POLITECNICO MILANO 1863

M.Sc. Computer Science and Engineering
Data Analysis for Smart Agriculture

A Study on Laying Hen

Davide Canali	-	10674880	-	davide1.canali@mail.polimi.it
Matteo Cordioli	-	10611332	-	matteo.cordioli@mail.polimi.it
Federico Camilletti	-	10619856	-	federico.camilletti@mail.polimi.it
Shakiba Shahidiani	-	10826424	-	shakiba.shahidiani@mail.polimi.it

December 19, 2022

Contents

1 Introduction **2**

2 Data **2**
2.1 Data strategy: research and mining 2
2.2 Data Collection 2
2.3 Data Cleaning 3

3 Analysis of Each Cycle **3**
3.1 Clustering Techniques 3
3.2 Cycle X1 3
3.3 Cycle X2 4
3.4 Cycle Y 5
3.5 Cycle Z 6
3.6 Cycle A 7
3.7 Cycle B 8
3.8 Cycle C 9

4 Common features **10**

5 Organic vs non-organic cycles **12**

6 Death-season correlation **15**

7 Cycle C **17**

1 Introduction

In this study, we will expose the data analysis done starting from data gathered from an eggs farm near Mantova to see if it's possible to improve both animal welfare and farmer revenue.

The farm under analysis is "Boccola Allevamenti Avicoli Soc Agricola SS" and has around 40'000 chickens that produce organic eggs. They provided us with data starting from 2014, the production of eggs is divided into cycles lasting about 13 - 15 months each. The analysis will consider 5 complete cycles and the current 2022 cycle. The first 2 cycles (called X and Y) are non-organic which means the chickens are treated differently from the last 4 cycles (Z, A, B, C) which are organic.

The chickens are divided into 2 equal-sized barns, both of which receive the same treatment and operations.

The farmer is in an agistment that takes care of refurbishing the feed and the livestock.

Upon talking with the farmer we focus our attention on 3 main topics which involve:

- Understanding the mortality between different cycles and organic with non-organic.
- Improve the welfare of the chickens.
- Quantify the losses in cycle C

2 Data

2.1 Data strategy: research and mining

Approaching our data analysis we had to decide whether to start with research and mining or with use cases. The first option implies finding a dataset from which we could extract a use case. The second one consists in having a use case and then finding data useful to reach the objective. Since at the beginning of the process our knowledge of the farming industry was limited and not concrete it was difficult for us to find a real use case. For this reason, we decided to start with research and mining. Our idea was to contact some farming industries with unexploited data at their disposal and then ask them if they needed some kind of data analysis.

The research and mining approach was successful and helped us to find a farm that has around 40,000 chickens for egg production. As a result of direct dialogue with the farmer, we understood that he must register chickens' data because of regulations, but the data were completely unexploited. Furthermore, we noticed that the breeding process was a kind of function with inputs like feed, water, temperatures and eggs as the main output. It was clear that we could have the possibility to optimize this function.

At this point, our research brought us a dataset and an objective. So we decided to start our data analysis in this field.

2.2 Data Collection

The farmer collected the data thanks to sensors and by inspecting the barns each day. Before cycle A the data were only collected in paper spreadsheets and to be used were needed to be imported in a digital format. The data on paper had 5 features: date, death count, current chickens in the barn and eggs produced and sold when the data became saved digitally 2 extra features were collected which are the amount of feed and water consumed. As a result of direct dialogue with the farmer, we understood that some of these elements are not entirely accurate and have been partially measured. Due to the limitations of non-mechanized poultry farms, the water consumed by chickens could not be accurately recorded, so this element should not be given much weight in the analysis. According to our analysis of factors affecting chicken health, which was based on reading articles and consulting with chicken farmers, the information recorded by chicken farmers should be combined with weather data in the region. We also include in our analysis the market data from the "Mercato avicunicolo di Forlì" which is the reference point of the farmer.

2.3 Data Cleaning

Before starting to analyze the data we cleaned them. In all the cycles we performed these next steps:

- Removed tail and head (20 rows each more or less).
- Adjusted feed and water consumption to correct typos and switchboard errors.
- Adjusted eggs production when % of laid when above 100%, distributed in the days before assuming that those eggs were not collected.
- Adjusted death count to match the delta of death at the end of the cycle.
- Checked that the difference between the total amount of eggs produced and the number of eggs sold was less than 0.6%.

Specifically, in cycle A we also adjust a series of missing data regarding the water consumption adding them following the cycle distribution. Cycles X and Y needed some normalization in the number of eggs per day, many days had the same amount of eggs, which is statistically impossible. Talking with farmers we understand that those numbers were placed for easiness to sell the eggs, and a replacement of all those numbers was needed. We replaced the value of those days with the average of the 7 days before.

3 Analysis of Each Cycle

3.1 Clustering Techniques

We used 3 different clustering techniques K-Means, Hierarchical and Density Based Clustering. To choose the parameters for these methods we used different solutions:

- K-Means: silhouette technique to find the optimal number of clusters.
- Hierarchical: plot the dendrogram to find the optimal number of clusters.
- Density: plot the distance of k-NN to find the optimal ϵ.

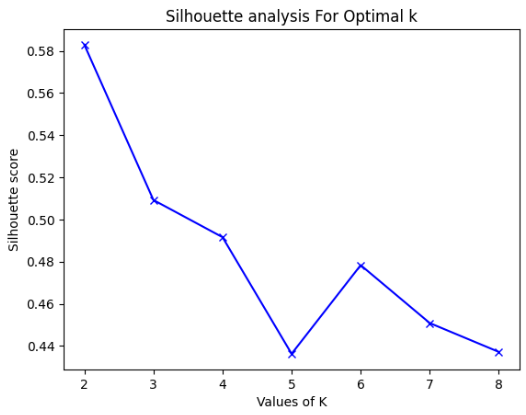
3.2 Cycle X1

Arrival Date	#Chickens	Frist Laid	End of Cycle	Organic	#Eggs
20/01/2014	33.743	18/03/2014	08/07/2015	No	12.375.840

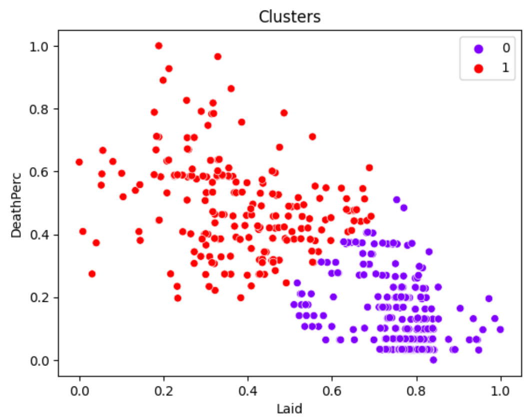
Significant features:

- Average Percentage of Laid Eggs each day: 80,69%
- Average Percentage of Death each day: 0,032%
- Average Temperature during the cycle: 15,34 °C
- Average Humidity during the cycle: 72,83%

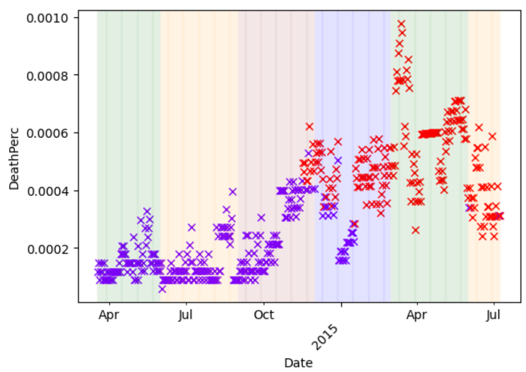
Clusters:



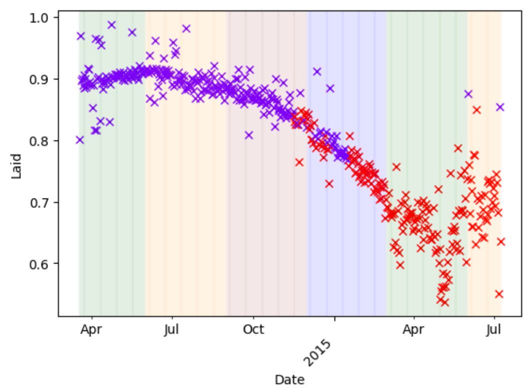
Silhouette



K-Means cluster scatter plot



Percentage of death



Percentage of laid eggs

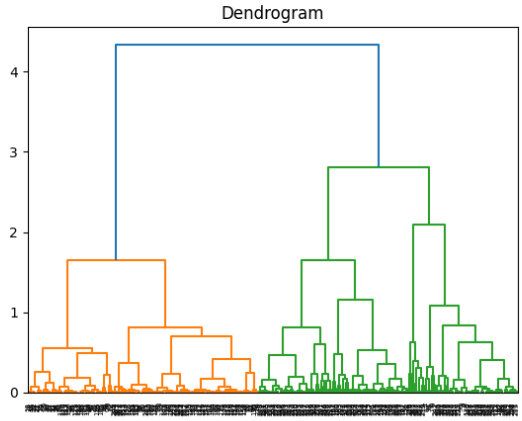
3.3 Cycle X2

Arrival Date	#Chickens	Frist Laid	End of Cycle	Organic	#Eggs
26/05/2014	23.898	15/07/2014	21/06/2015	No	7.558.799

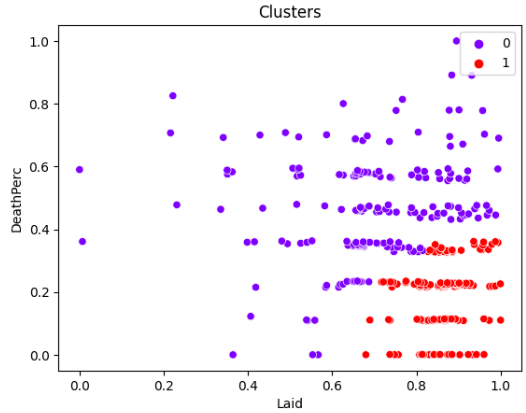
Significant features:

- Average Percentage of Laid Eggs each day: 95,36%
- Average Percentage of Death each day: 0,021%
- Average Temperature during the cycle: 13,76 °C
- Average Humidity during the cycle: 76,04%

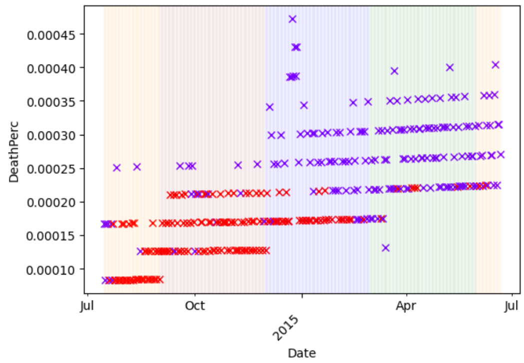
Clusters:



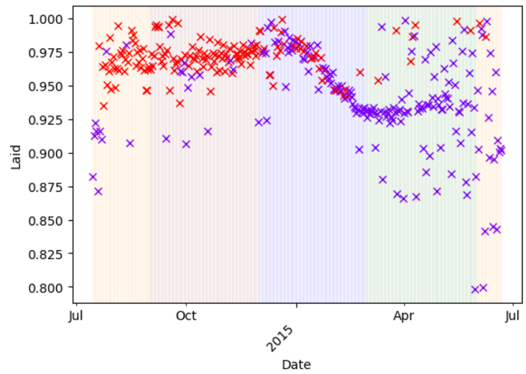
Dendrogram



Hierarchical cluster scatter plot



Percentage of death



Percentage of laid eggs

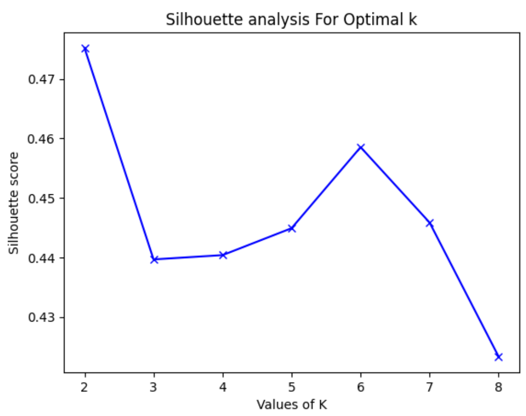
3.4 Cycle Y

Arrival Date	#Chickens	Frist Laid	End of Cycle	Organic	#Eggs
11/08/2015	57.346	05/10/2015	27/09/2016	No	16.759.240

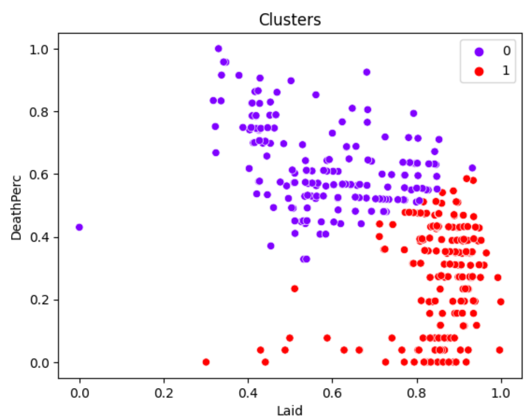
Significant features:

- Average Percentage of Laid Eggs each day: 83.70%
- Average Percentage of Death each day: 0,022%
- Average Temperature during the cycle: 14,20 °C
- Average Humidity during the cycle: 74,54%

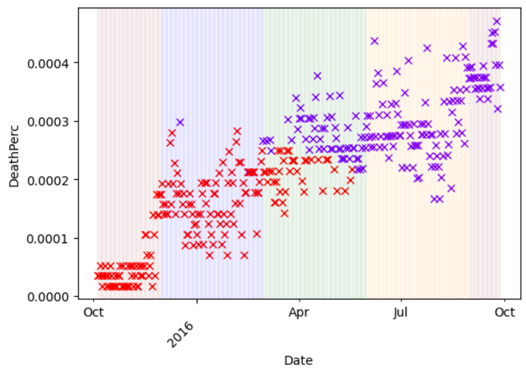
Clusters:



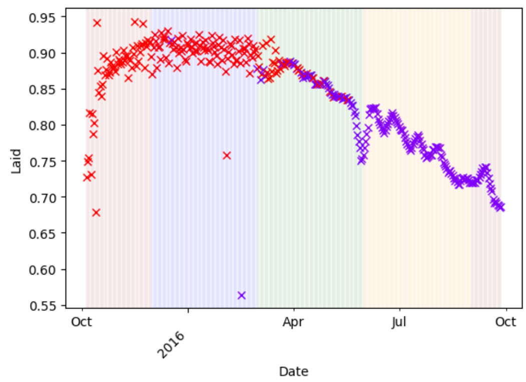
Silhouette



K-means cluster scatter plot



Percentage of death



Percentage of laid eggs

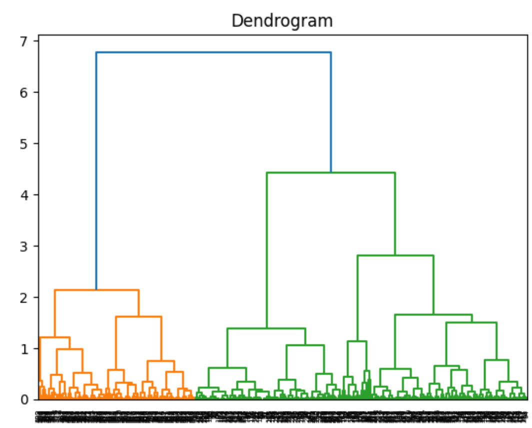
3.5 Cycle Z

Arrival Date	#Chickens	Frist Laid	End of Cycle	Organic	#Eggs
17/11/2016	42.130	08/01/2017	27/05/2018	Yes	17.721.240

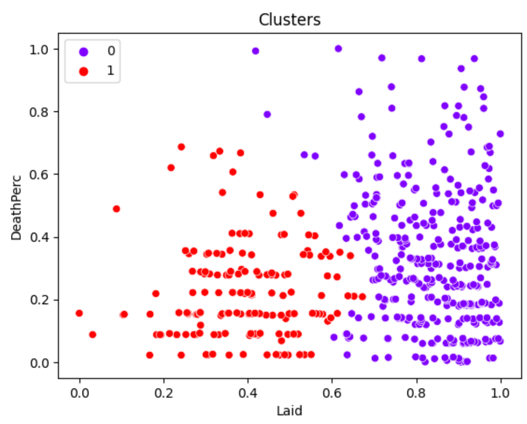
Significant features:

- Average Percentage of Laid Eggs each day: 87,78%
- Average Percentage of Death each day: 0,021%
- Average Temperature during the cycle: 13,07 °C
- Average Humidity during the cycle: 74,27%

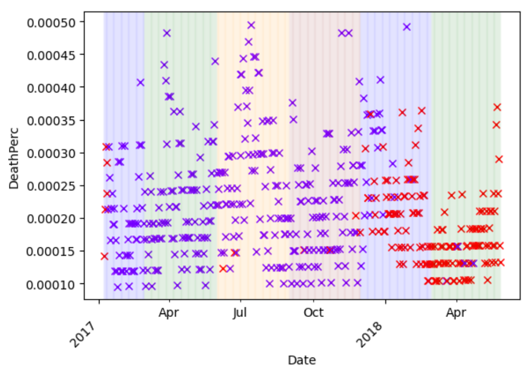
Clusters:



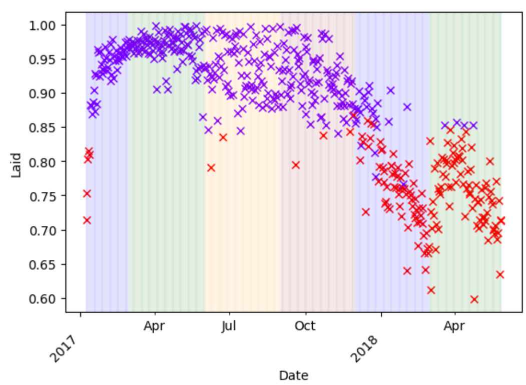
Dendrogram



Hierarchical cluster scatter plot



Percentage of death



Percentage of laid eggs

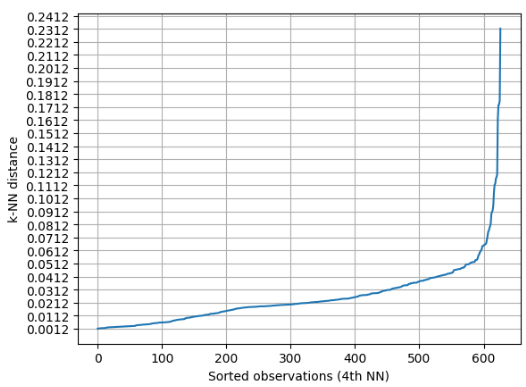
3.6 Cycle A

Arrival Date	#Chickens	Frist Laid	End of Cycle	Organic	#Eggs
19/7/2018	42.009	01/9/2018	19/5/2020	Yes	20.208.086

Significant features:

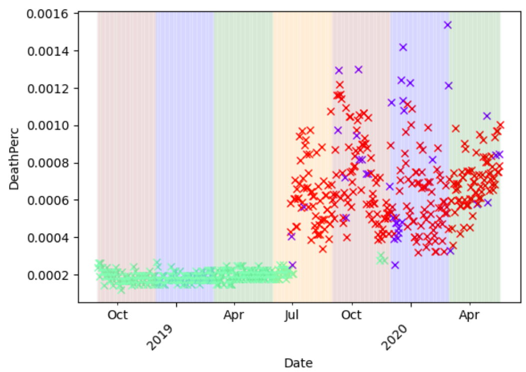
- Average Percentage of Laid Eggs each day: 83,8%
- Average Percentage of Death each day: 0,044%
- Average Temperature during the cycle: 12,99 °C
- Average Humidity during the cycle: 74,55%

Clusters:

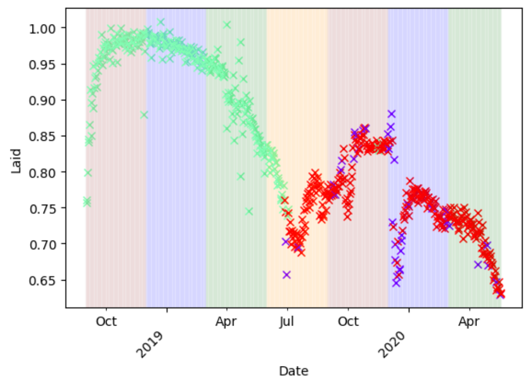


k-NN distance to find ϵ

Density scatter plot merged



Percentage of death



Percentage of laid eggs

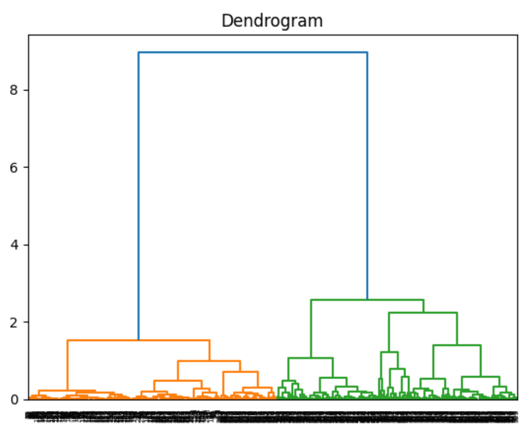
3.7 Cycle B

Arrival Date	**#Chickens**	**Frist Laid**	**End of Cycle**	**Organic**	**#Eggs**
09/08/2020	42.098	24/09/2020	02/05/2022	Yes	18.392.640

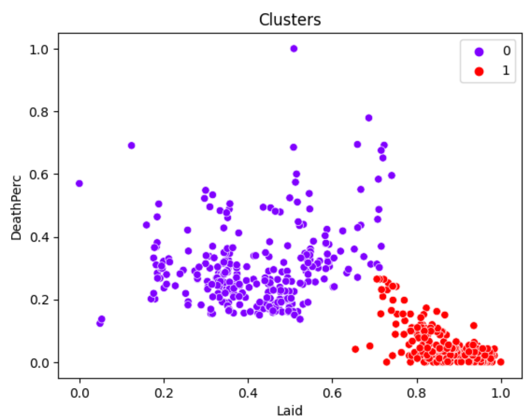
Significant features:

- Average Percentage of Laid Eggs each day: 81,39%
- Average Percentage of Death each day: 0,0494%
- Average Temperature during the cycle: 11,81 °C
- Average Humidity during the cycle: 73,24%

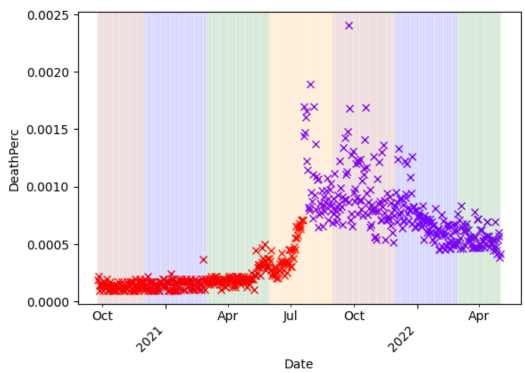
Clusters:



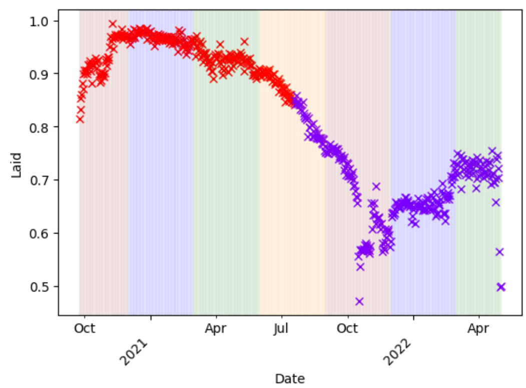
Dendrogram



Hierarchical cluster scatter plot



Percentage of death



Percentage of laid eggs

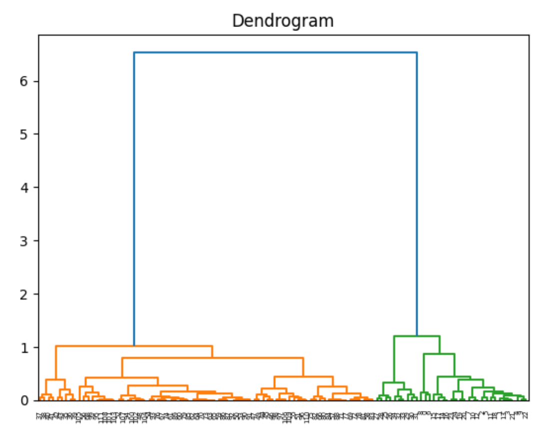
3.8 Cycle C

Arrival Date	#Chickens	Frist Laid	End of Cycle	Organic	#Eggs
20/06/2022	42.098	08/08/2022	In progress	Yes	In progress

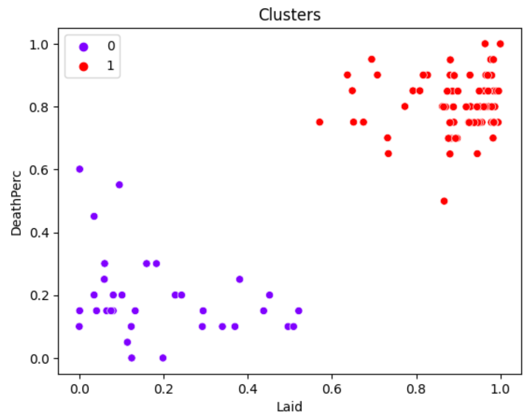
Significant features:

- Average Percentage of Laid Eggs each day: 55,85%
- Average Percentage of Death each day: 0,043%
- Average Temperature during the cycle: 18,61 °C
- Average Humidity during the cycle: 70,05%

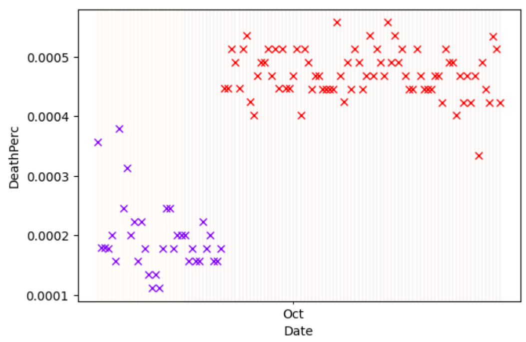
Clusters:



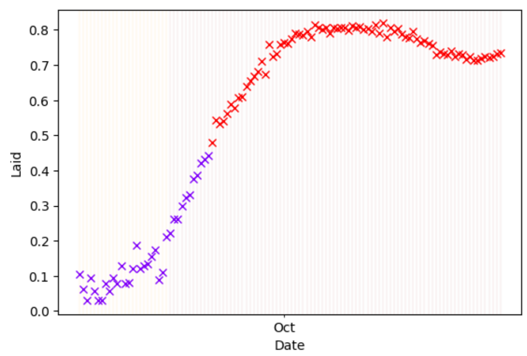
Dendrogram



Hierarchical cluster scatter plot



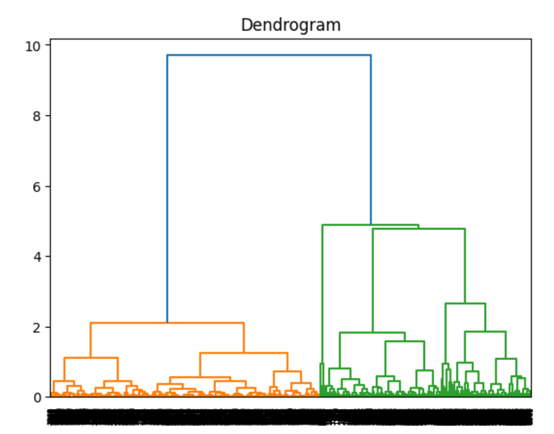
Percentage of death



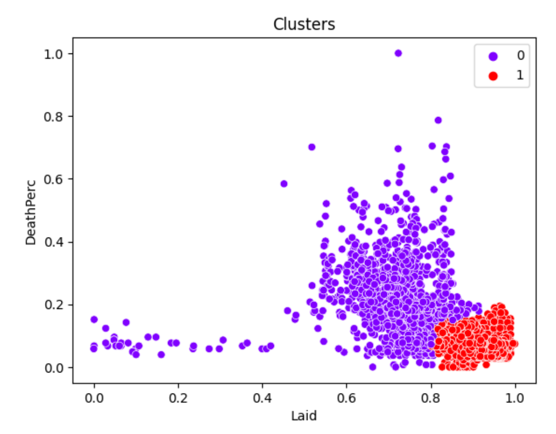
Percentage of laid eggs

4 Common features

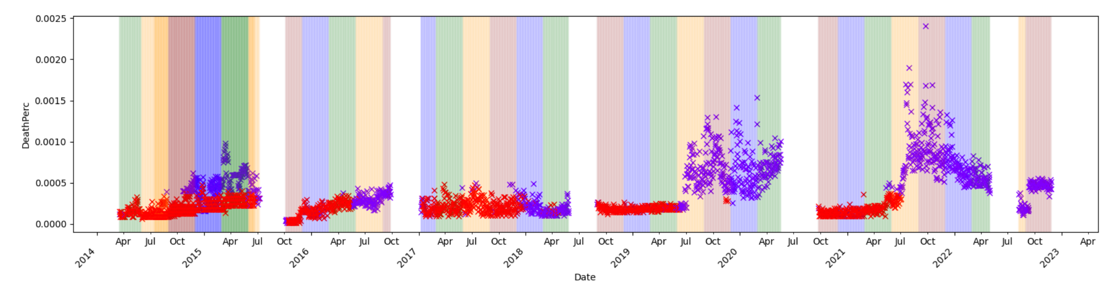
In this section, we will discuss what we discovered by comparing all the cycles together. After seeing that each cycle can be clustered in two parts, those clusters also divide the cycle into two distinct sections in a plot based on date. We tried to apply the same idea in a dataset composed of all the cycles together.



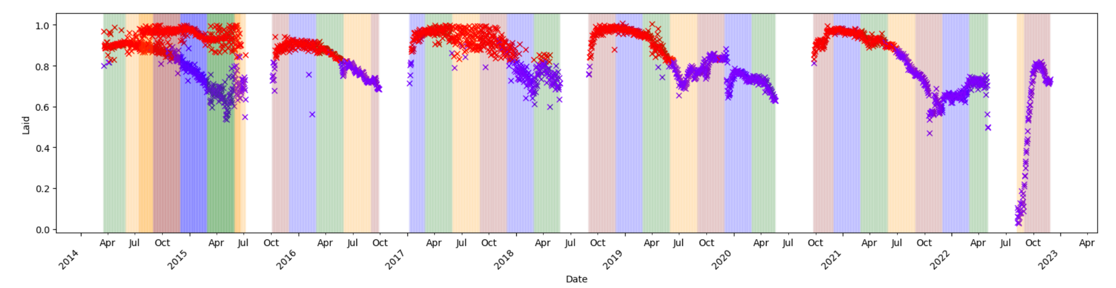
Dendrogram



Hierarchical cluster scatter plot



Percentage of death

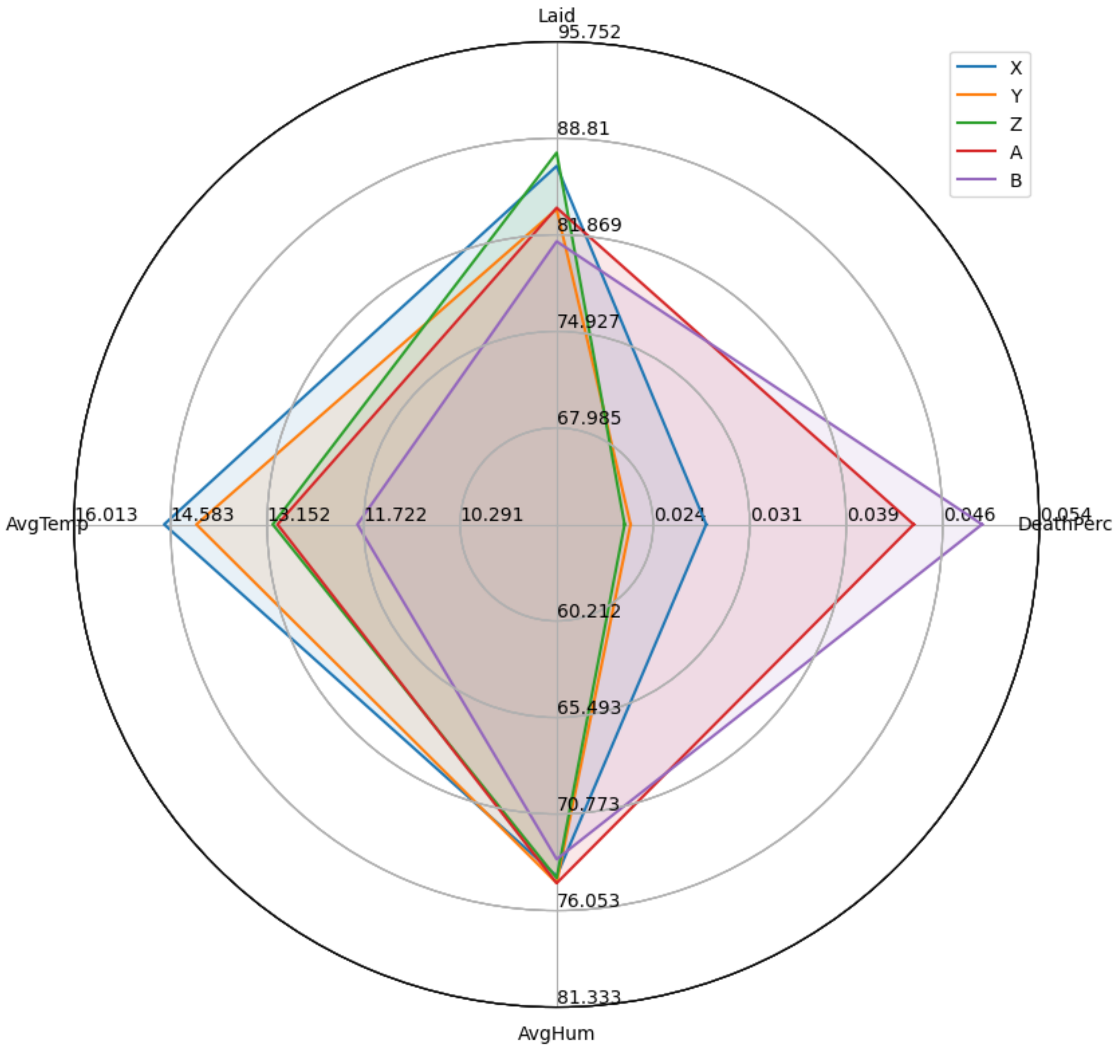


Percentage of laid eggs

Looking at the clusters we find that indeed the division of each cycle is respected. We find also that the only one that didn't respect this division is cycle C, which could be expected since it just started, but the interesting thing is that the whole cycle C is clustered as cluster 1 which is the one that identifies the end of every other cycle so we can see how bad is performing this new cycle both in death and laid rate. This can show how the ban on beak-cutting affected the welfare of the chickens, with the beak the violence between animals increased a lot and not only increase the death rate but increased a lot the stress of the chickens reducing their productivity.

To better compare the different cycles we made a spider plot with 4 features:

- Average Percentage of Laid Eggs each day
- Average Percentage of Death each day
- Average Temperature during the cycle
- Average Humidity during the cycle



5 Organic vs non-organic cycles

One of the main differences in the cycles under analysis is the breeding method. Cycles A, B, C and Z are grown with organic farming, a method regulated by the European Union with the regulation 589/2008 modified from text 2168/2017 while the cycles Y, X1 and X2 are non-organic. We can summarize the main differences:

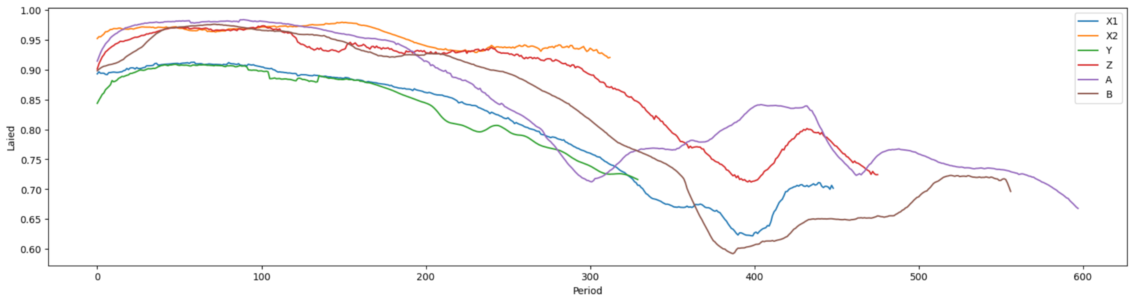
Characteristics	Non-Organic	Organic
Chickens/m^2	9	6
Outdoor access	✓	✓
Feed requirements	×	✓

Given the differences in the breeding methods, we analyze if there are implications in the cycles' behavior. Since the non-organic ones have no information on feed and water consumption, we can't go deep into the analysis because of the lack of inputs. For this reason, we will focus mainly on the outputs like the percentage of eggs laid and deaths.

One of the most important outputs is the productivity of a cycle expressed as EggsProduced#Chickens. In the following results, we exclude cycle C since it is in progress. Furthermore, we round the data to the second digit.

Cycle	Organic	Eggs per chicken	Breeding days	Productivity
X1	×	366.77	478	0.80
X2	×	316.29	342	0.95
Y	×	292.25	359	0.83
Z	✓	420.63	505	0.87
A	✓	481.04	627	0.84
B	✓	436.90	586	0.81

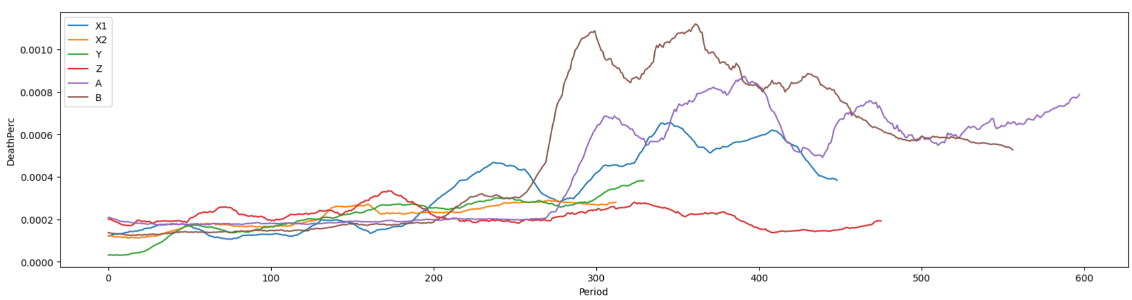
We can notice that non-organic cycles have a lower EggsProduced/#Chickens rate than the organic ones, but considering the days of breeding the non-organic cycles are better. The medium organic cycle productivity is 84% while the non-organic one is 86%.



Beyond the productivity in terms of eggs produced there is another important aspect to consider which is the chickens' health. Since it is a difficult feature to value, we decided to consider the chickens' death percentage on the number of chickens as a health indicator.

Cycle	Organic	Deaths on chickens	Breeding days	Daily deaths average
X1	×	0.14	478	10.10
X2	×	0.07	342	4.93
Y	×	0.07	359	12.03
Z	✓	0.10	505	8.53
A	✓	0.24	627	16.18
B	✓	0.25	586	18.07

In organic cycles, the 'Daily deaths average' is higher than in non-organic ones: on average it is 5,76 higher since it is 14.78 in organic cycles and 9.02 in non-organic ones.



Since non-organic cycles have both better productivity and a lower death rate it would be interesting to understand the reason. The parameters that could have an impact on this are various:

- Alimentation: organic breeding has strict rules also in terms of feed
- Temperature: if a year is particularly hot chickens could suffer more
- Space: in organic cycles chickens have more space
- Time: the duration of a cycle could affect chickens' health

Since alimentation is one of the most important factors in the health of a living being and these data are missing in every cycle except for cycles A and B it could be difficult to answer the question about chickens' health. However, we tried to give some explanation based on our dataset and expert knowledge.

First of all, in almost every cycle, the correlation between the percentage of deaths and the period of breeding is high:

- Cycle X1: 0.76
- Cycle X2: 0.68
- Cycle Y: 0.86
- Cycle Z: -0.14
- Cycle A: 0.73
- Cycle B: 0.63

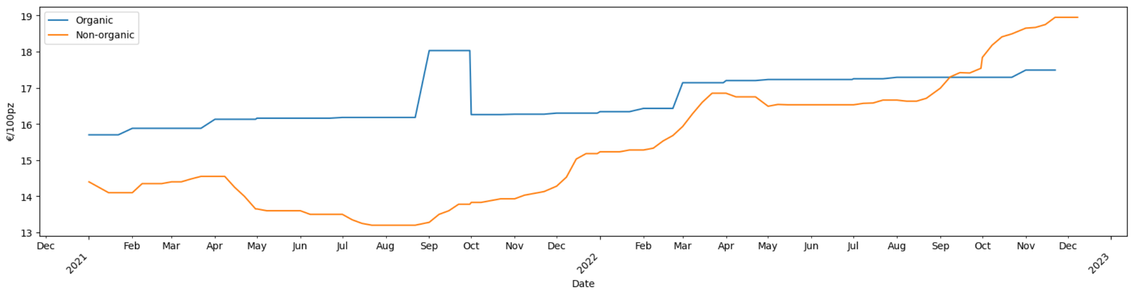
The correlation of deaths with temperature is not as significant if considered in the entire cycle. It becomes more interesting if it is considered just the part of the cycle in which the temperature rises:

- Cycle X1: -0.12
- Cycle X2: 0.3
- Cycle Y: 0.6
- Cycle Z: 0.34
- Cycle A: 0.61
- Cycle B: 0.61

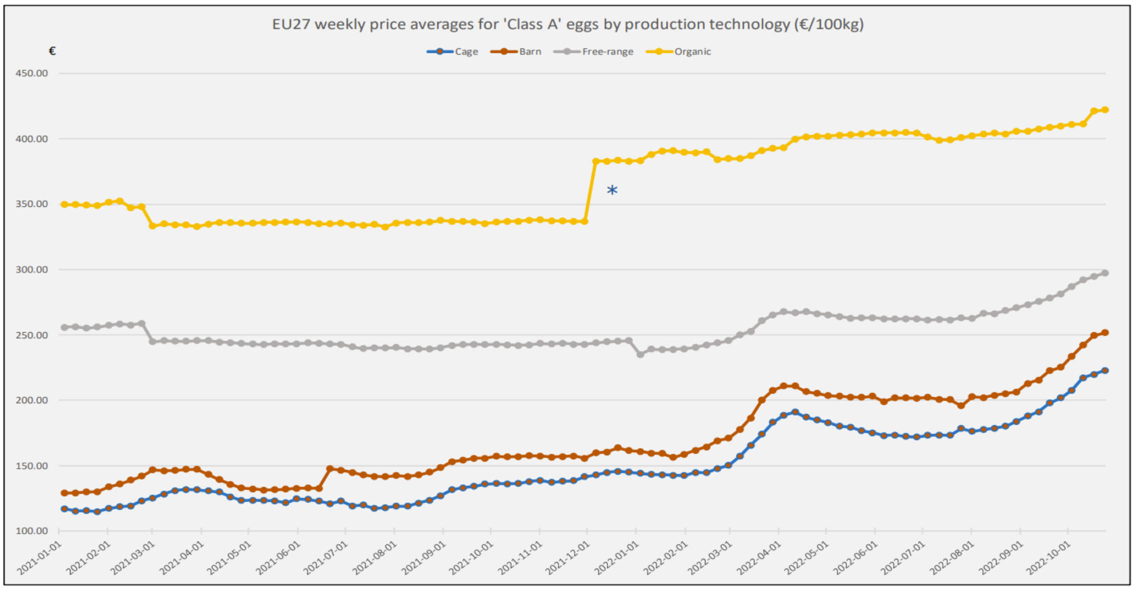
The square meters per chicken haven't shown any significant correlation since the number of chickens is constant for both the organing and the non-organic cycles.

It can be observed that the cycles that start at the end of the summer (A, B, Y) have a high correlation between deaths and temperature, while the ones that starts in winter or spring (Z, X1) have a lower correlation. The results suggest that there is a high correlation between the chickens' status and the time of breeding, but also that the temperature is another important aspect to consider. For this reason, we decided to direct the analysis to a deeper comprehension of these observations.

The last variable that could help the farmer in the decision-making process is the egg price.



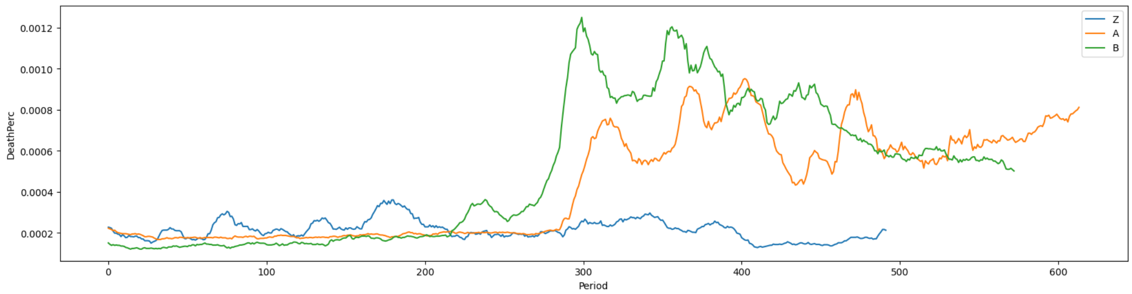
As we can see in the Forlì market the non-organic eggs price is higher than the organic ones. However, it is not a parameter shared by the entire European market:



6 Death-season correlation

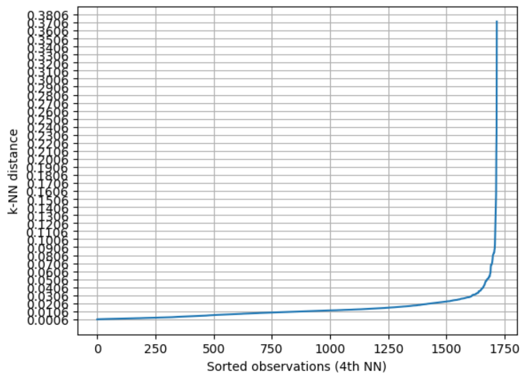
In this section, we will analyze how the period when the chickens are introduced into the barn affects the mortality rate. This effect will be more visible in the organic cycles where the chickens are more fragile due to a lack of nutrients.

First, we plot the Death Rate of all the organic cycles (Z, A, B).

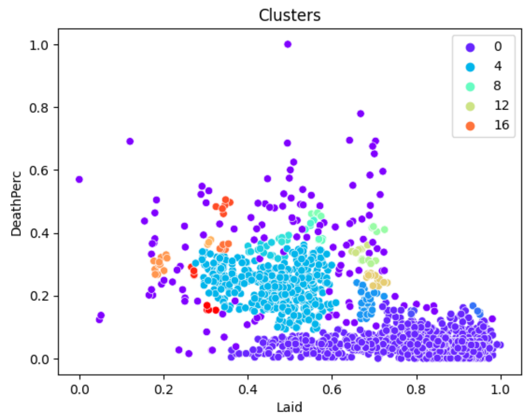


It's clear that the Z cycle outperforms the other two by a lot, so we will analyze what are the differences between those 3 cycles.

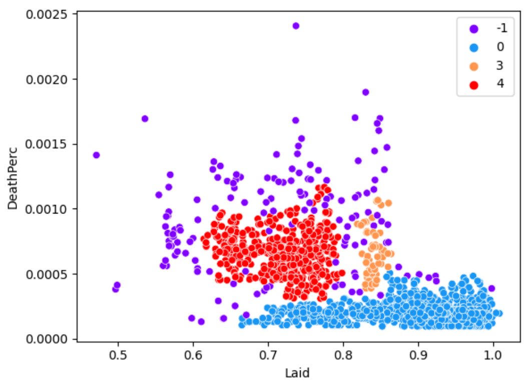
The number of chickens at the start of the cycle is identical ($\sim$ 42.000), and the feed in an organic farm is standard and checked so it doesn't change between those cycles. The difference in mortality cannot be explained with just "a lucky cycle" so we tried to analyze those graphs deeper.



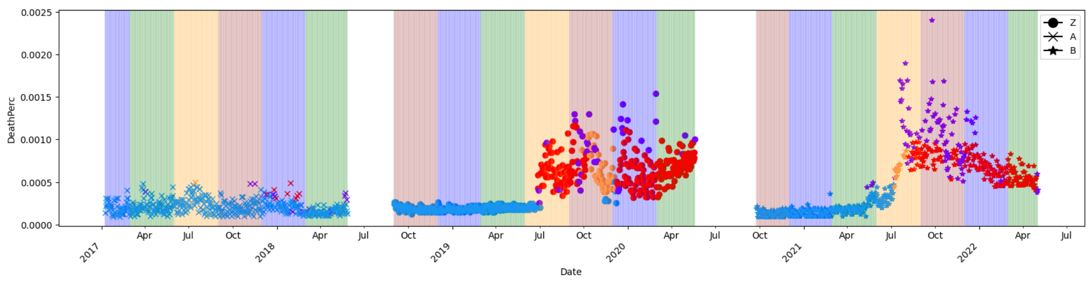
k-NN distance to find ϵ



Raw Density



Density Cleaned

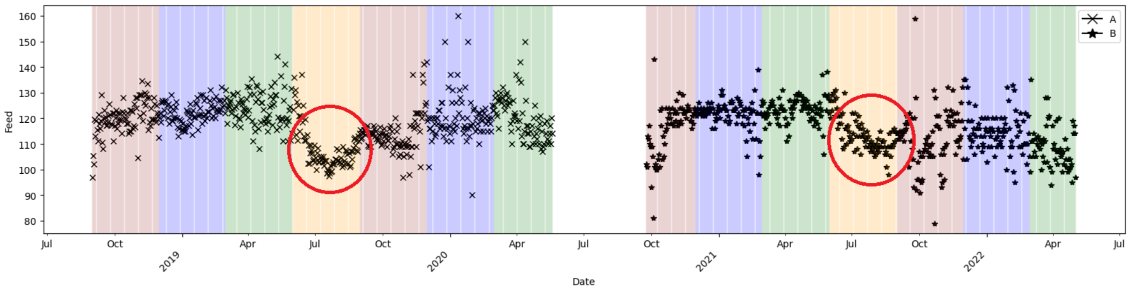


Percentage of death

We see as expected that cycle Z is clustered the same as the start of the other cycles. But the interesting analysis is over the season in which events happen.

- Both cycles A and B start to see an increase in mortality around mid-summer whereas cycle Z just saw a small peak.
- Cycles A and B started around early Autumn whereas cycle Z started in mid-winter

What also happens in the summer is that chickens start to eat less and so become weaker.



After discussing it with the farmer we conclude that the summer is the most challenging period for the chickens since:

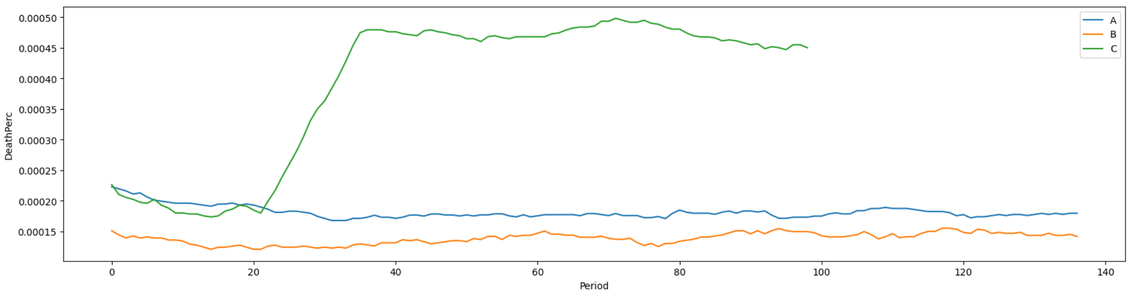
- The barns don't have air conditioning and the temperature can be quite high.
- The chickens are weaker due to age and the quality of nutrients in the organic feed.

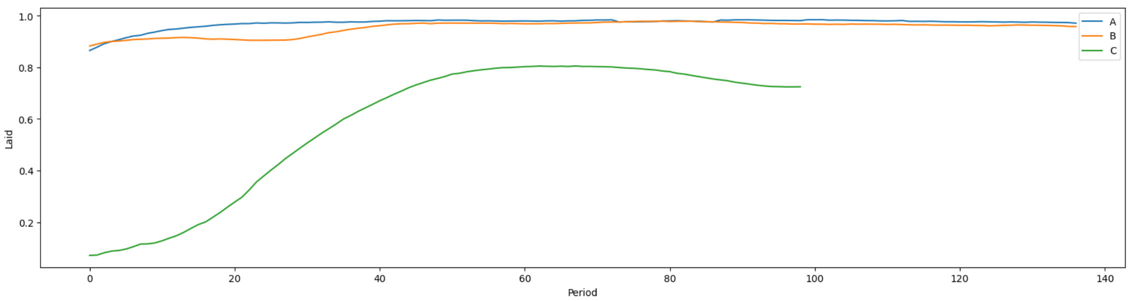
To overcome those problems it seems that starting the cycle in mid-winter improves the mortality rate since during the summer the chickens are younger and stronger.

To put it in perspective, cycle Z had 4308 deaths compared to cycles A and B which had respectively 7449 deaths and 9130 deaths across the same period, a reduction of 42,1% and 52,8%. This help the production of cycle Z which, in the same period, produced 500.000 more eggs than cycle A and 1.200.000 more eggs than cycle B. Which is respectively an increment of 2% and a 7% in eggs produced. So we can say that on average starting the cycle in mid-winter can increase the production by 4,5% while reducing the mortality by 46,4%. Using the current price of organic eggs in the Forlì Exchange Hall of 15.63€/100 eggs we can estimate an increase in revenue of 132.000€.

7 Cycle C

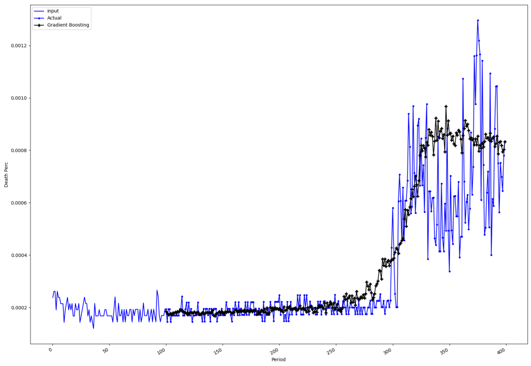
Cycle C is a particular cycle, it is an organic cycle so it has the same characteristics as cycles A and B but from this one, a new regulation has been introduced. This new law banned the cut of the beak from the chick which increased the violence between the animals and this led to a higher mortality rate.

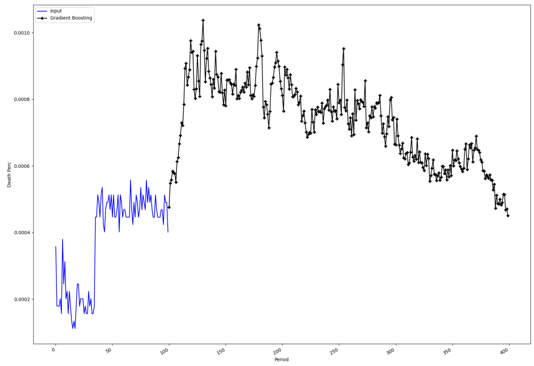




A difference of 0,03% means on average that in this period every day die 12 chickens more than the normal rate. This has an effect also on production since the chickens have poorer welfare they produce less.

Using a gradient boost regression we were able to guess the death that this cycle will have in the first 400 days based on the data that we have with the other 2 cycles A and B.





Using this prediction we say that we guess that in the first 400 days the farmer will lose 9322 compared to the 4903 of cycle A and 6369 of cycle B. If we, instead of using a predictor, assume that cycle C follows a constant mortality rate in its first 270 days and then has a raise, like the other cycles, we obtain a similar result with a death count of around 9000 chickens. This increase in mortality of respectively 90% and 46% will inevitably reduce the number of eggs produced. If the cycle will be able to keep this rate of production, which seems to be its peak, after 400 days the total number of eggs produced will be 10.059.107 much less than 14.285.849 and 14.270.250 of cycles A and B, this means a loss of around 650.000€.