratio for both the Average Temperature and Precipitation are less than one and their 95% and confidence intervals do not contain 1, indicating a strong relationship

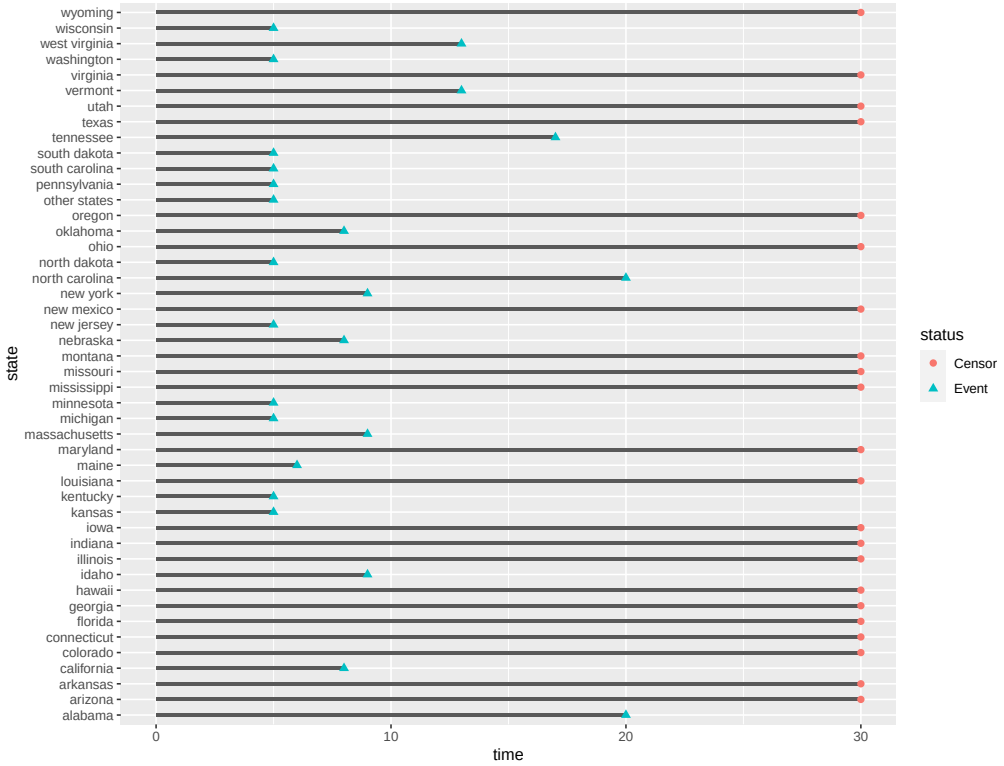
Table 5: Test Results

| **Test** | **Statistic** | **df** | **p-value** |
|---|---|---|---|
| Likelihood Ratio | 50.51 | 5 | $1 \times 10^{-9}$ |
| Wald | 29.06 | 5 | $2 \times 10^{-5}$ |
| Score (logrank) | 50.72 | 5 | $1 \times 10^{-9}$ |

between these covariates and decreased risk of death. The hazard ratio of PDSI is greater than 1 and the 95% CI is [0.977532; 1.5740]. Because the confidence interval for HR includes 1, PDSI makes a smaller contribution to the difference in the HR after adjusting for the other covariates.

We interpret the result as following: as expected states that belong to cluster2 (meaning being a state with harsh winter) have higher survival probability than states belonging to cluster1(meaning being a state with warmer winter). Moreover, holding the other covariates constant, a higher value for Precipitation is associated with a better survival. The higher the precipitation the lower the loss, and the same goes for Average Temperature. So, Average Temperature and Precipitation are good prognostic factor.

We can visualize the hazard ratio and its 95% confidence interval in the following plot:



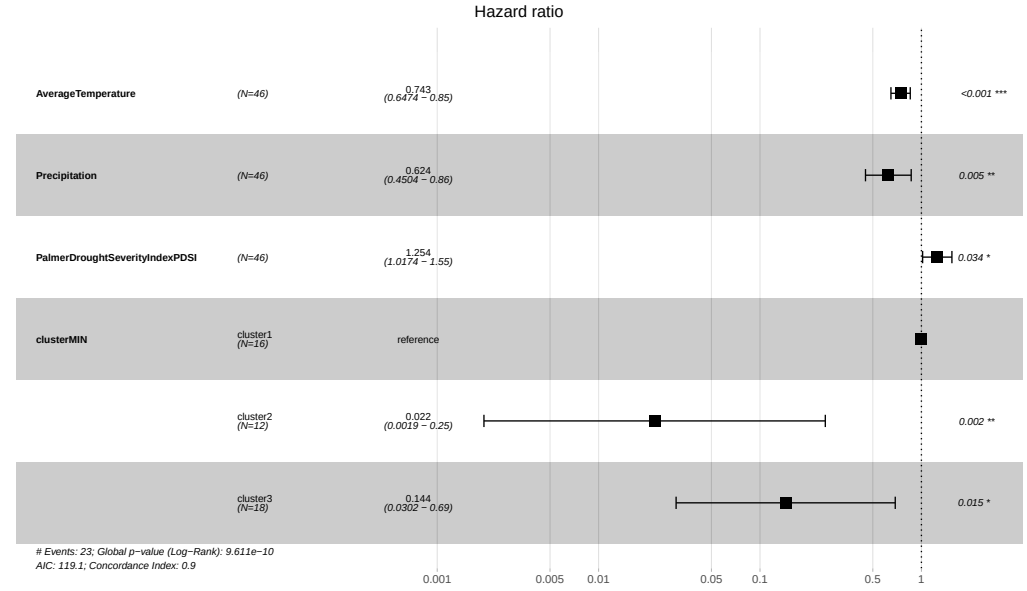



Figure 12: Forest Plot

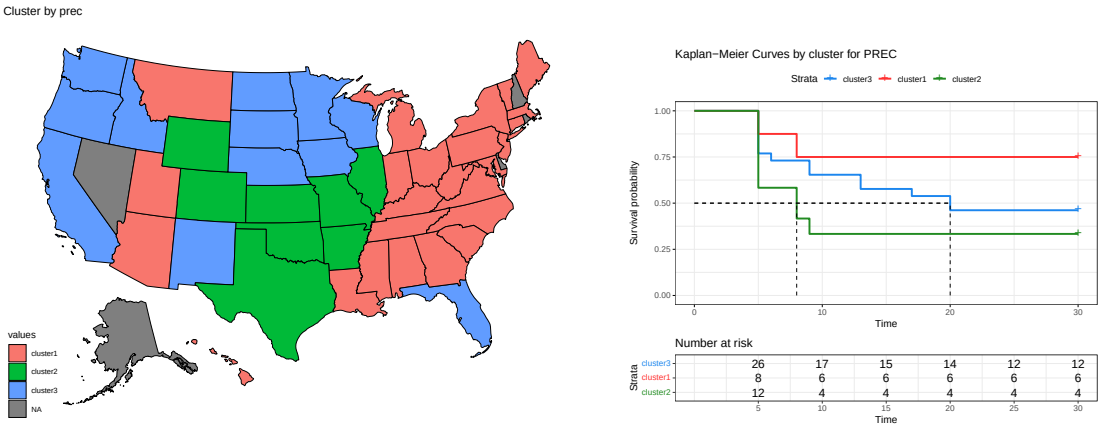



Figure 13: Kaplan-Maier Curves by precipitation-based groups

## 5.2 Economic Quantification

To make an assessment on the economic impact of bee colony losses, we want to try another model that allows a visualization of how much money is lost over time due to the death of colonies.

Since we are dealing with spatio-temporal data, we adopt a method (suggested in [7] and [8]) for the analysis of functional data with multiple dependencies over complex geometries. The models are based on the idea of regression with partial differential regularizations: we consider here two roughness penalties that account separately for the regularity of the field in space and in time.

**Assumptions**:

- Simplification on the nature of the data: we consider them in the framework of geostatistical functional data instead of the one of functional areal data. We assume the datum is observable in principle in any point of the domain, so our areal data collected on the surface of each state in US are assigned to the centroid of the state.

- Functional data are available continuously over time, even if we have available data in discrete time interval. With this assumption, we assign each data in time to the beginning of the time interval it refers to.

For the purpose, we consider a penalized regression model for spatial functional data, to **estimate the amount of money lost due to the loss of colonies**. This loss refers to the monetary gain it could be achieved selling the honey that would have been produced by the colonies lost.

As showed in the data description, we use here also the information on the price of honey and on the efficiency of colonies in terms of honey production.

### 5.2.1 Penalized Semiparametric Regression Model for Spatial Functional Data

Let $\{\mathbf{p}_i = (x_i, y_i)\,; i = 1, \ldots, n\}$ be a set of 44 spatial points on the US nation domain $\Omega \subset \mathbb{R}^2$. Each spatial location is the centroid of a given state (expressed with longitude and latitude), excluding Hawaii, Alaska and few others were the data was missing. Let $\{t_j; j = 1, \ldots, 29\}$ be a set of 29 quarters in the time interval $[T_1, T_2] \subset \mathbb{R}$, from 2015-Q1 to 2022-Q2 and excluding 2019-Q2.

Let $z_{ij}$ be the amount of money loss in Kdollars due to the presence of 100 colonies in the state $i$ at quarter $t_j$. The data $z_{ij}$ can be seen as a sampling of time dependent surfaces on $\Omega$.

Moreover, $\mathbf{w}_{ij}$ is a vector of 2 covariates, stressors varroa and Pesticides, associated to the observation $z_{ij}$, at location $\mathbf{p}_i$ and time instant $t_j$, and $\boldsymbol{\beta}$ is a vector of 2 regression coefficients.

We assume that $\{z_{ij}; i = 1, \ldots, n; j = 1, \ldots, m\}$ are noisy observations of an underlying spatio-temporal smooth function $f(\mathbf{p}, t)$ added to the linear combination of stressors covariates.

The semiparametric generalized additive model, including space-time varying covariates, is the following:

$$z_{ij} = \mathbf{w}_{ij}^T \boldsymbol{\beta} + f\left(\mathbf{p}_i, t_j\right) + \epsilon_{ij} \quad i = 1, \ldots, 44, j = 1, \ldots, 29, \tag{8}$$

where $\{\epsilon_{ij}; i = 1, \ldots, 44; j = 1, \ldots, 29\}$ are independently distributed residuals with mean zero and constant variance $\sigma^2$. We work here conditionally on covariates $\mathbf{w}_{ij}, \mathbf{p}_i$ and $t_j$.

We can jointly estimate the vector of regression coefficient $\boldsymbol{\beta}$ and the spatio-temporal field $f(\mathbf{p}, t)$ by minimizing a penalized sum of square error functional $J(f, \boldsymbol{\beta})$, where the penalization takes into account separately the regularity of the function in the spatial and temporal domains.

$$\begin{aligned} J(f, \boldsymbol{\beta}) = & \sum_{i=1}^{44} \sum_{j=1}^{29} \left( z_{ij} - \mathbf{w}_{ij}^T \boldsymbol{\beta} - f\left(\mathbf{p}_i, t_j\right) \right)^2 \\ & + \lambda_S \int_{T_1}^{T_2} \int_{\Omega} (\Delta f(\mathbf{p}, t))^2 d\mathbf{p} dt \\ & + \lambda_T \int_{\Omega} \int_{T_1}^{T_2} \left( \frac{\partial^2 f(\mathbf{p}, t)}{\partial t^2} \right)^2 dt d\mathbf{p}. \end{aligned}$$

Where $\lambda_S > 0$ and $\lambda_T > 0$ are the two smoothing parameters that weight the penalizations respectively in space and time.

Note that we are here considering an isotropic smoothing. If we could profit of more detailed problem-specific information, we could model the spatio-temporal behavior of the phenomenon using a time-dependent PDE, and perform a more accurate anisotropic smoothing.
Moreover, in this smoother, the roughness penalty in space is integrated only over the region of interest thanks to a finite element formulation (see section 5.2.1).

On the other side, in the case of tensor products, the roughness penalty is integrated over the rectangular multidimensional domain of variables.

### Basis system in space-time

We represent the spatio-temporal field $f(\mathbf{p}, t)$ as an expansion on a separable space-time basis system.

In this case, we use in space a finite element basis on a triangulation $\mathrm{X_s}$ of the spatial domain $\Omega$ of interest. In figure 14 is reported the triangulation we generated for the spatial domain of the problem. This allows to handle efficiently data distributed over irregularly shaped domains, that is crucial when the shape of the domain influences the phenomenon under study, as in many applied problems.

The linear finite element basis is composed by globally continuous functions that coincide with a polynomial of degree 1 on each element of the domain triangulation (piecewise linear functions).

For the temporal dimension, we use a cubic B-spline basis with penalization of the second derivative, with knots coinciding with the sampling time instants in the data (each quarter). Finally, the values of the smoothing parameters $\lambda_S$ and $\lambda_T$ are chosen via Generalized Cross-Validation (GCV).

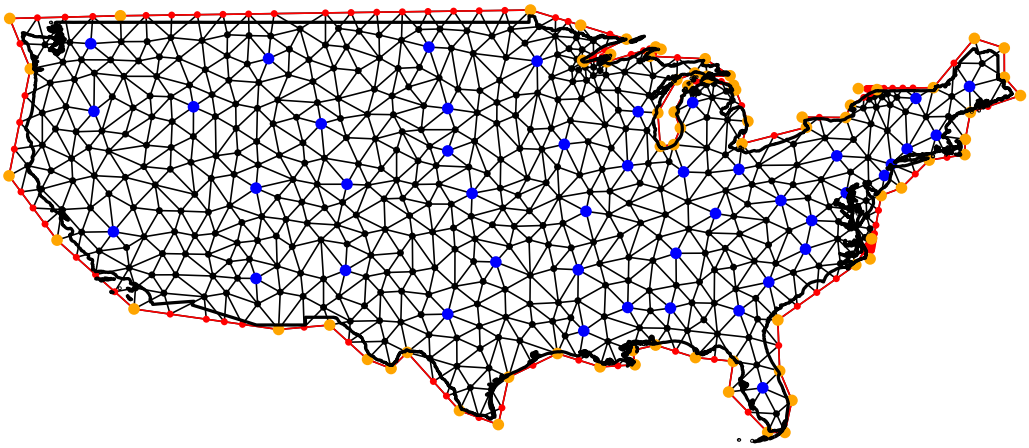


Figure 14: Triangulation of US domain

### Model Choice

The choice of this model is motivated in the literature ([7], [9] and [8]), where in many simulation studies and real cases, the approach presented seems to outperform other space-time models based on differential regularization with tensor-product approach. For instance, wavelet-based smoothing, tensor product splines and thin plate splines are not appropriate for these data, since they do not take into account the shape of the domain and also smooth across concave boundary regions;

All these available techniques naturally work over rectangular or tensorized domain, but in our case we thought that the shape of US domain could influence the spatio-temporal money loss field, and hence must be explicitly considered during the estimation process. That's why we adopted this penalized regression model for spatial functional data, such that the non-tensor product basis allows to handle efficiently data distributed over complex domains where the shape influences the phenomenon's behavior.

### 5.2.2 Results

The resulting model trained on the data seems to provide satisfying results. Fixing a time instant, we can plot how our variable of interest is distributed over US. For instance, in figure 15 we reported a 3dimensional representation of the estimated field at the first quarter of 2016. It's visible from there that the surface seems to be reasonable compared to the observed points which are represented in black. Indeed, if we evaluate the model computing the RMSE through k-fold cross validation in space, with k=4, we get a value of 5.52. This corresponds to a cv-error of around 5.5 kDollars while the observed money loss values range in an interval of around [0, 35]. Therefore, we get an error of 15.7%, which suggests of possible improvements that could be achieved in future steps

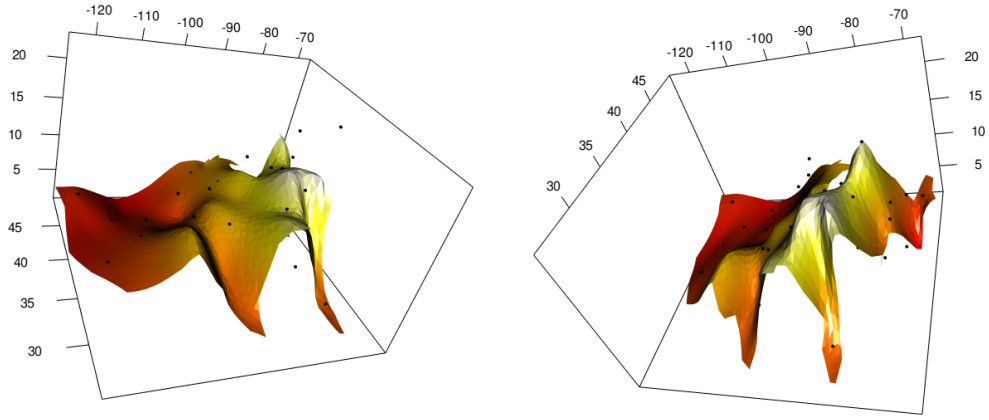



Figure 15: Estimated spatio-temporal field for the money lost (in Kdollars) every 100 colonies during the first quarter of 2016

but still is promising.

Nevertheless, we can exploit the results to have an overview on how relevant is the death of bee colonies on the economy of US. Checking the figure 16, we can see how the economic impact changes over time and how is important for the government to make some interventions to reduce this phenomenon due to the huge economic losses it's generating.

Moreover, it's important to highlight the difference between the two main regression models we employed for the project. We are not making a comparison between the two, but we are simply employing both of them for different analyses:

- with GAM model in section 4 we studied the impact of environments next to the one of stressing and ecological factors on the loss of colonies;
- with the penalized spatial functional regression model we want to provide to our stakeholder a reliable visualization and quantification of the impact that this phenomenon has on US economy; This visualization takes into account the complexity and shape of the spatial domain, which is discarded from models relying on tensorized domains, but doesn't allow to account for the nonlinear effect of stressors.

### Limitation of the model and possible improvements

- The main approximation done in the model was to assign each spatial datum to the state centroid. This assumption makes sense according to [7], but still a better approach would be to account for both areal observations in space and interval observations in time.

  The following model, proposed in [8], would theoretically fits better to the data we have available:

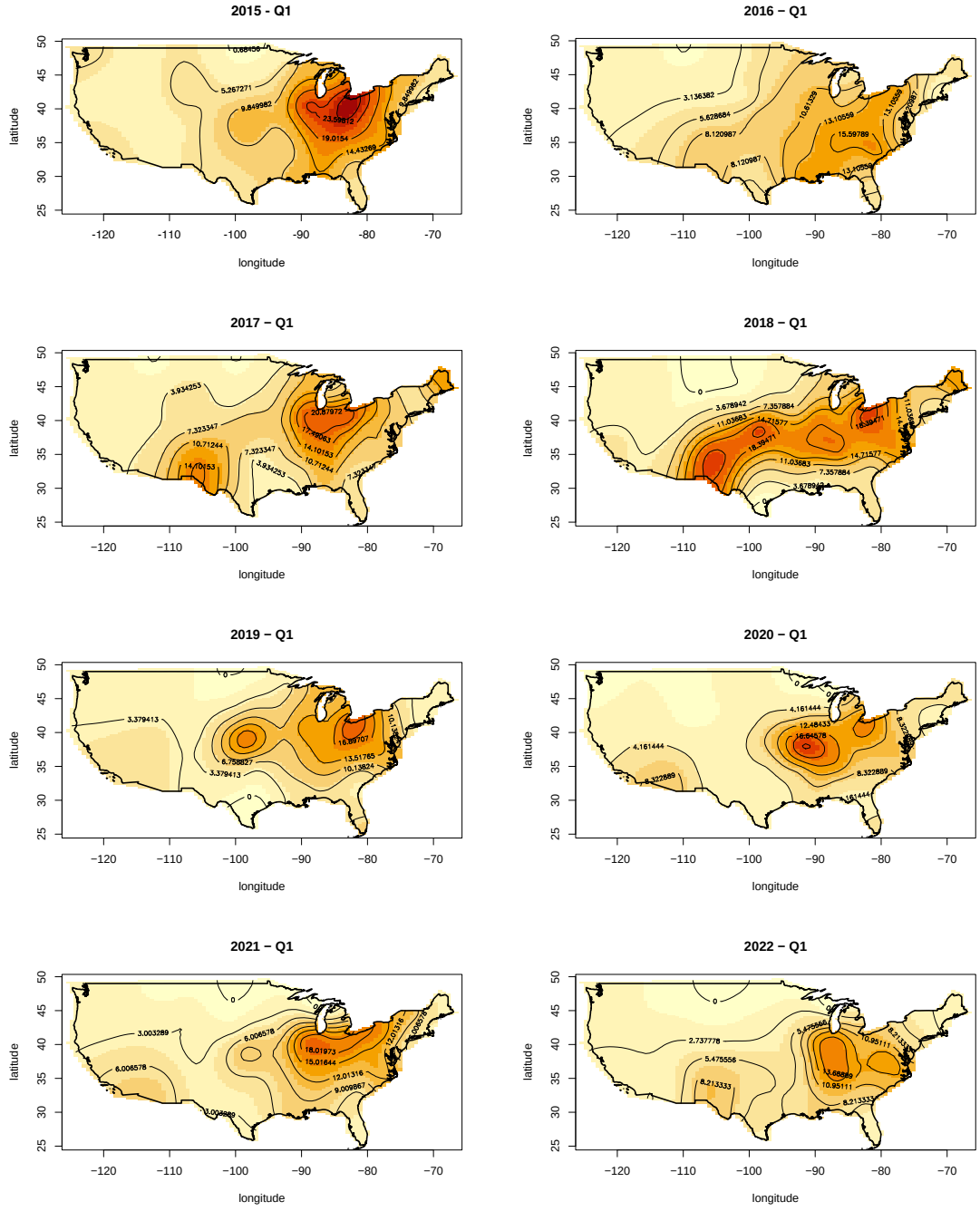



Figure 16: Amount of money lost in Kdollars, in the given quarter, every 100 colonies of honeybees

Let $D_1, \ldots, D_n \subset \Omega$ be 44 disjoint spatial subdomains, and let $T_1, \ldots, T_m \subset [T_1, T]$ be 28 disjoint temporal intervals. Assume that for all $i \in \{1, \ldots, 44\}$ and $j \in \{1, \ldots, 28\}$,

$$z_{ij} = \mathbf{w}_{ij}^T \boldsymbol{\beta} + \frac{1}{|D_i|\,|T_j|} \int_{T_j} \int_{D_i} f_0(\mathbf{p}, t) d\mathbf{p} dt + \varepsilon_{ij},$$

where the errors $\varepsilon_{ij}$ are independent, with zero mean, and variance proportional to $1/\left(|D_i|\,|T_j|\right)$.

- In our case we use the latitude and longitude as spatial covariates, but further improvements could be achieved employing the UTM coordinate system, which allows to compute the distance between two points on the Earth's surface by means of the Euclidean distance instead of the geodesic distance.

- According to [10], the model could be improved including a priori information available on the phenomenon under study, to account for non-stationarities and anisotropies in space and/or time. This can be achieved using a more complex differential regularizations, modelling the space and time behavior of the phenomenon using a time-dependent PDE.

### Hypothesis testing on $\beta$ coefficients

We expressed the model (8) using only 2 stressors in the parametric part, which were selected using some testing procedures to check whether linear components given by the different stressors have an effect on the target $z_{ij}$ given by the money loss.

We are thus interested in the system of hypotheses

$$\mathrm{H}_0 : \boldsymbol{\beta} = 0 \quad \text{versus} \quad \mathrm{H}_1 : \boldsymbol{\beta} \neq 0 \tag{9}$$

The inference analysis for estimator and testing were performed in [11] in the framework of Spatial Regression with differential regularization, but one of tests, eigen sign-flip, could be extended for spatio-temporal regression models.

Focusing first on a **penalized spatial regression setting**, it is shown that the estimator of $\boldsymbol{\beta}$ is asymptotically normal, such that Wald type inference could be performed. However, the empirical variance of the estimator is biased due to the presence of the regularization. This may result in poor control of the Type-I error and in general in an under-conservative behavior of the test (there is a high chance to overestimate the variance).

To avoid these issues, a nonparametric eigen sign-flip score test has been designed in [11]. It is a modification of the sign-flip test based on a spectral decomposition of the smoothing matrix, since it loses power if a strong spatial correlation in the score component, the residuals $(\mathbf{z} - W\boldsymbol{\beta} - \Psi\mathbf{f})$, is present.

where $\Psi$ is the matrix of spatial basis function evalutated in the different space points, and W is the matrix of spatial covariates.

**Algorithm 1** Simple sign-flipping, from [12]

1: Compute the score components $(\mathbf{z} - W\boldsymbol{\beta} - \Psi\mathbf{f})$ under $H_0$
2: Compute the the observed test statistic $T^{obs}$
3: **for** i $\in 1, \ldots, B$ **do**
4: Generate a sign flipping matrix $\Pi^i$
5: Compute the test statistic

$$T^i = W^\top \Pi^i (\mathbf{z} - W\boldsymbol{\beta} - \Psi\mathbf{f})$$

6: **end for**
7: Use the $T^1, \ldots, T^B$ to obtained the $p$-value

### 5.2.3 Nonparametric Eigen sign-flip score test

The nonparametric Eigen sign-flip score test is an asymptotically exact test which preserves the finite sample invariance of the covariance structure of the test statistics under random sign-flips (this does not hold for the simple sign-flip test). It exploits an appropriate decomposition of the matrix $\Psi B_n \Psi^T$, with $B_n = (\Psi\Psi^T/n + \lambda P)^{-1}$ and P discretization of penalizing term, in order to reduce the effect of the spatial dependence, without any parametric assumption on the form of the correlation structure. This leads, even in the more complicated cases, to a very good control of Type-I error, accompanied by an high power.

In our case, we performed these tests in the model (8), relying on the generalization of Eigen sign-flip score test in the case of **space-time models**. We find out that the only two stressors that were relevant in the parameteric part were "Varroa Mite" and "Pesticides", since for the other ones there was not enough evidence at 5% to reject H0 in (9). While the other stressors were removed one by one with univariate testing, when only Varroa and Pesticides were left, in the simultaneous test with the two of them we obtained a pvalue of almost 0, meaning that there is evidence to say that vector $\boldsymbol{\beta} = (\beta_1, \beta_2)$ is significantly different from 0, even if its components are really low. The linear effect of the two main stressors is indeed slightly relevant, even if not that high. The fact that the other stressors should not be included in the linear terms agrees with the idea that distribution of loss is the result of complex nonlinear interaction between the morphology of the region, the environmental conditions and and the timing of events. Moreover, we showed at the beginning of the report that the stressors are all correlated, so including them all in the parametric part of the model would have led to multi-collinearity issues.

# Final conclusions

We have tried to provide possible explanations of these curves associated with losses. In general, states in the Southwest and South-Central are most at risk.

These are the states with higher Varroa concentrations, warmer autumns, and medium to low precipitation.

In contrast, states in the Northwest are those with lower Varroa concentrations, cooler autumns and good rainfall.

Finally, the eastern states show average losses and stressor level.

To summarize what we have done in our project:

1. we analyze the risk factors that can be associated with the disappearance of colonies,
2. estimate the economic impact associated with lost honey, (although losses from missed pollination could be even more significant),
3. show how these losses are not bearable for beekeepers.

   Finally, this information was gathered to provide suggestions for public authorities to better develop resilience plans.

# Code

The code is available at the following Github page.

# References


[1] Lucas A Garibaldi et al. 'Los polinizadores en la agricultura'. In: *Ciencia hoy* 21.126 (2012), pp. 34–43.

[2] Vidal-Naquet. *Honeybee Veterinary Medicine*. 2018.

[3] Otis GW. Mattila HR. 'Dwindling pollen resources trigger the transition to broodless populations of long-lived honeybees each autumn.' In: (2007).

[4] Erik Stokstad. 'Pesticides can harm bees twice—as larvae and adults'. In: *Science.org* (2019). DOI: 10.1126/science.acx9706.

[5] Kirsten S. Traynor et al. 'Varroa destructor: A Complex Parasite, Crippling Honey Bees Worldwide'. In: *Trends in Parasitology* 36.7 (2020), pp. 592–606. ISSN: 1471-4922. DOI: https://doi.org/10.1016/j.pt.2020.04.004. URL: https://www.sciencedirect.com/science/article/pii/S147149222030101X.

[6] Elena N. Ieno Alain F. Zuur. *Mixed effects models and extensions in ecology with R*. 2008.

[7] Mara S Bernardi et al. 'A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province'. In: *Stochastic environmental research and risk assessment* 31 (2017), pp. 23–38.

[8] Eleonora Arnone et al. ‘Modeling spatially dependent functional data via regression with differential regularization’. In: *Journal of Multivariate Analysis* 170 (2019), pp. 275–295.

[9] Laura M Sangalli, James O Ramsay and Timothy O Ramsay. ‘Spatial spline regression models’. In: *Journal of the Royal Statistical Society: SERIES B: Statistical Methodology* (2013), pp. 681–703.

[10] Laura Azzimonti et al. ‘Blood flow velocity field estimation via spatial regression with PDE penalization’. In: *Journal of the American Statistical Association* 110.511 (2015), pp. 1057–1071.

[11] Federico Ferraccioli, Laura M Sangalli and Livio Finos. ‘Some first inferential tools for spatial regression with differential regularization’. In: *Journal of Multivariate Analysis* 189 (2022), p. 104866.

[12] Ferraccioli Federico. *Nonparametric methods for complex spatial domains: density estimation and hypothesis testing*.