

Check Image Skip Analysis

Kyle Nessen

2025-06-12

Background

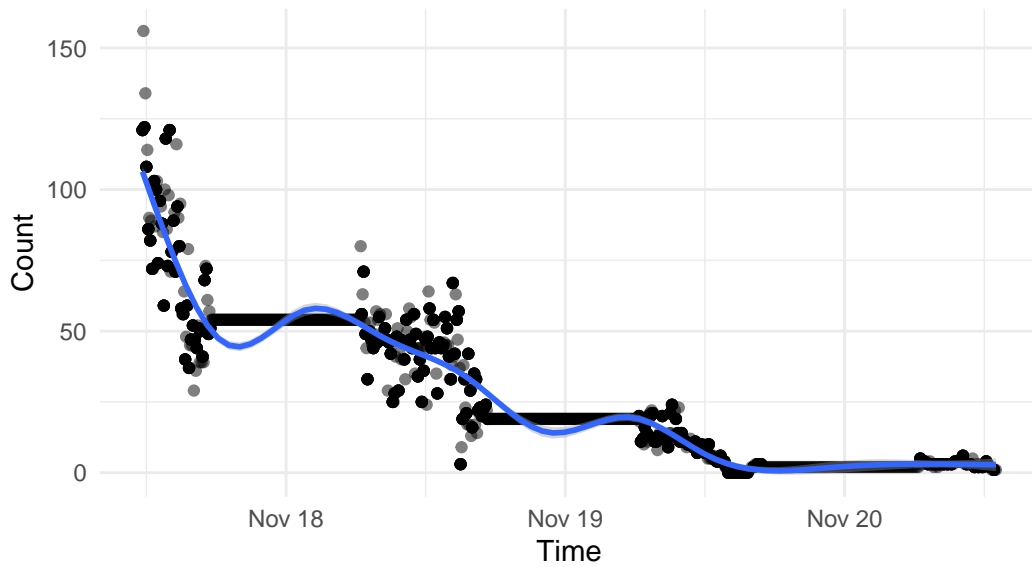
The purpose of this script is that it is taking a long time to classify the images. And part of the problem there is I set the interval at a really high frequency. So for most of the deployments it was 10 minutes. But for the first two that I have classified, it was every five minutes. And what I'm hoping to discover in this notebook is whether I can skip some number of images and have it tell the same story, thereby reducing the total number of images that I have to manually label.

Visualize Existing Data

Here I'll plot the raw monarch count data as a function of time from the few deployments that we have data for. So SC1 and SC2, which were classified by Skylar and Vincent respectively, are complete. This is the total monarch seen. Anything past that is zero butterflies. And SC4 is being done by me and I have not finished this particular deployment yet.

SC1

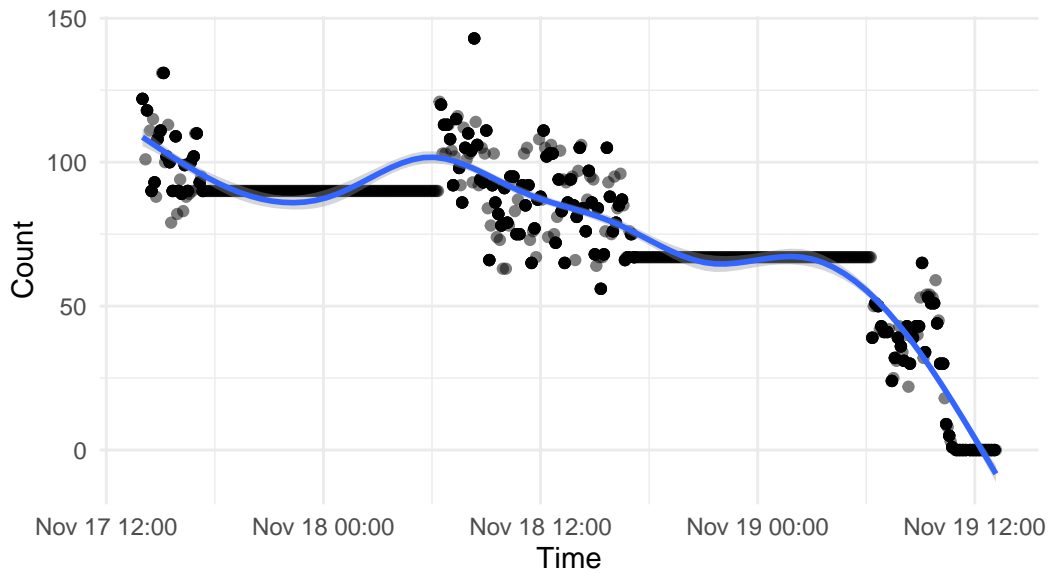
Monarch Counts Over Time – SC1



Source: Trailcam Images

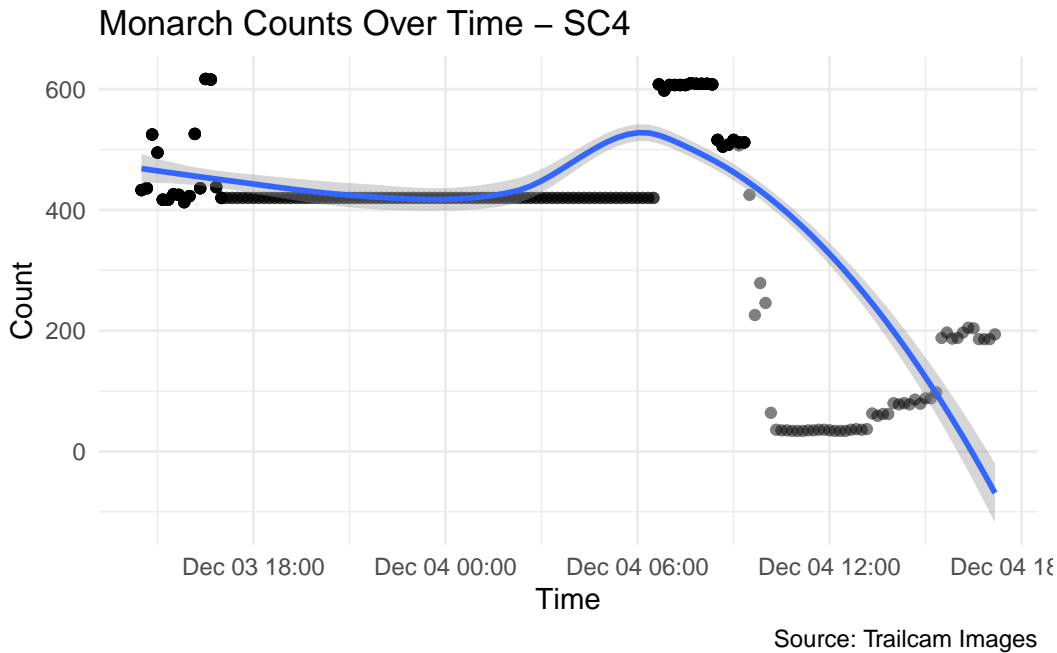
SC2

Monarch Counts Over Time – SC2



Source: Trailcam Images

SC4



It's interesting that there's a lot of spread among the different counts per photo. You can see that most clearly in mine SC4 where I'm using the hundreds category more frequently and that's creating a lot more variance. I think telling a story that may or may not be true. And this like tracks with what I noticed while I was doing the actual labeling leaves would shift and turn and that would change the categories as I was making them. And I was worried I was generating a lot of noise, which seems like may be the case.

For SC1 and SC2, they also have some amount of noise, but it's much more predictable and it looks cleaner. And I think that's partly because they were sampling more frequently. So these photos were every five minutes versus every 10 minutes like in mine. And also the cameras are much closer to the butterflies, so you could actually see the them much more clearly. The scope of the scene was much more reduced as compared to SC4 where you could see more of the landscape. And as we continue along these different deployments, the overall field of view that I took with my camera gets bigger and bigger. So I'm a little concerned that it just might get more noisy as we get further along.

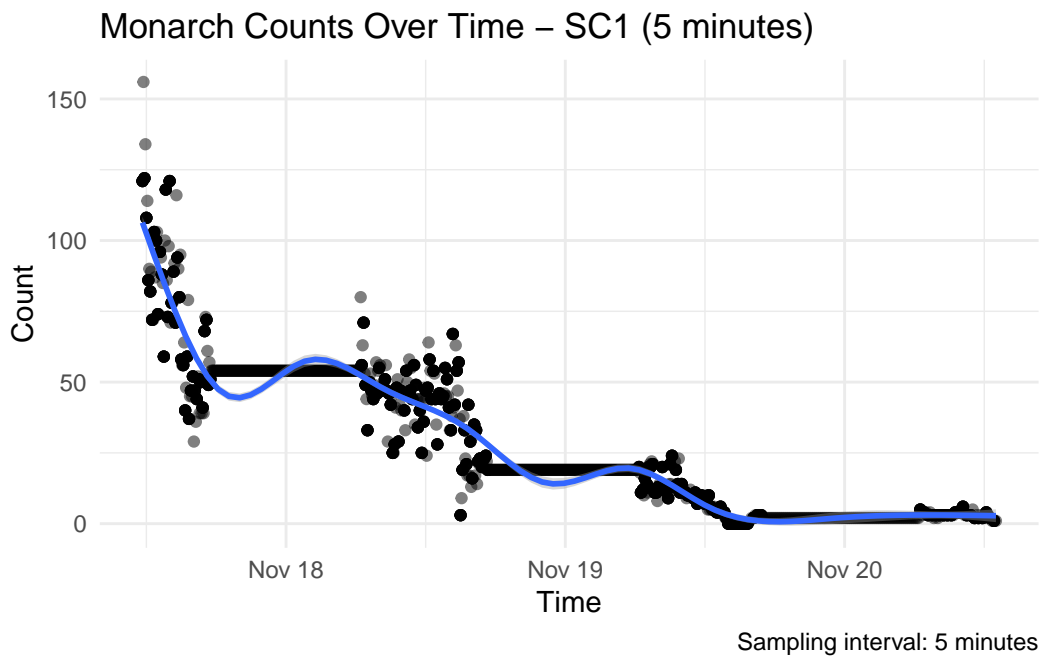
You'll also notice that there are these long series of identical measurements. Those are night-time photos and that came from my explicit directions to the students and for myself which was it's difficult to observe the butterflies at night because the light sort switches it goes from. That is ambient that is more favorable for contrast to a direct flash from the camera itself, which really causes the butterflies to blend in with the leaves. It may be possible. to sort it out, but I think it'd be much more difficult. And my suspicion, at least before seeing these

graphs, was that they don't move at night. But maybe there's some change or maybe it's measurement error. So anyway, I think for what I have now, I'm probably just going to remove the night photos and remove them from future graphs. And it's worth a discussion whether this is something we want to try to extract information from or if it's simply just not going to be part of our study.

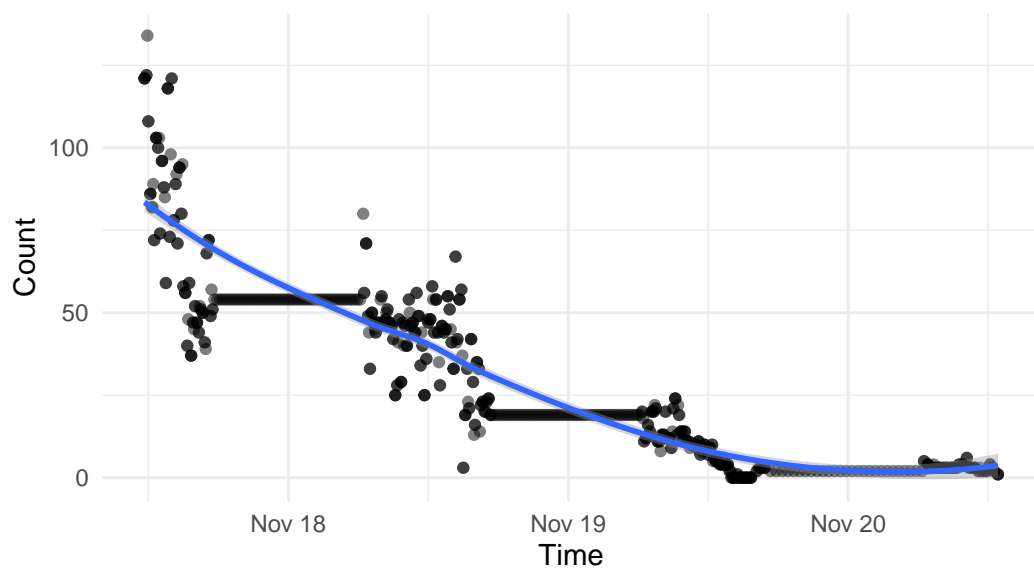
Trying different intervals

Let's analyze how different sampling intervals affect our ability to detect patterns. Here, I'll sample the data at 5-minute (original), 15-minute, 30-minute, and 60-minute intervals.

SC1 with different sampling intervals

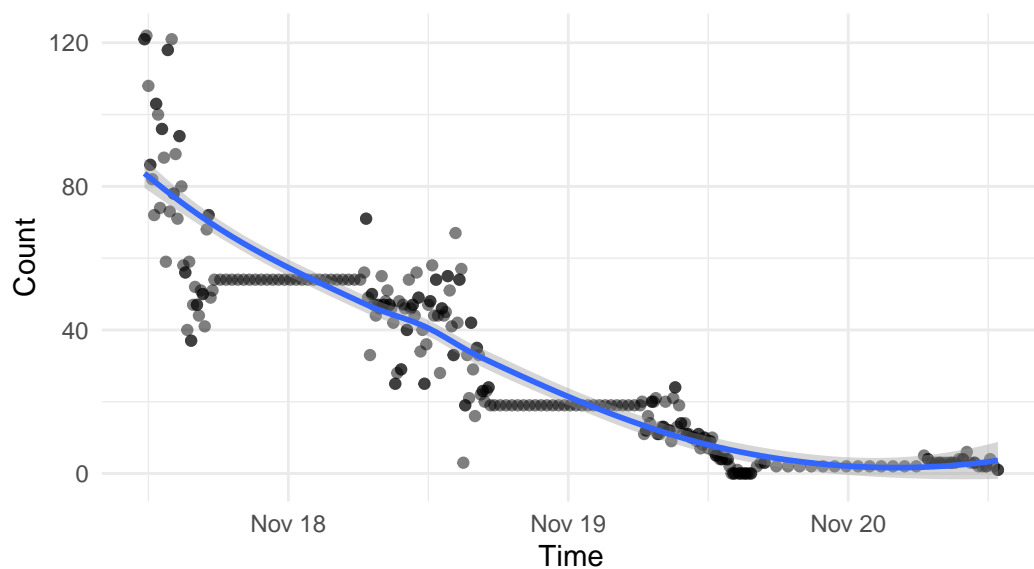


Monarch Counts Over Time – SC1 (15 minutes)



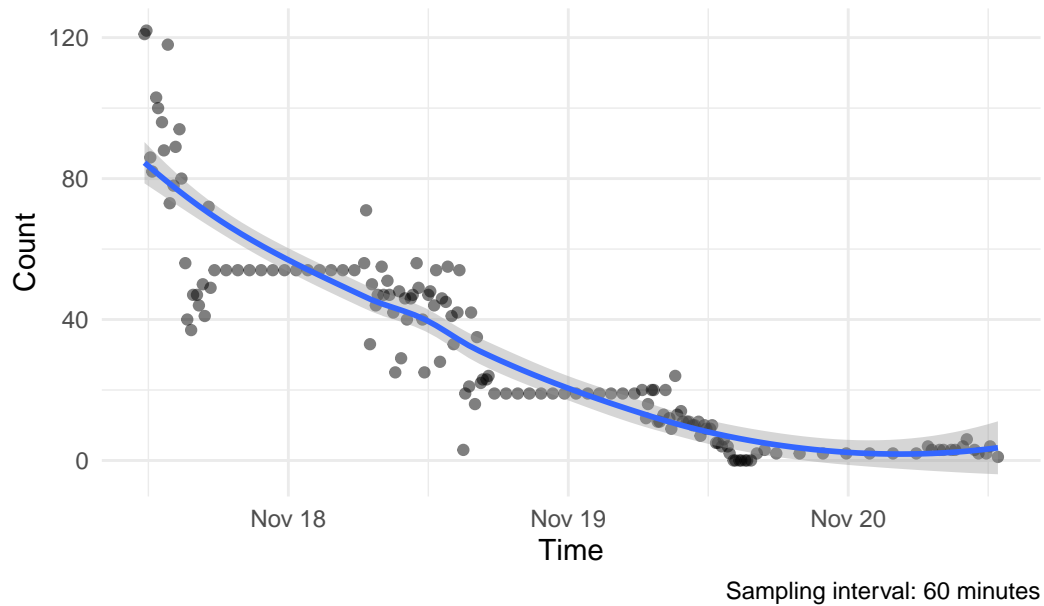
Sampling interval: 15 minutes

Monarch Counts Over Time – SC1 (30 minutes)



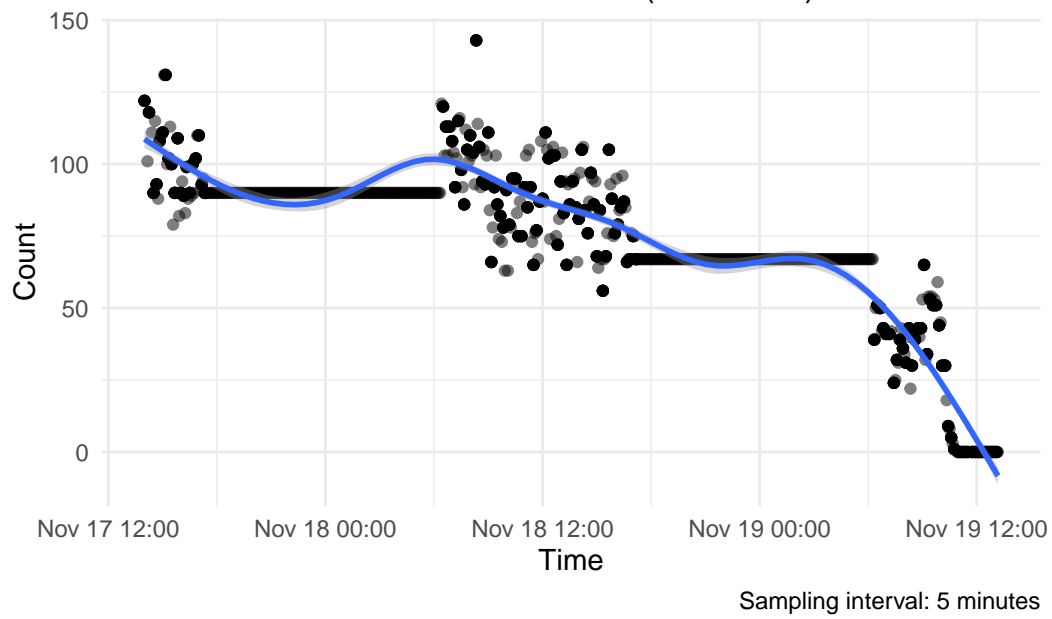
Sampling interval: 30 minutes

Monarch Counts Over Time – SC1 (60 minutes)

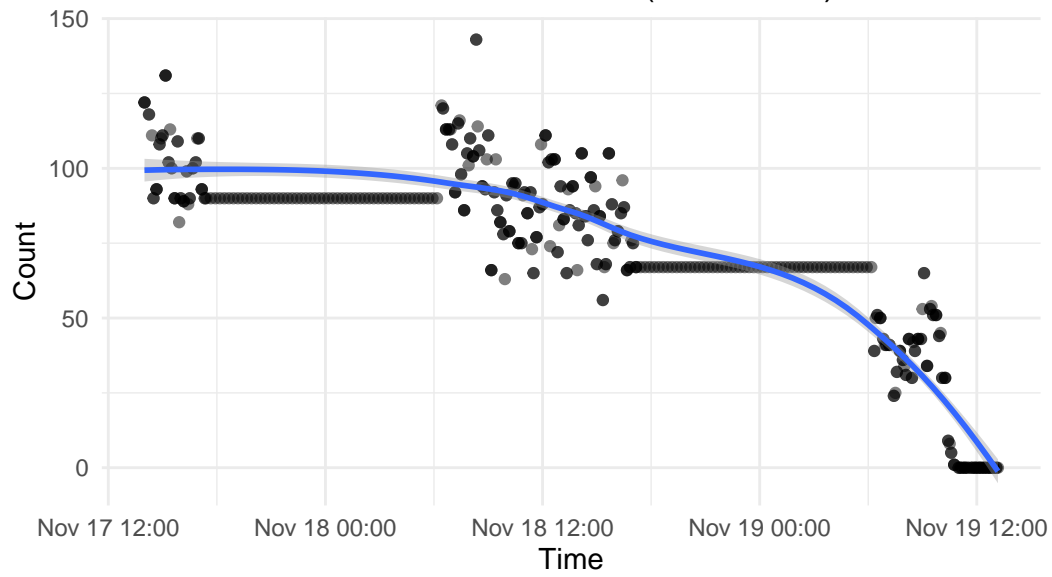


SC2 with different sampling intervals

Monarch Counts Over Time – SC2 (5 minutes)

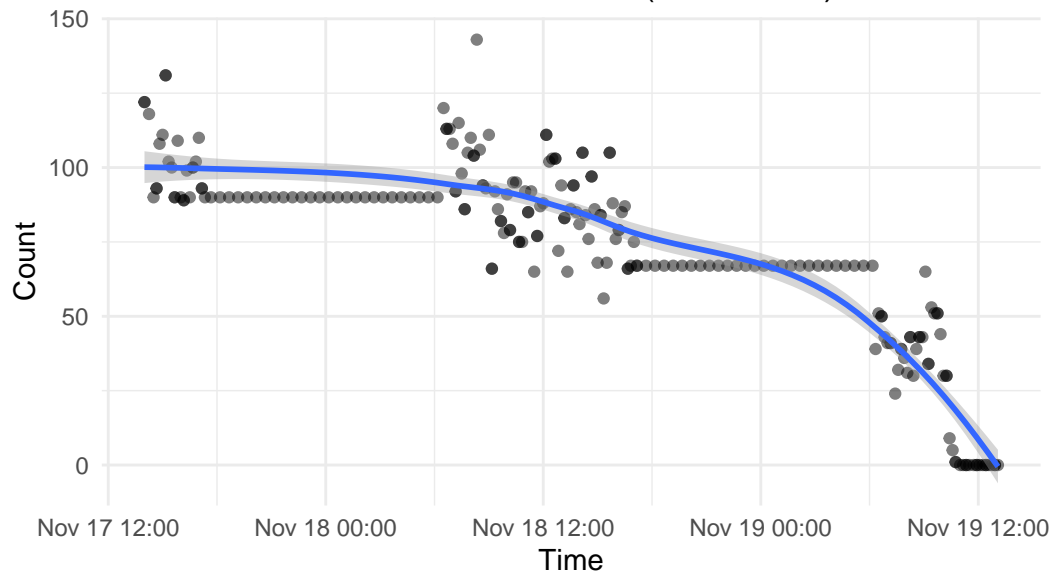


Monarch Counts Over Time – SC2 (15 minutes)



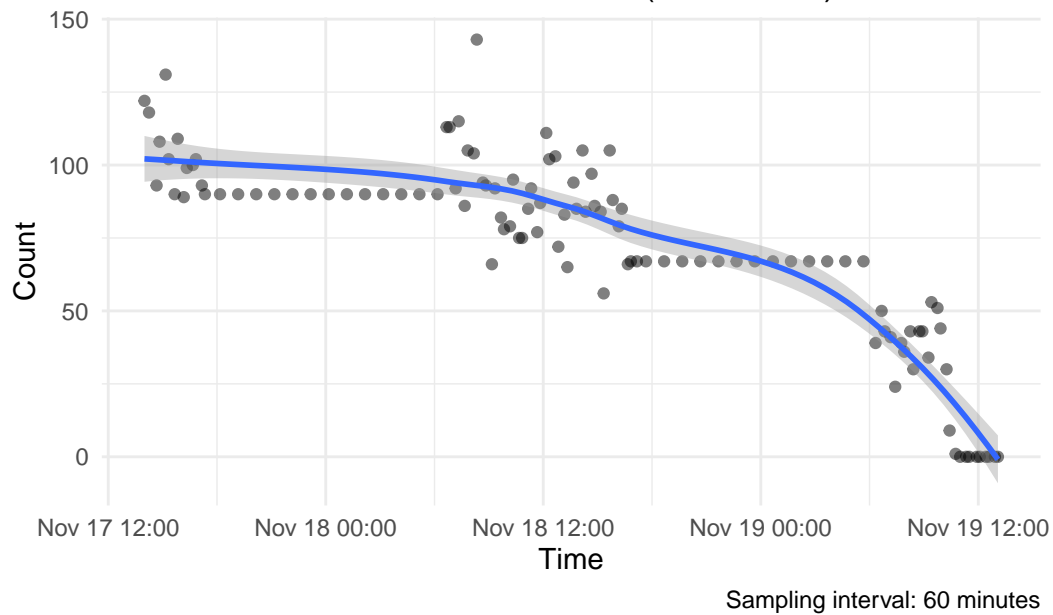
Sampling interval: 15 minutes

Monarch Counts Over Time – SC2 (30 minutes)



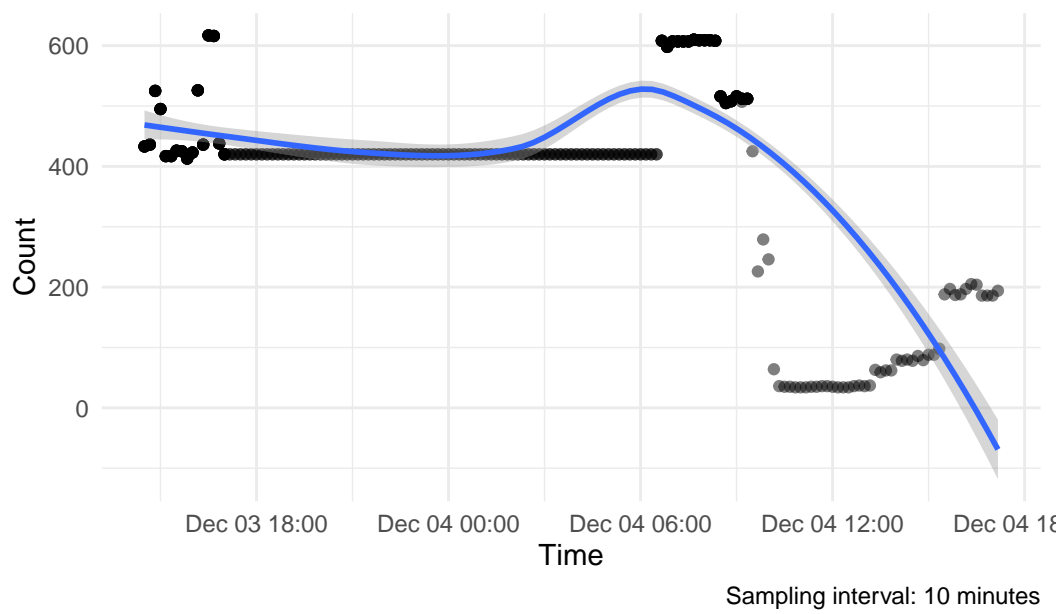
Sampling interval: 30 minutes

Monarch Counts Over Time – SC2 (60 minutes)

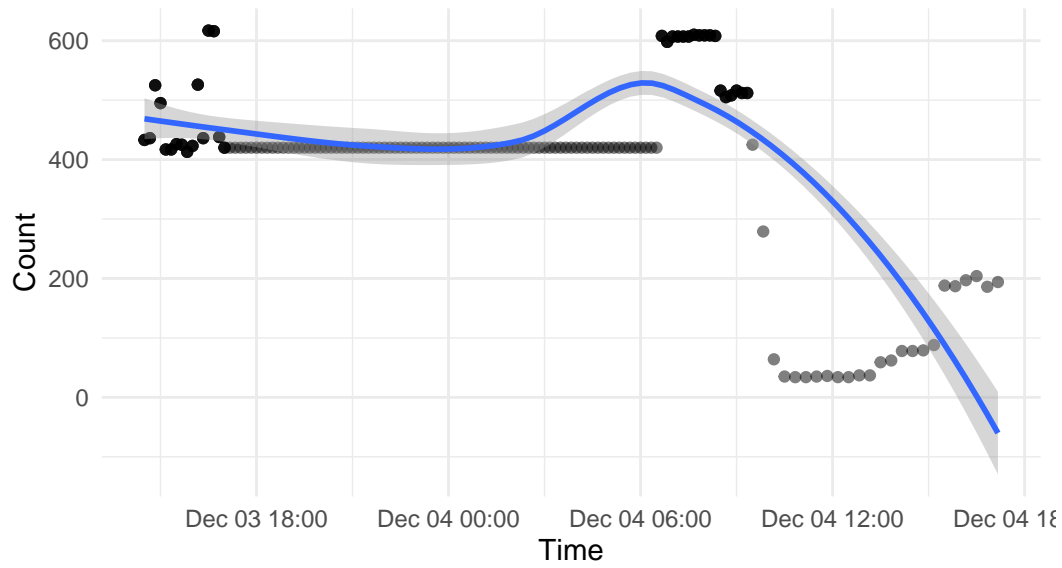


SC4 with different sampling intervals

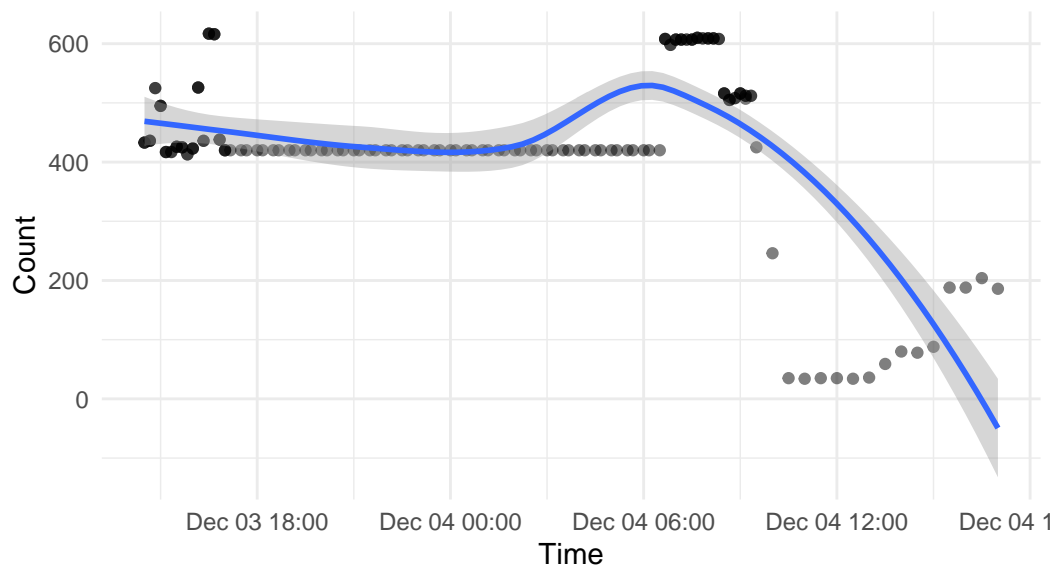
Monarch Counts Over Time – SC4 (10 minutes)

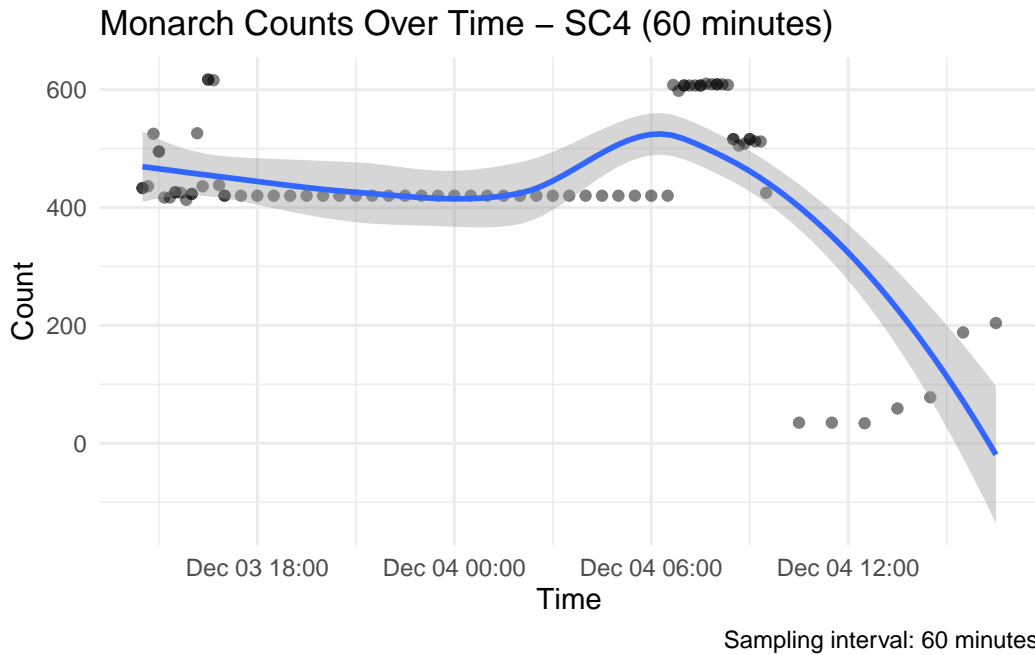


Monarch Counts Over Time – SC4 (20 minutes)



Monarch Counts Over Time – SC4 (30 minutes)





It's encouraging to see that reducing the number of images from, say, every five or ten minutes to every hour doesn't seem to change the trend line of the data itself. I think this is a viable way to reduce the number of images that need to be manually classified.

Analysis of Different Sampling Intervals

To better compare the sampling intervals, let's examine how well each sampling rate captures the key patterns in the data:

Table 1: SC1 - Summary Statistics by Sampling Interval

| interval | mean_count | median_count | min_count | max_count | std_dev | n_samples |
|----------|------------|--------------|-----------|-----------|----------|-----------|
| 5 min | 32.00868 | 22.0 | 0 | 156 | 27.73242 | 1844 |
| 15 min | 32.18537 | 23.0 | 0 | 134 | 28.00142 | 615 |
| 30 min | 32.24026 | 23.0 | 0 | 122 | 28.02978 | 308 |
| 60 min | 32.01948 | 22.5 | 0 | 122 | 28.22169 | 154 |

Table 2: SC2 - Summary Statistics by Sampling Interval

| interval | mean_count | median_count | min_count | max_count | std_dev | n_samples |
|----------|------------|--------------|-----------|-----------|----------|-----------|
| 5 min | 72.52896 | 83.5 | 0 | 143 | 32.13616 | 1312 |
| 15 min | 72.65068 | 83.0 | 0 | 143 | 32.25937 | 438 |

| interval | mean_count | median_count | min_count | max_count | std_dev | n_samples |
|----------|------------|--------------|-----------|-----------|----------|-----------|
| 30 min | 72.70776 | 83.0 | 0 | 143 | 32.07283 | 219 |
| 60 min | 72.40909 | 83.5 | 0 | 143 | 32.81930 | 110 |

Table 3: SC4 - Summary Statistics by Sampling Interval

| interval | mean_count | median_count | min_count | max_count | std_dev | n_samples |
|----------|------------|--------------|-----------|-----------|----------|-----------|
| 10 min | 445.4863 | 420.0 | 34 | 617 | 138.1609 | 473 |
| 20 min | 445.3502 | 420.0 | 34 | 617 | 138.4727 | 237 |
| 30 min | 446.7595 | 421.5 | 34 | 617 | 137.0585 | 158 |
| 60 min | 446.5570 | 423.0 | 34 | 617 | 136.2667 | 79 |

Looking at it in a different way from a descriptive statistics perspective, it seems like the number of samples or the mean count and the standard deviation looks really good compared to different sampling intervals. This is worth considering further.