Point Process Trends of Maryland Bee Observations

MTH 6384 - Spatial Data Analysis

Brady Lamson

Fall 2025

CU Denver

Introduction

Background Info

The topic of bee populations in the United States is not a new one. We have been aware of the dwindling bee population in this country for decades. Though this topic has a lot of surface level popular appeal through the ”save the bees” movement, there is a lot that still isn’t understood about this trend. For starters, it’s difficult to pin down exact numbers on this decline as it depends on location, species and many other factors. There are also many known overlapping challenges facing native wild bee populations and non-native human managed honeybee colonies. Native wild bee populations have been harmed by habitat loss, urbanization, pesticide use, and climate change to name a few. Honeybee colonies have been ravaged by parasites, disease, other non-native bugs such as the small hive beetle and a bizarre phenomenon called Colony Collapse Disorder (Dr. Underwood, source). This is a complicated and multifaceted issue; there isn’t just a single simple cause driving these populations down.

The importance of this decline goes beyond a simple adoration of the fluffy insects. According to Brianna Randall from the NRCS, ”More than 80 percent of the world’s flowering plants need a pollinator to reproduce; and we need pollinators too, since most of our food comes from flowering plants. One out of every three bites of our food, including fruits, vegetables, chocolate, coffee, nuts, and spices, is created with the help of pollinators.” (source). There are so many managed honeybee colonies in the US because they are crucial for the agricultural branch of our economy to function. As bee populations struggle there are powerful ripple effects that are felt both on the local ecological scale and at the urban level. Referring back to Brianna again, ”pollinators’ ecological service is valued at $200 billion each year”. Meanwhile native bees are the backbone of every states ecological wellbeing by pollinating native flora and being essential parts of the food chain.

I am particularly interested in the challenges to habitat that bees are facing. As climate change and urbanization destroy hives and make previous areas infeasible, I am curious to see if this is reflected in spatial data of bees. If we look at maps over time, do the spatial distributions of bees change? Do we see clustering in new areas or an absence of clustering in areas from prior years? Do the spatial scales of clustering change over time and if so, what does that mean?

Data Source

The data I am using for this project comes from the Global Biodiversity Information Facility (GBIF). This is an international group funded by the world's governments that has the explicit goal of ”providing anyone, anywhere, open access to data about all types of life on Earth” (source). The dataset I'm working with specifically is a collection of insect occurrence records across all of the United States of America, and some areas outside of it, with a focus on bees (source).

This dataset is an aggregation of many different projects and collection efforts across various groups. These groups include USGS employees, federal workers, volunteers, private groups and civilian scientists. It contains an enormous amount of species occurrence records for ”native and non-native bees, wasps and other insects”. These data were collected using a variety of methods such as pan, malaise and vane trapping. This dataset features occurrences going all the way back to 1996 complete with latitude and longitude information, so I'll be treating this as point process data and examining trends across decades.

Results

Data Exploration

The full dataset contains 586 thousand rows and 50 columns. Each row represents an individual insect occurrence. The columns include information on the specific insect observed, information on the exact location it was observed and then a ton of administrative info about the groups collecting and processing this data. Much of this information is suplerfluous and will not be used in this project. For the insect specific info it gets extremely precise, including the kingdom, phylum, class, order, family and more. On the temporal side of things this dataset contains occurrences dating all the way back to 1964, though there are not many rows from before the year 2000.

Filtering and Data Cleaning

The first thing I did to trim the dataset down was to filter for only bees in the dataset. There are 101 distinct insect families in this dataset with only about 5 of those being bee families. For location, I wanted to perform my analysis at the state level. Checking the various states, Maryland has the healthiest dataset featuring over 190 thousand bee occurrences with coordinates. Other states checked, like Colorado, had far smaller sample sizes, an extremely limited variety of location occurrences and data that did not go back as far as I desired. For Maryland's data I could comfortably go back to the early 2000s and occurrences covered most of the state's map.

Looking at the spatial variables, Maryland data had two problems to fix. First were points labeled as in Maryland that fell far outside the bounds of the state. Second were a ton of overlapping points. Of the roughly 190 thousand occurrences in the data, only about 2747 latitude and longitude combinations were unique. This problem I handled by using the stjitter function from the sf package to add a small amount of random error to the location of every point in the data. From there to address the first problem I simply filtered for the points in the dataset that fell inside of Maryland's boundaries using the stintersects function from the sf package.

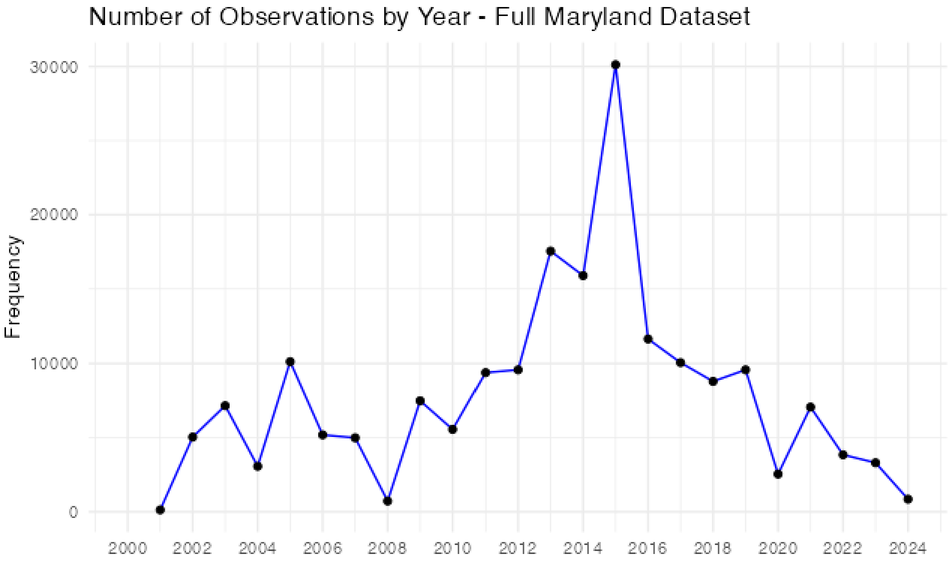


Figure 1: Yearly frequency trends

Next up was handling unbalanced yearly frequencies in the data. As we can see in figure 1 there is an enormous spike in occurrences from 2013 to 2015. I did not want these years to overpower the results of my analysis so I did two things. First I created a new column, decade, and then I created a stratified random sample of the data using that decade column. By doing this I can give all three decades equal weight in my analysis with some additional benefits. This reduced the size of my dataset substantially which enhanced map readability while also reducing computational complexity. To keep the analysis feasible I gave each decade 500 rows giving me a final dataset of 1500 rows. Below is a map showing the random sample used for this analysis.

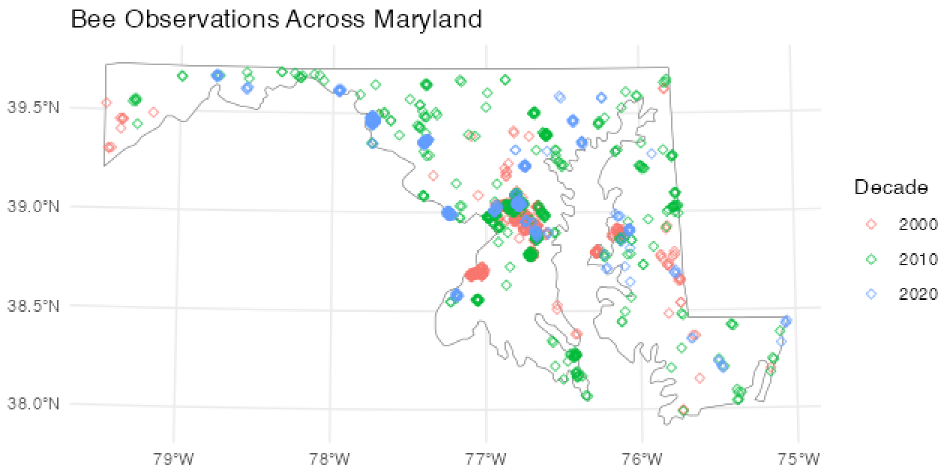


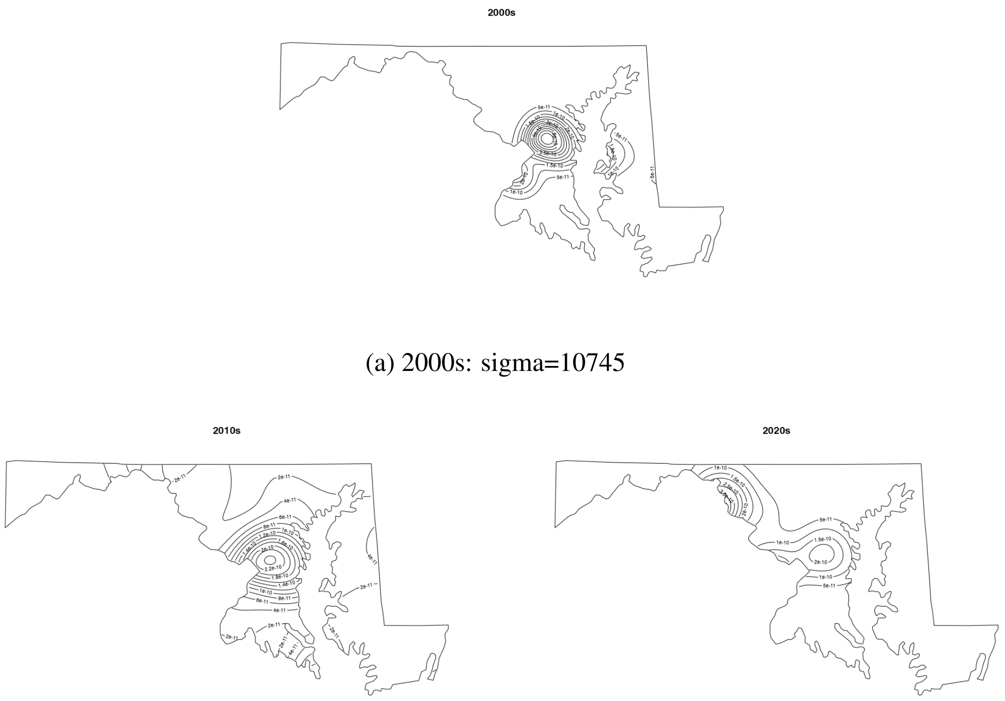
Figure 2: Map of the random sample by decade

This plot is included primarily as a point of reference for later results. Due to the overlapping nature of these points it's difficult to identify a ton of trends from just this image alone. A quick look shows the red points representing the 2000s seem to be found only in around 3 spots on the map. The 2010s show a ton of variety in their locations as do the 2020s.

Spatial Densities by Decade

Before getting into direct comparisons it is helpful to show each decades contour plots as they will be an invaluable reference point for later visualizations. Each of these contour plots was created using Scott's Rule for bandwidth selection. A separate isotropic bandwidth was chosen for each decade to simplify some downstream computations that will be covered soon.

Just looking at the contour plots in Figure 3 already shows some noticable differences.



(a) 2000s: sigma=10745

(b) 2010s: sigma=16674

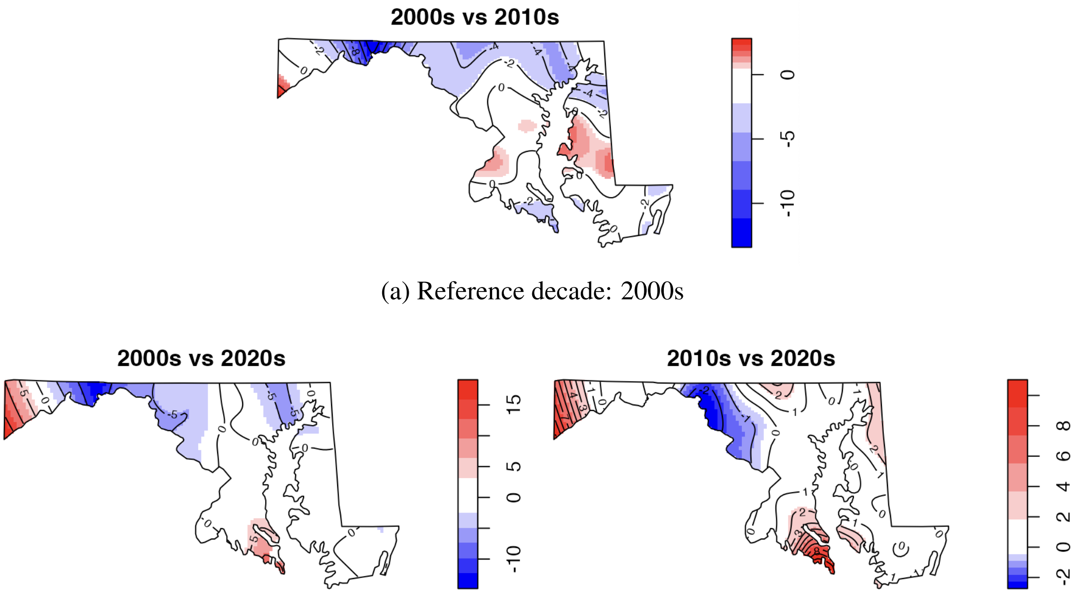
(c) 2020s: sigma=14607

Figure 3: Contour Plots

The density for the 2000s is contained almost entirely in the center of Maryland, between Washington DC and Baltimore. The 2010s meanwhile have densities covering most of the map. The bulk of that density is in the same location of the 2000s, but it's branching out a lot more especially in the north. The 2020s are interesting there is a far more pronounced area of density in the northern part of the state bordering West Virginia. We also see some different bandwidth values for each of the decades, with the 2000s having by far the smallest and the 2010s having the largest.

Comparing Decades - Tolerance Contour Plots

Moving onto actual comparisons, in Figure 4 I have three plots. Each one represents a tolerance envelope contour plot comparing two decades. For these, the bandwidth chosen is the mean of each decade's bandwidth. So the bandwidth of the 2000s vs the 2010s is $(10745 + 16674)/2 \approx 13709$. To aid in basic interpretation, areas of red represent spatial clustering of the reference decade relative to the other decade. Blue would represent spatial clustering of the other decade



(a) Reference decade: 2000s

(b) Reference decade: 2000s

(c) Reference decade: 2010s

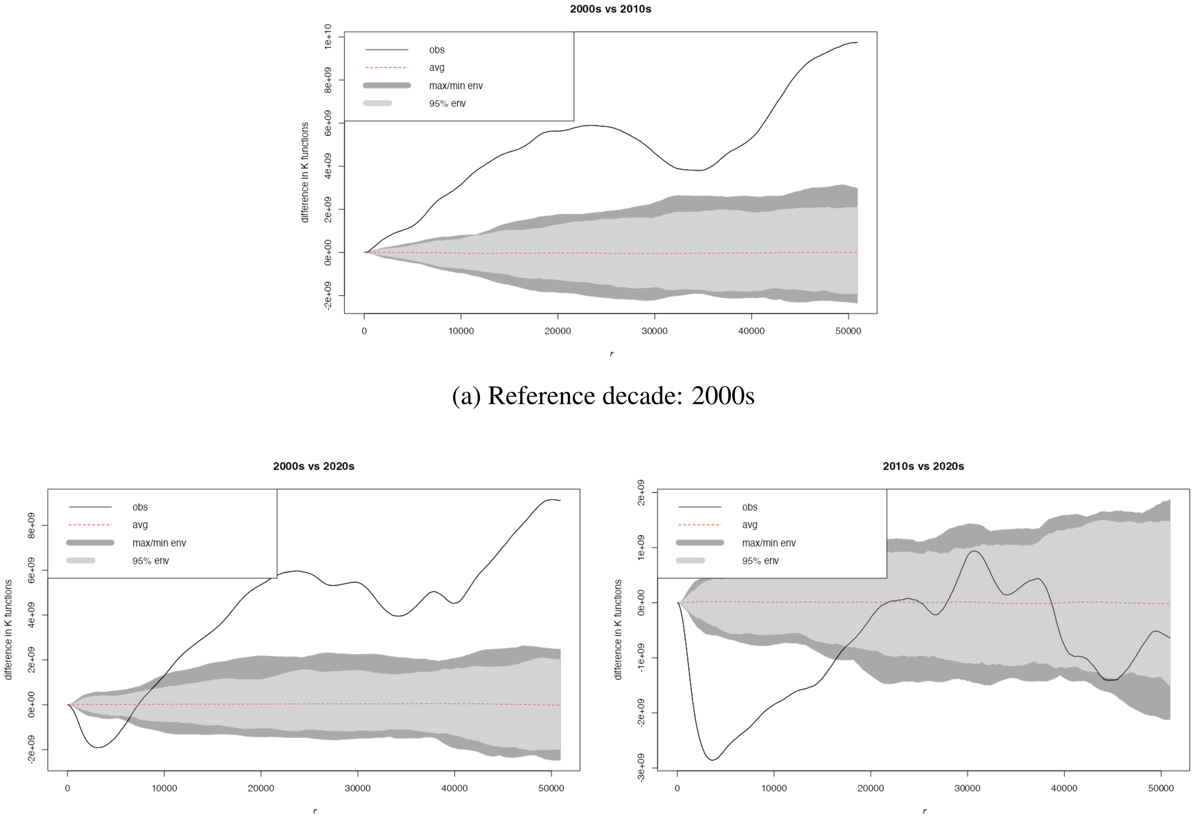
Figure 4: Tolerance Envelope Contour Plots

relative to the reference.

When interpreting these kinds of plots it is important to remember that areas can show up here that may not be present in the contour plots of Figure 3. This is because these plots show clustering of one group relative to the other. This is how we see clustering of the 2000s with respect to the 2020s in the bottom left plot of Figure 4 despite there being no density there in Figure 3. Referring back to Figure 2 here shows a small collection of points there for the 2000s and none in the 2020s, explaining the visual.

The general takeaway here is that all three decades show about the same degree of clustering in the middle of the state. It isn't exact but that region is mostly white in all three plots. The big differences here appear in the northern part of the state in particular, with both the 2010s and 2020s showing clustering with respect to the 2000s in that area. The 2020s is highly clustered in that northwest area bordering West Virginia relative to both other decades which is very interesting. This change in spatial densities through the decades could represent either a movement in bee populations to the northern portion of the state, or a shift in where human researchers chose to collect data. It's difficult to tell.

Comparing Decades - Difference in K-Functions



(a) Reference decade: 2000s

(b) Reference decade: 2000s

(c) Reference decade: 2010s

Figure 5: Difference in K-Functions

Case	Control	Test Statistic	p-value
2000s	2010s	3291	0.05
2000s	2020s	2386	0.05
2010s	2020s	-1294	1

Table 1: K-function Differences - $H_a : KD(r) > 0$

The k-function results show more evidence of the differences between these decades. I will be referring to the summary output not listed here for exact ranges. Of note that r values listed are in meters. Starting from the top, we see $KD(r)$ being larger than the upper envelope limit from 199 to 50921 meters. So there is evidence of clustering of the 2000s relative to the 2010s for nearly the entire range of spatial scales examined. We see similar results comparing the 2000s to the 2020s, though the 2020s appear to potentially show more clustering in the lower spatial scales from 99 to 6464 meters. The 2000s pull ahead from 8752 to 50921 meters. Lastly and most interestingly, the 2020s show clustering relative to the 2010s from 99 to 17305 meters before they both start to behave similarly for the rest of the spatial scales.

I think what is so interesting here is how pronounced the difference is between the 2000s and the other two decades when the 2010s and 2020s show such similar clustering behaviors at most of the spatial scales tested. I believe this is related to both of those decades branching out away from the central location that the 2000s lies in. The 2000s are far more concentrated in their occurrences, explaining the large scale of clustering seen here. Again though, it is difficult to ascertain the exact reason for the change of spatial patterns here. I think it is tempting to assume this is due to bees being forced to branch away from the main cluster due to environmental factors, but this is also easily explained by a branching out in research locations as more people get involved.

Q Nearest Neighbors

Case: 2000s

Contrast	p-value
T150-100	0.005
T200-150	0.110
T250-200	0.470

Case: 2010s

Contrast	p-value
T40-25	0.005
T55-40	0.765
T70-55	0.990

Case: 2020s

Contrast	p-value
T175-150	0.020
T200-175	0.005
T225-200	0.005
T250-225	1.000

Table 2: QNN test statistic contrasts. All 3 tables use the remaining decades as control.

Q nearest neighbor tests were ran as another way to gauge what kind of clustering each decade exhibits. The goal here was to examine at what neighbor counts the clustering began to break down. The approach here was different to the previous tests. Whereas previous sections were direct comparisons between decades, using one as the case and another as the control, here we use the two non-case decades as the controls. This was done for practical reasons, as doing the direct decade comparisons would require separate qnn tests for both controls across all 3 cases. We can't rely on the bidirectional interpretation the other tests provide us here. This does hinder the transfer of interpretation to previous tests in this case, as the qnn tests compare against different distributions than they would using the direct comparison method.

Regardless, Table 2 shows the test statistic contrast results. What's important about these tables is precisely at what points the p-values become large. What we see from the 2000s is that up to 150 nearest neighbors, there are more cases than would be expected under the random labeling hypothesis. This is no longer the case somewhere up to 200 nearest neighbors, where any clustering is mostly explained by the prior threshold. The 2010s in comparison see this breakdown much sooner, only having more cases than expected for up to 40 neighbors. The 2020s actually show the largest values of the decades, showing more cases than expected up to 225 neighbors. This result is surprising when we compare it to the difference in k-function results, and may be a symptom of the problems that arise from the inconsistency in methodology here.

Spatial Scan Results

Decade	Radius	Events	Cases	Ex	p
2000s	96709.9	964	470	321.3	0.005
2010s	51117.7	97	86	32.3	0.005
2010s	74902.7	238	149	79.3	0.005
2010s	4097.6	49	48	16.3	0.005
2010s	28398.4	26	26	8.7	0.005
2010s	1265.7	23	21	7.7	0.005
2010s	26004.3	11	11	3.7	0.020
2020s	54640.6	306	257	102.0	0.005
2020s	1325.9	108	86	36.0	0.005
2020s	1804.0	14	14	4.7	0.005
2020s	13704.2	14	14	4.7	0.005
2020s	2207.1	13	13	4.3	0.005
2020s	1696.3	9	9	3.0	0.080

Table 3: Combined spatial scan cluster summaries.

Spatial scan tests were also performed in the same manner as the qnn tests with one decade being the case and the other two functioning as the controls. This is justified using the same logic, though limitations in this strategy are taken into account for interpretation.

Table 3 I feel provides a lot of insight into what we've been seeing from our previous results. The 2000s has a single likely cluster, and it contains $470/500$ of its points. The radius for this cluster is the largest out of all the others in the table as well. The 2010s and 2020s both show large clusters as well, but alongside many far smaller clusters. We can see the locations of these

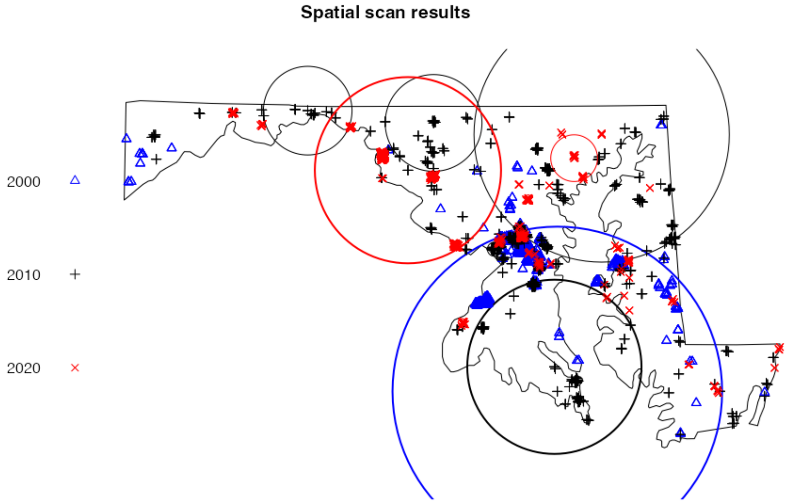


Figure 6: Spatial Scan Clusters by Decade

clusters in Figure 6, though some are not visible due to their small radius and the resolution of the provided image.

The radius doesn't tell the whole story though. When we ignore that and look at cases in each cluster, we see a reinforcement of what was shown in the qnn tests. Both the 2000s and 2020s have clusters with an extremely large number of cases and all of the likely clusters shown in the 2010s have far smaller case counts relative to those. The general trend we've seen overall is that while the 2000s is the most condensed of the decades, the 2010s seems to be the most spread out of the three. The 2020s is more spread out than the 2000s, but still focused primarily on a couple key areas.