General features of sampling design

Jens Åström

2018-10-04

Replikering og kilder til variasjon

Arealrepresentativitet – rutenettverk kontra habitatkartlegging

Generelle ting, men også eksempler på tilfeldige rutevalg. Eksempelkart på lands, region og lokal nivå. Utfordringer ved automatiske valg. Mikromiljøer, etc.

Antall habitater/regioner/forklaringsvariabler å estimere effekt av

Locations and number of visits

Variasjon mellom plasser, mellom år innen plasser, størrelser på nedganger og variasjon.

The simplest design for time series data is to pick a number of sites that you revisit a number of times, revisiting all sites each time. Revisiting is needed to identify time trends, and the weaker the time trend, the more years have to pass before you can detect a difference.

Figure 1 shows a hypothetical example of 5 sites that differ in their overall values, but follow the same time trend. Since they follow the same trend, you don't have to visit many locations to correctly identify the underlying common trend. If you like to know estimate the overall abundance level, however, you might want to sample more locations to account for the variation in baseline levels for the different sites.

```
set.seed(1234)

test5yearsTrend <- createOccNorm(map10km, intercept = 10, sigmaFylke = 0, sigmaKommune = 0,
    sigmaGrid = 0.2, nYears = 5, interceptTrend = -0.05, sigmaFylkeTrend = 0,
    sigmaKommuneTrend = 0, sdInterceptTrend = 0, sortGrid = F, sortFylke = F,
    sortKommune = F)
```

```
test5yearsTrend$map %>% filter(ssbid %in% sample(map10km$ssbid, 5)) %>% select(ssbid,
    norm, year) %>% ggplot(.) + geom_smooth(aes(x = year, y = norm, color = ssbid),
    stat = "summary", fun.y = "mean", lwd = 2) + ylab("Abundance")
```

But the locations might not follow the same time trend so closely, and in reality they will always differ to some degree. If the locations have varying trends, as in figure 2, it becomes harder to identify a potential underlying, common trend. In this case, we need to visit more individual locations to be able to identify the underlying trend.

```
set.seed(1234)

varying5yearsTrend <- createOccNorm(map10km, intercept = 10, sigmaFylke = 0,
    sigmaKommune = 0, sigmaGrid = 0.2, sigmaSurvey = 0, nYears = 5, interceptTrend = -0.1,
    sigmaFylkeTrend = 0, sigmaKommuneTrend = 0, sdInterceptTrend = 0.05, sortGrid = F,
    sortFylke = F, sortKommune = F)
```

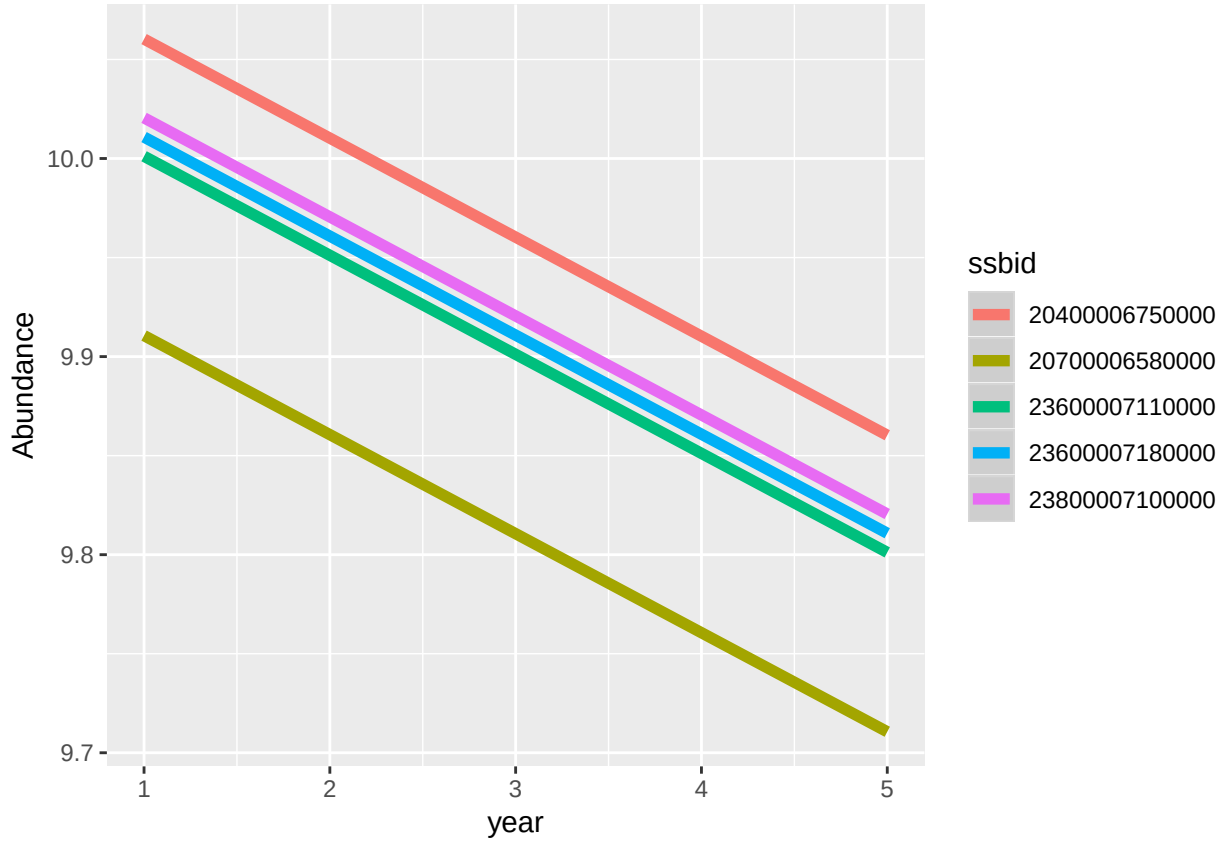

Figure 1: Hypothetical time trends for 4 sites (ssbid grid cells) that follow the same time trend but have diffent overall levels.

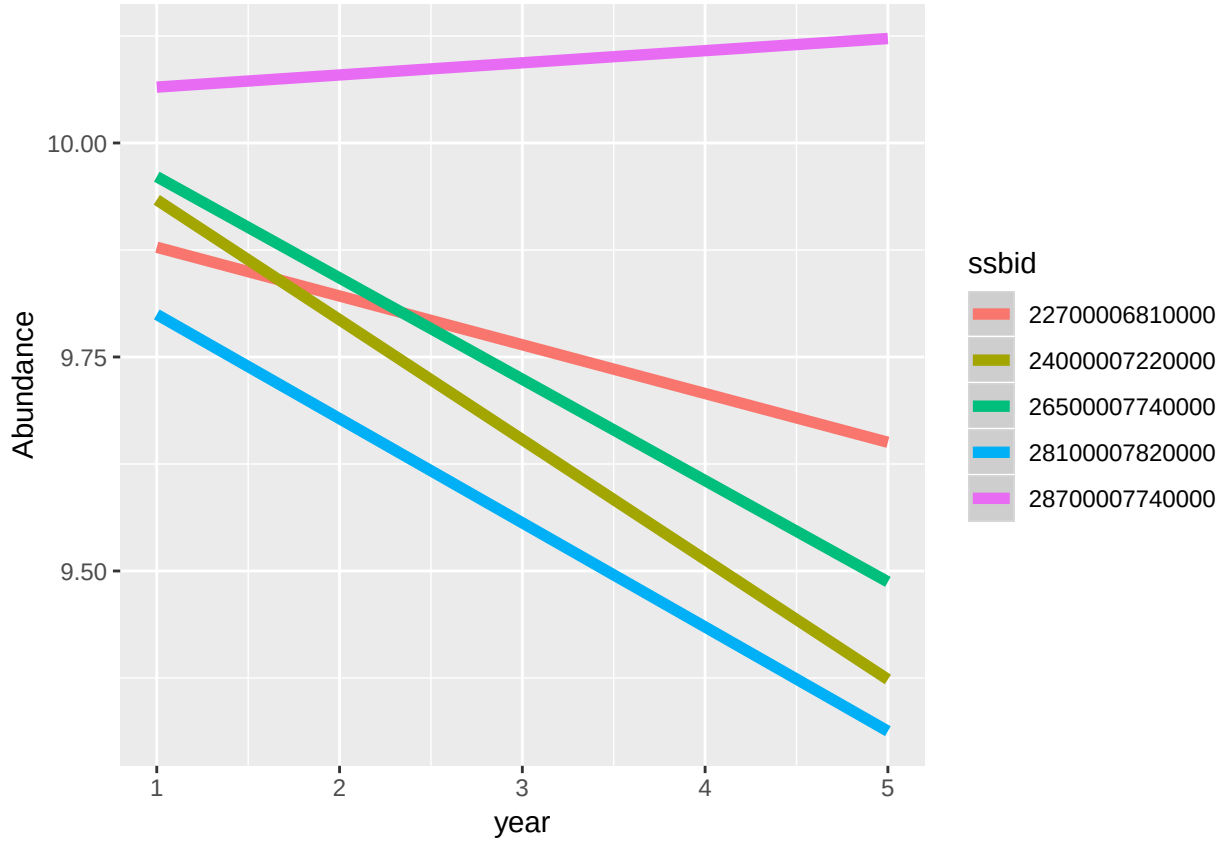

Figure 2: Hypothetical time trends for 4 sites (ssb grid cells) that follow the same underlying time trend but have diffent overall levels, and also vary in individual time trends.

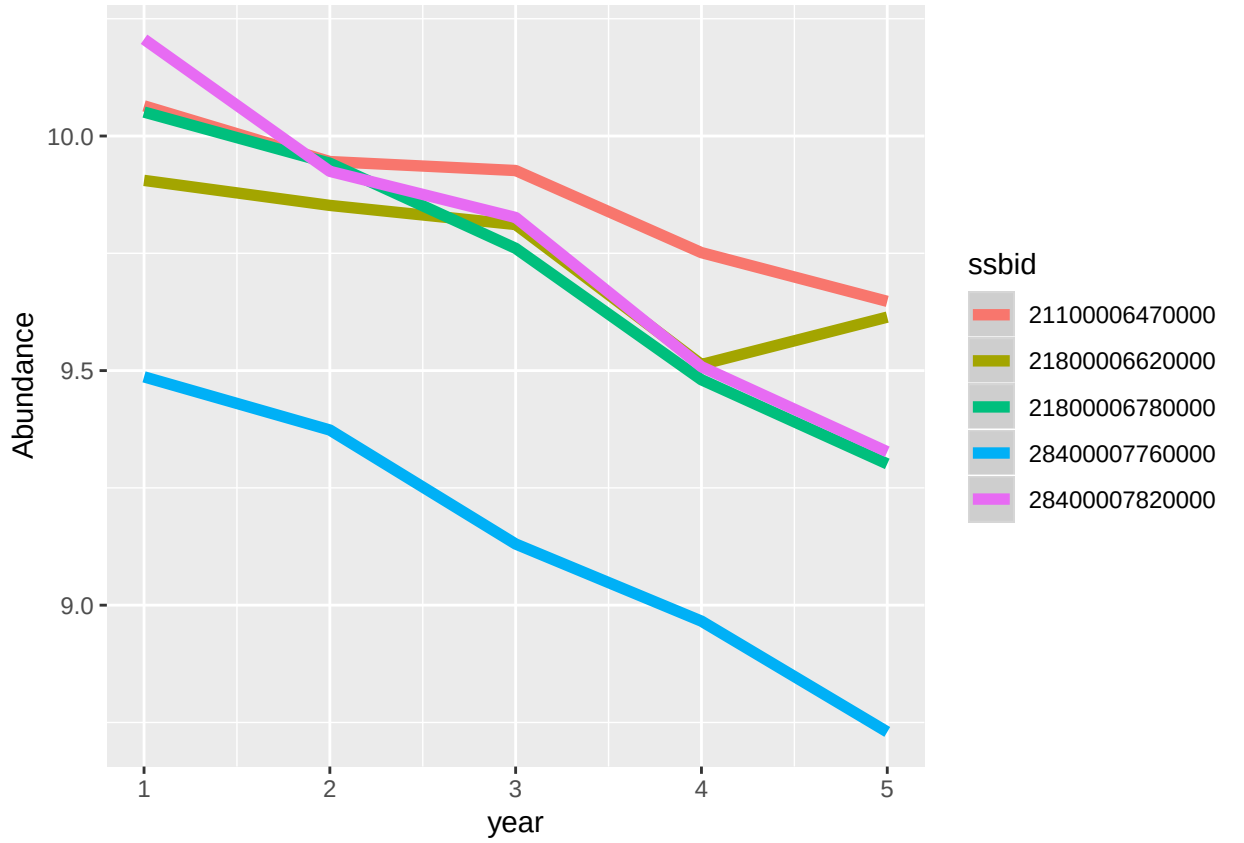

Figure 3: Hypothetical time trends for 4 sites (ssb grid cells) that follow the same underlying time trend but have diffent overall levels, and also vary in individual time trends and have vary around their individual time trends.

On the other hand, there is always some backround random variation for the observations, so that the time trends might not look so straight (figure 3). In these cases, we need to sample each location several times to accurately estimate its time trend. If there we no such variation, we could actually just survey each location twice to estimate the slope. In reality, of course, there is always some additional variation.

```
set.seed(1234)

sigmaVarying5yearsTrend <- createOccNorm(map10km, intercept = 10, sigmaFylke = 0,
    sigmaKommune = 0, sigmaGrid = 0.2, sigmaSurvey = 0.1, nYears = 5, interceptTrend = -0.1,
    sigmaFylkeTrend = 0, sigmaKommuneTrend = 0, sdInterceptTrend = 0.05, sortGrid = F,
    sortFylke = F, sortKommune = F)
```

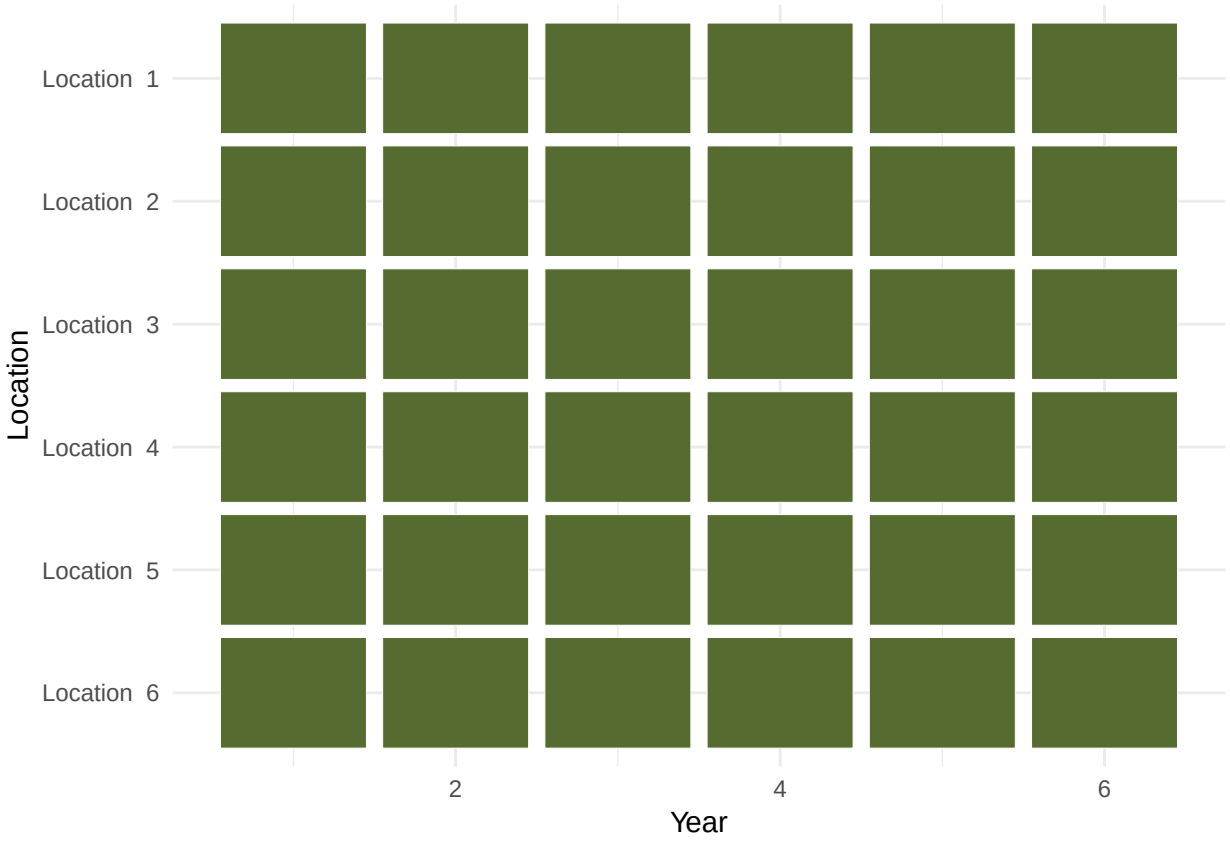

Figure 4: 6 locations visited each year. 6 samples per location for a total of 36 samples.

Staggered surveying and variation in time and space

A typical survey program will have a limited yearly budget or work capacity that determine how many sites or samples that can be processed. This means you end up with a yearly capacity of sites to cover. But this yearly capacity can be put to use in different ways. Should we for example revisit each site each year, or should we alternate the sites and thereby span over a wider range of sites? We could for example choose to revisit each location every third year, and span over 3 times as many sites. The penaly is that you only get a third of the replicates per site.

The best choice here will depend on the relative size of the variation between sites and the yearly variation. If sites are more or less the same, we don't have to visit many different sites to accurately represent the underlying population. If they vary a lot, any small subset of sites risk being unrepresentative of the underlying population and we should cover a wider range of sites to capture this variation. If there is little between-year variation, any visit vill be representative of the overall level or trend for each location, and we require fewer samples per site. But when between-year variation is higher, any visits will be influenced by random fluctuations and we need many visits per site to accurately capture its mean level or trend.

This means we end up with a trade-off between the number of visits per site and the number of sites to visit, given that we can only visit so many sites each year. Figures 4, 5, 6, and 7 shows the various ways we could allocate a yearly capacity of surveying 6 locations a year, for 6 years, while keeping the number of times each location is visited the same for all locations (a balanced data set).

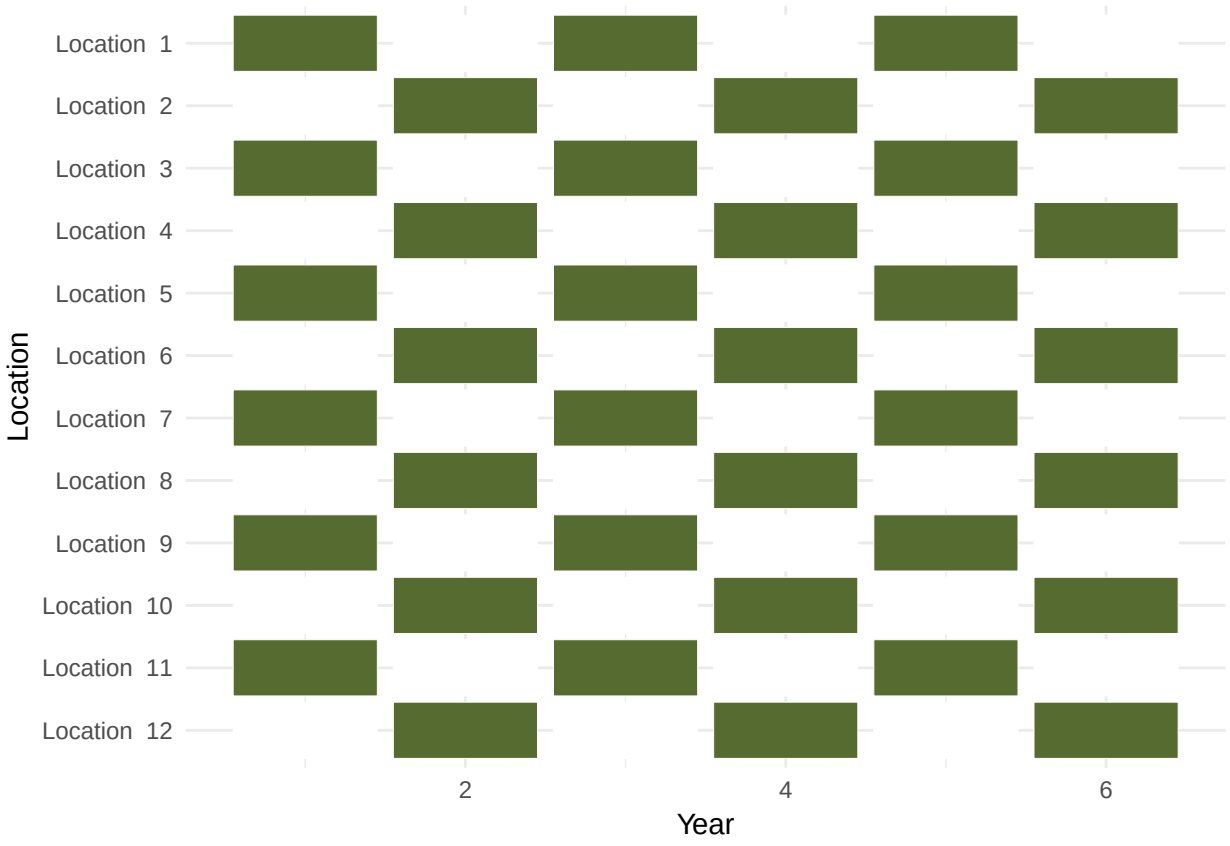

Figure 5: 12 locations visited every other year. 3 samples per location for a total of 36 samples.

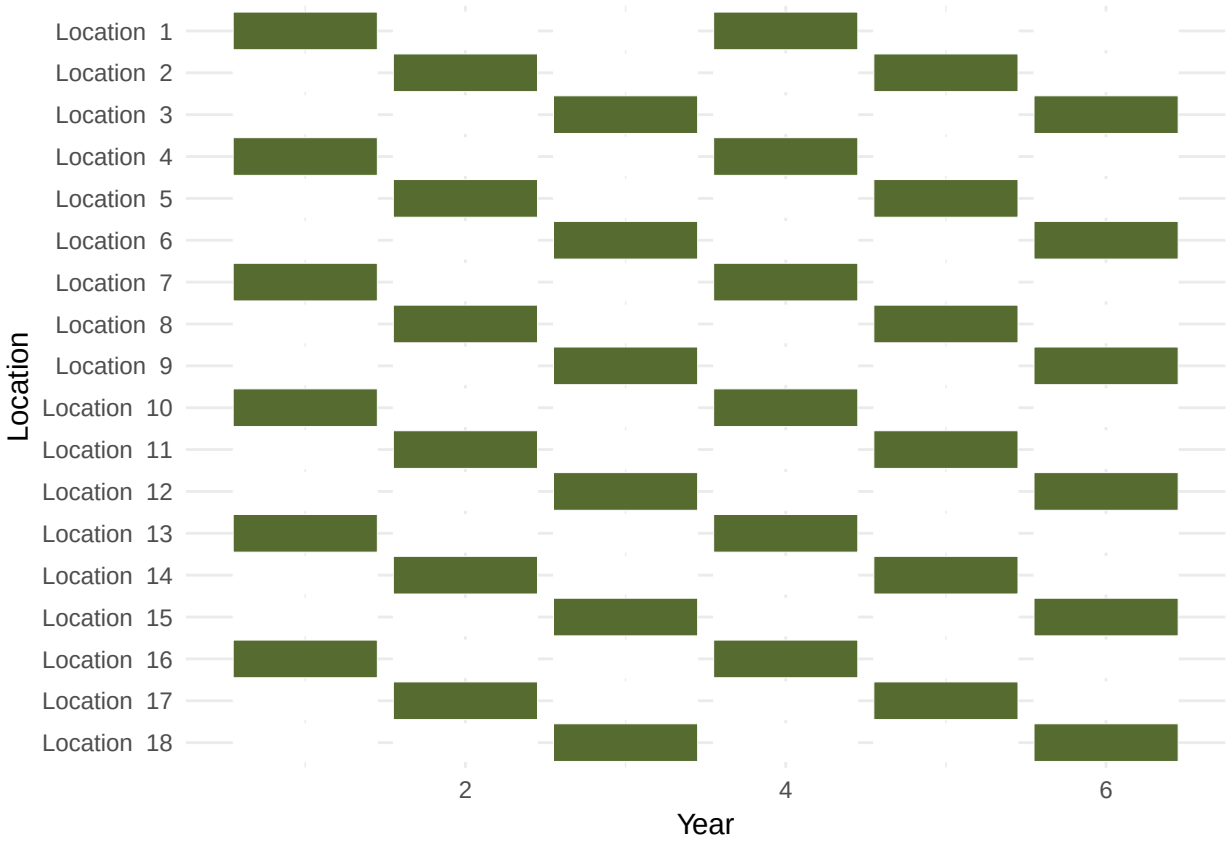

Figure 6: 18 locations visited every third year. 2 samples per location for a total of 36 samples.

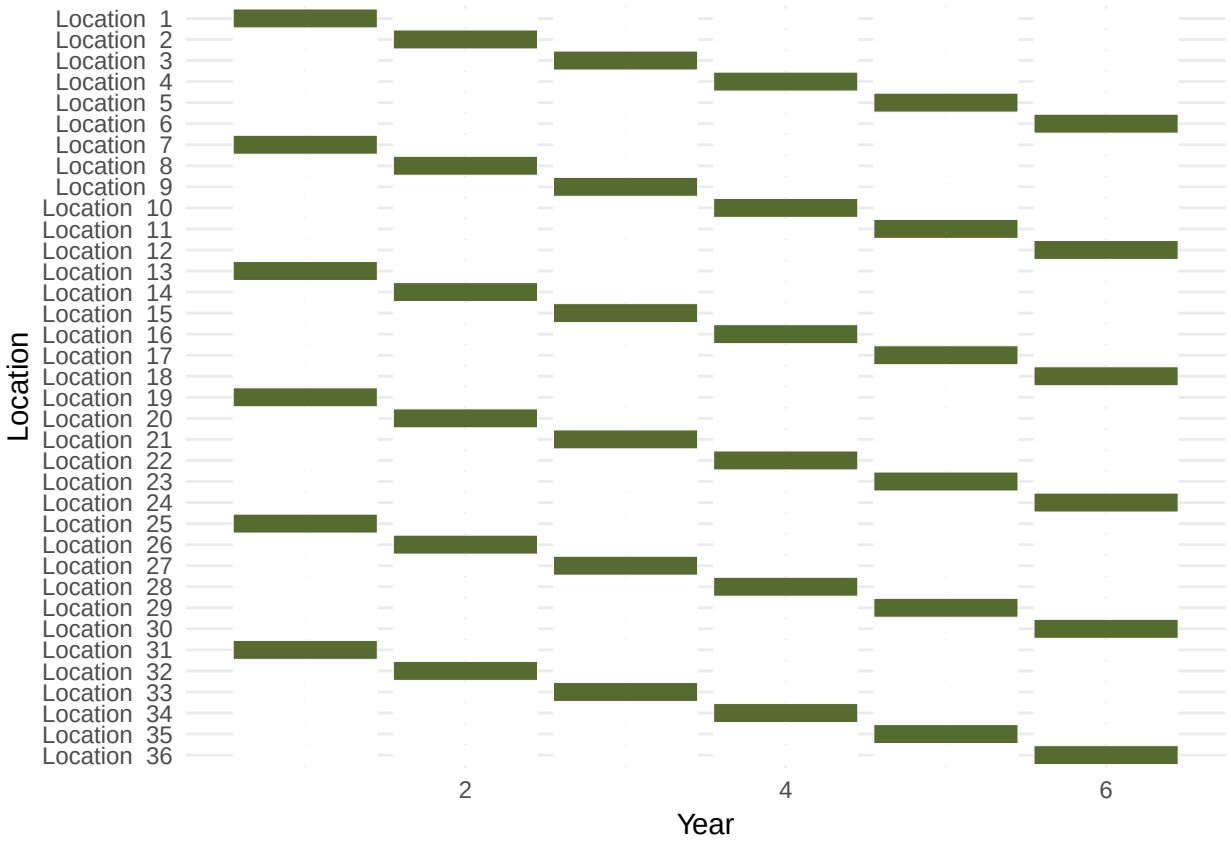

Figure 7: 36 locations visited once. 1 sample per location for a total of 36 samples.

As long as we keep to a balanced design, where we revisit each site the same number of thimes, there are some simple rules of how such survey schemes can be set up. For simplicity, we limit our explorations to balanced cases.

We can define these parameters:

$$\begin{aligned} Sampling\ capacity\ each\ year &: a \\ Total\ timespan\ of\ survey &: T \\ Resampling\ interval\ per\ locality &: t \\ Number\ of\ localities &: l \\ Total\ number\ of\ samples &: s \\ Replicates\ per\ location &: r \end{aligned}$$

The number of localities that we can survey is then

$$l = t * a$$

, which can be maximised by maximising t to $t = T$. The minimum timespan between revisits in a location is

$$t = T/a$$

, which also minimizes the total number of locations. However, the number of samples per location r is at the same time maximized as this follows

$$r = T/t$$

. The total number of samples simply is

$$s = T * a$$

. Lastly, balanced survey schemes relies on the total survey length being evenly divisible by the yearly total survey capacity, i.e. as long as $T \bmod a = 0$.

The possible survey schemes from relationships can be calculated by the `sampleAlternatives` for convenience, which creates an object with its own plot method. Figure 8 show how the number of possible locations and replicates (visits) per location depend on the yearly survey capacity and the time lag between visits to each location. As the figure shows, you can survey a staggering amount of locations in 40 years, if you are willing drastically lower the number of revisits. The trade-off between the number of locations and the number of visits per location is pretty sharp, and it is difficult to achieve high numbers in both, even with a high yearly capacity.

```
testSampleAlt <- sampleAlternatives(maxTime = 40, maxCapacity = 400, stepsCapacity = 100)

# class(testSampleAlt)

plot(testSampleAlt, color = "resampleTime", allTicks = F)
```

It is important to keep in mind the time span on the survey effort, within which we expect to evaluate the results. For most programs, it would be unreasonably long to have to wait 40 years to evaluate possible time trends. Figures 9 and 10 show a more reasonable time span of 9 years, with low and high yearly capacities, respectively.

```
plot(sampleAlternatives(maxTime = 9, maxCapacity = 3, stepsCapacity = 1), color = "totSamples")
```

```
plot(sampleAlternatives(maxTime = 9, maxCapacity = 300, stepsCapacity = 60),
    color = "totSamples", allTicks = F)
```

For an active real world example, we can plot the situation for the bumblebee and butterfly survey. This survey is done each year in 18 locations for each of three regions, together comprising 54 locations. The

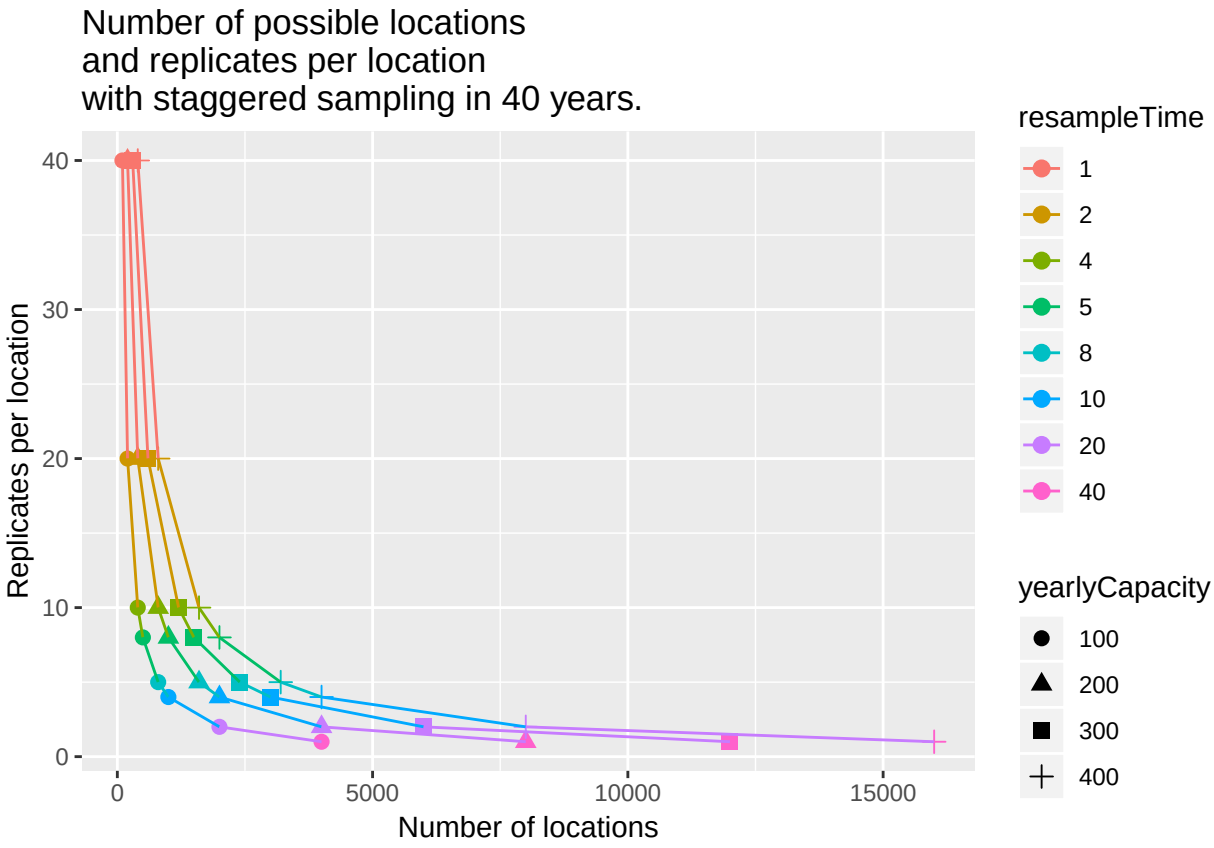

Figure 8: The relationship between number of surveyed locations and replicates per location for different survey staggering times and yearly survey capacities.

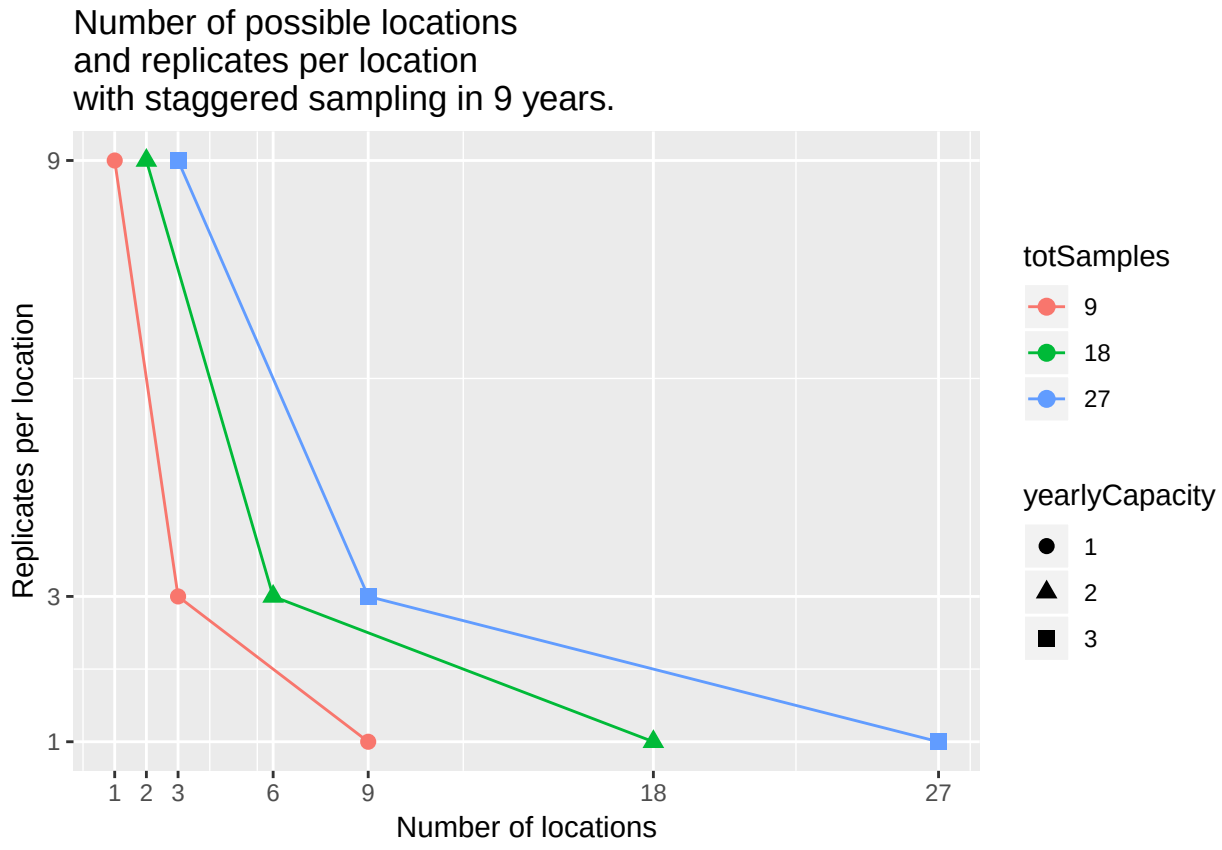

Figure 9: Possible survey schemes with low yearly capacity over 9 years. Colors represent the total number of samples, and symbol types represent different yearly capacities.

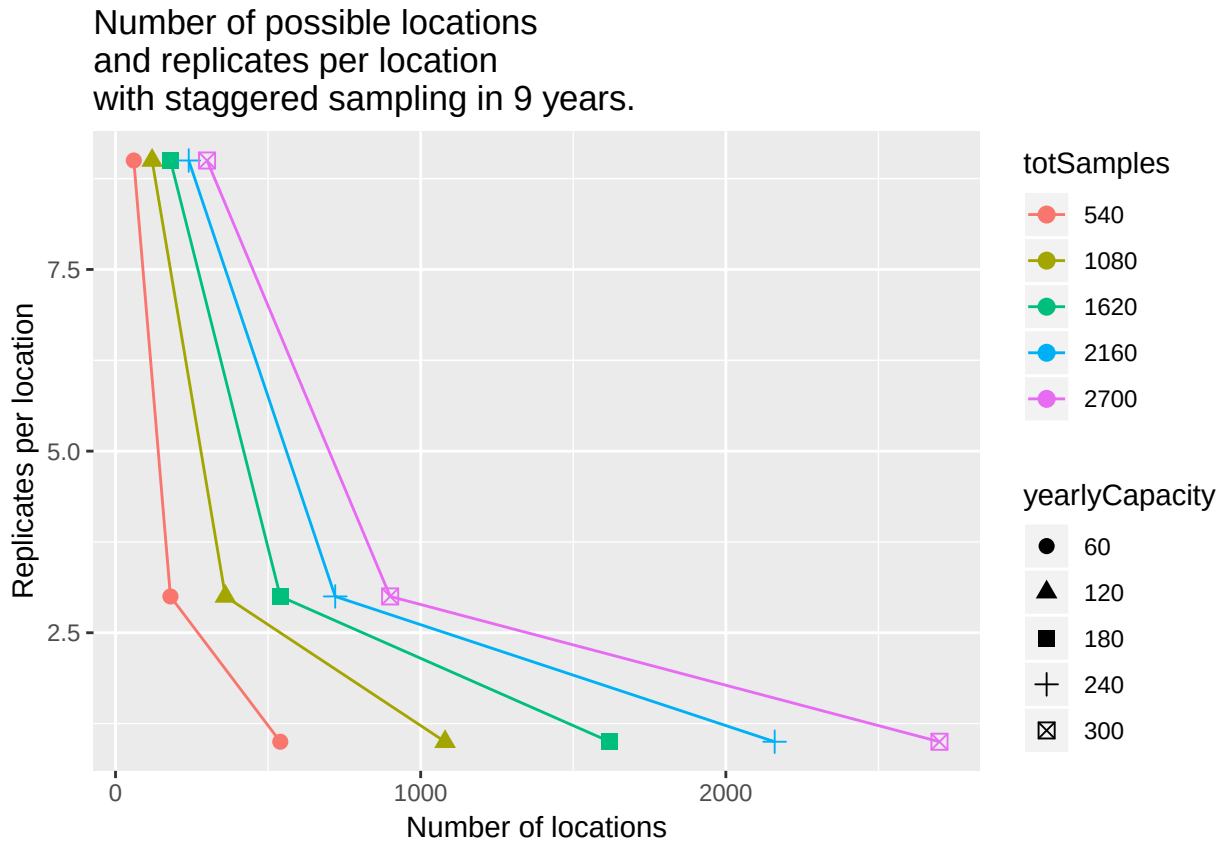

Figure 10: Possible survey schemes with high yearly capacity over 12 years. Colors represent the total number of samples, and symbol types represent different yearly capacities.

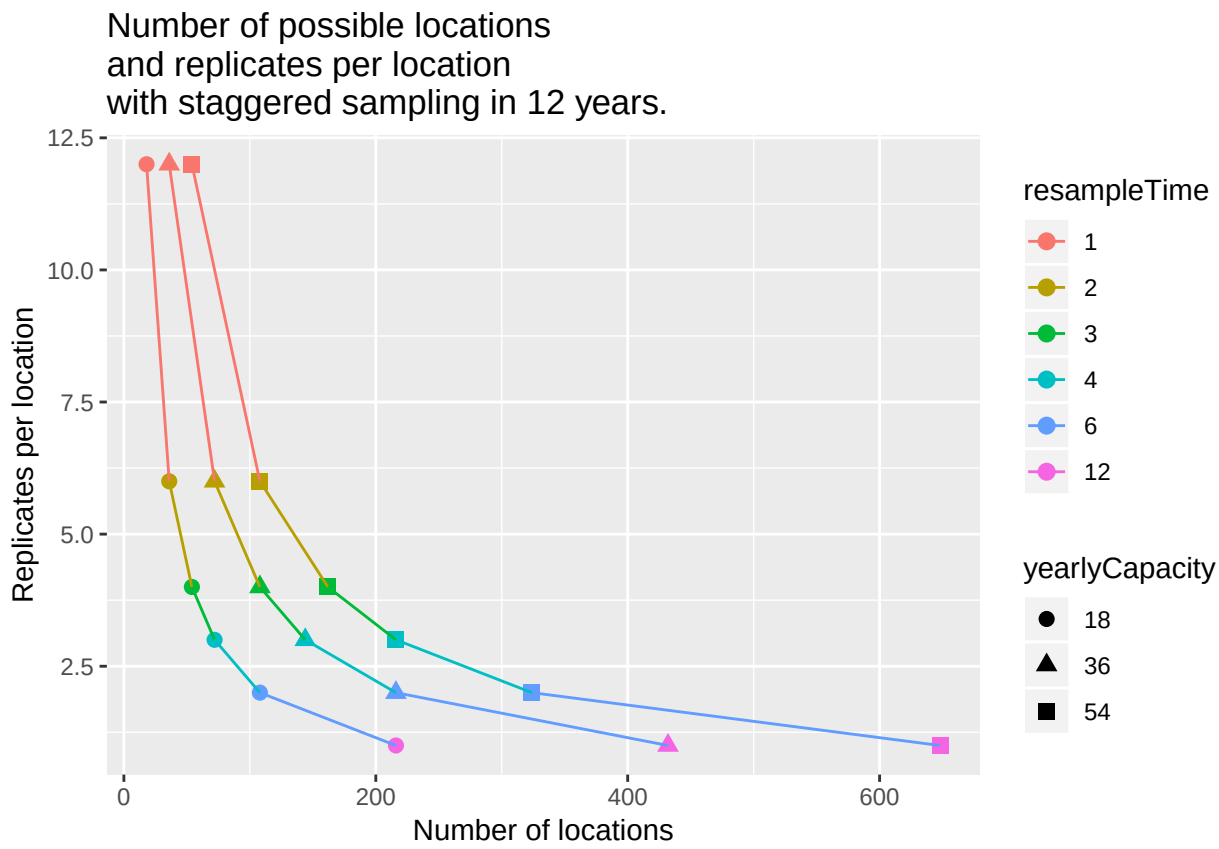

Figure 11: Possible survey schemes with a the same yearly capacity as the bumblebee and butterfly survey. This survey is done each year in 18 locations for each of three regions, together comprising 54 locations. The different lines/symbols thus represent including 1, 2, or all three regions in an analysis.

survey has been active since 2009 and thus exemplifies the results of potential schemes in the year 2021. Today, we revisit the same locations each year, resulting in 18 locations per region. But if we where to revisit each site only every third year, we could visit 54 locations 4 times in the same time span. Or for all regions, instead of visiting 54 locations 12 times, we could visit 162 locations 3 times each, figure 11.

```
plot(sampleAlternatives(maxTime = 12, maxCapacity = 54, stepsCapacity = 18),
    color = "resampleTime", allTicks = F)
```

The optimal strategy will vary depending on the various sources of variation. A high yearly variability increases the need for multiple samples per location, while a high variability between sites increases the need to visit many sites. If we are interested in the absolute numbers in each region, and this varies by sites, we must visit more sites per region. If we are interested in the trends in each region, and these trends vary by site, we also need to visit more sites per region. It is not easy working out the optimal strategy, or calculate the statistical power of various survey schemes analytically. We need to simulate various situations and test out different strategies.

Overvåking av sjeldne og kryptiske arter

Konsekvenser av lav sannsynlighet for oppdagbarhet og tilstedeværelse.