

# Format Cetaceans Data

## Contents

<b>Load occurrence dataset</b>	<b>2</b>
<b>Load the list of occurrences with morphological information</b>	<b>2</b>
<b>Explore the dataset</b>	<b>2</b>
Repartition through time . . . . .	2
Full fossil record . . . . .	2
Uniform subsampling . . . . .	4
Repartition among accepted ranks . . . . .	4
Pie chart . . . . .	5
Time repartition by rank . . . . .	5
BIAS CORRECTION N°1: Remove all Late Pleistocene and Holocene occurrences . . . . .	6
Redundancy of occurrences with the same accepted name . . . . .	6
<b>Mitigate the abundance bias</b>	<b>7</b>
Compute the occurrence density . . . . .	7
Count occurrences by accepted name . . . . .	7
BIAS CORRECTION N°2: Exclude occurrences with highly uncertain dating ( $> 10\text{My}$ ) . . . . .	8
Stratigraphic ranges = unique time interval combining age uncertainties of occurrences of the same taxon (excluding the most uncertain ones) . . . . .	11
Density distributions . . . . .	11
Correlation between occurrence density and age . . . . .	12
BIAS CORRECTION N°3: Subsample occurrences by homogenizing the combined occurrence density . . . . .	13
Compare densities by accepted name count (species only) . . . . .	13
Impact of correcting subsampling on density distributions (species only) . . . . .	14
Impact of subsampling on occurrences repartition (species only) . . . . .	14
Compare with a Poisson sampling process . . . . .	16
BIAS CORRECTION N°3bis: Subsample occurrences by aggregating similarly identified occurrences in each collection . . . . .	19
Wrong track: Aggregate similarly identified occurrences in each collection . . . . .	19
Aggregate similarly identified occurrences in each formation . . . . .	20
Wrong track: Aggregate occurrences without formation by country + early interval . . . . .	22
Aggregate occurrences without formation by geoplate + early interval . . . . .	25
Check that the sampling methods do not introduce biases in the repartition between Odontoceti and Mysticeti . . . . .	27
<b>Faster genus-level analysis</b>	<b>28</b>
<b>Conclusions</b>	<b>30</b>

Creation - jeremy.andreoletti@ens.fr - 13/04/2020

This notebook aims at exploring the Cetacean occurrence dataset from the PaleoBiology DataBase (PBDB), looking for potential biases invalidating to our model assumptions and how to mitigate them. Subsampling this dataset will allow (1) to reduce the computational burden of our analysis and (2) to compensate for the biases mentioned above.

## Load occurrence dataset

## Load the list of occurrences with morphological information

These are the specimens that are already included in the tree thanks to their morphological characters, and must therefore be removed from the occurrence dataset.

Combine these taxa with extant taxa to get all the species included in the tree.

Idem at the generic level for the genus analysis.

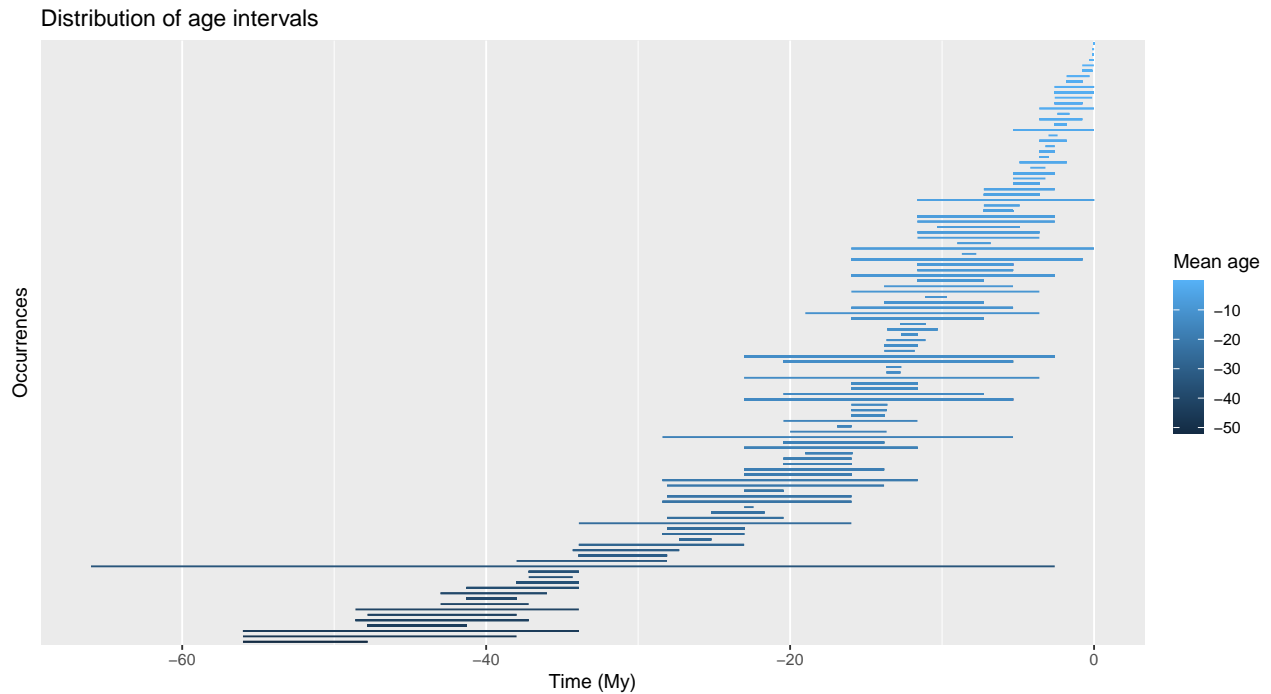
Remove those occurrences from our initial dataset to avoid redundancy.

## Explore the dataset

### Repartition through time

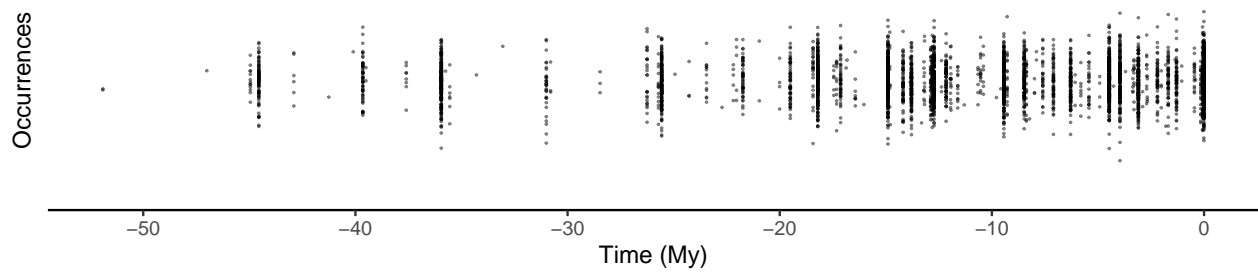
#### Full fossil record

Each occurrence is associated with a **stratigraphic age uncertainty interval** (minimum and maximum ages).

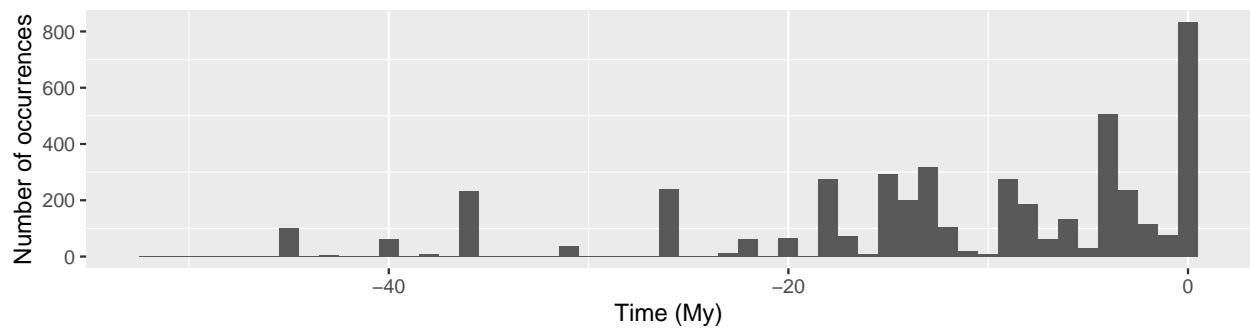


First, one can use the stratigraphic age uncertainty interval midpoint as the occurrence fixed age.

### Repartition of 4609 recorded occurrences through time – Midpoint

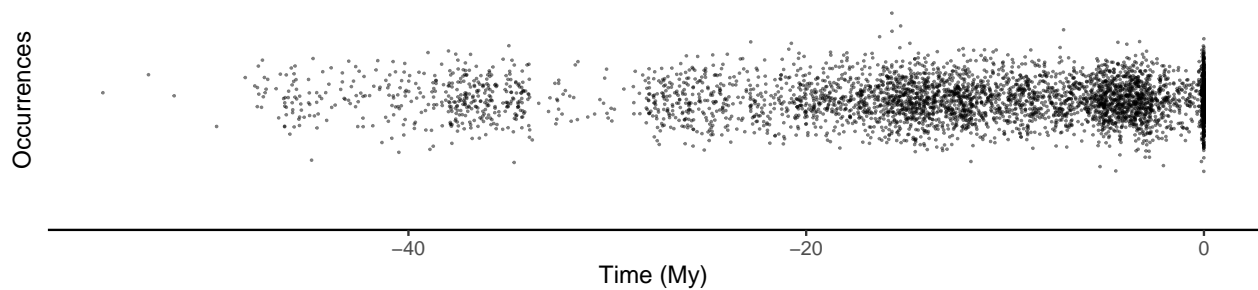


### Evolution of the number occurrences through time ( total of 4609 ) – Midpoint

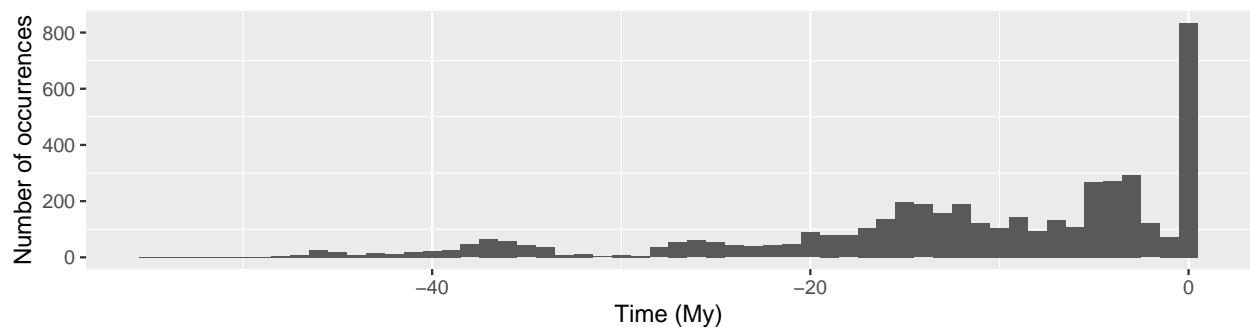


→ Numerous occurrences seem to have the same stratigraphic age uncertainty so in order to avoid clusters let's draw them uniformly in their interval rather than taking the midpoint.

### Repartition of 4609 recorded occurrences through time – Uniform draw



### Evolution of the number occurrences through time ( total of 4609 ) – Uniform draw

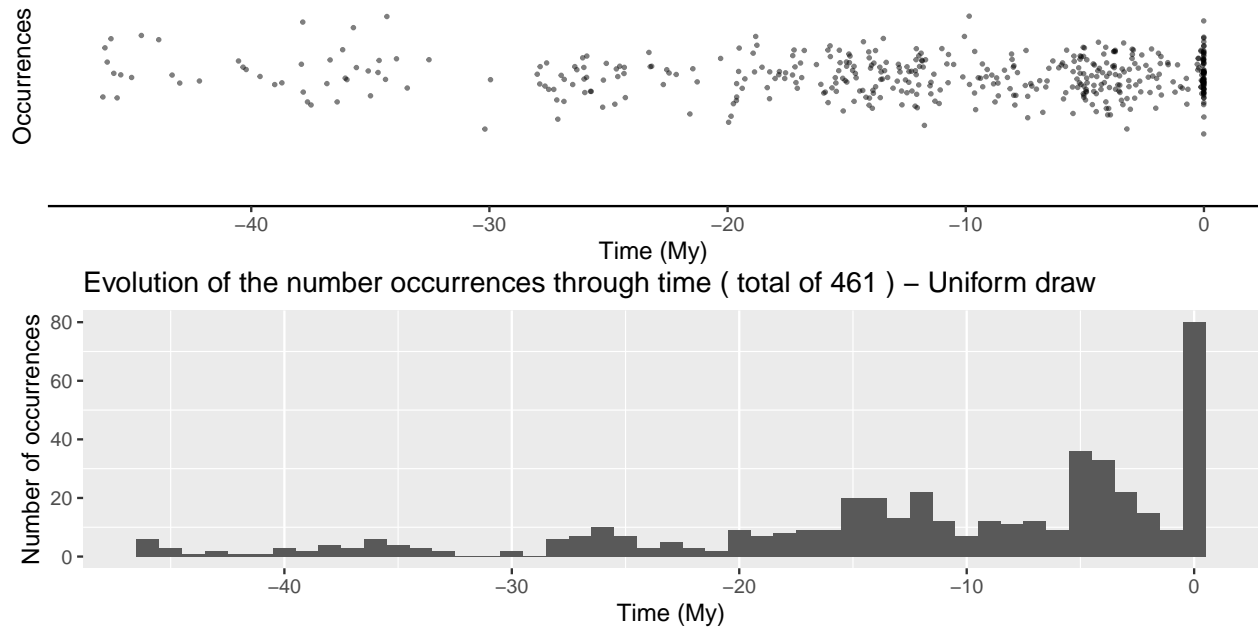


→ The distribution seems much smoother now.

## Uniform subsampling

These occurrences are too numerous to perform a quick preliminary analysis, so let's randomly subsample a fraction of them.

### Repartition of 461 recorded occurrences through time – Uniform draw



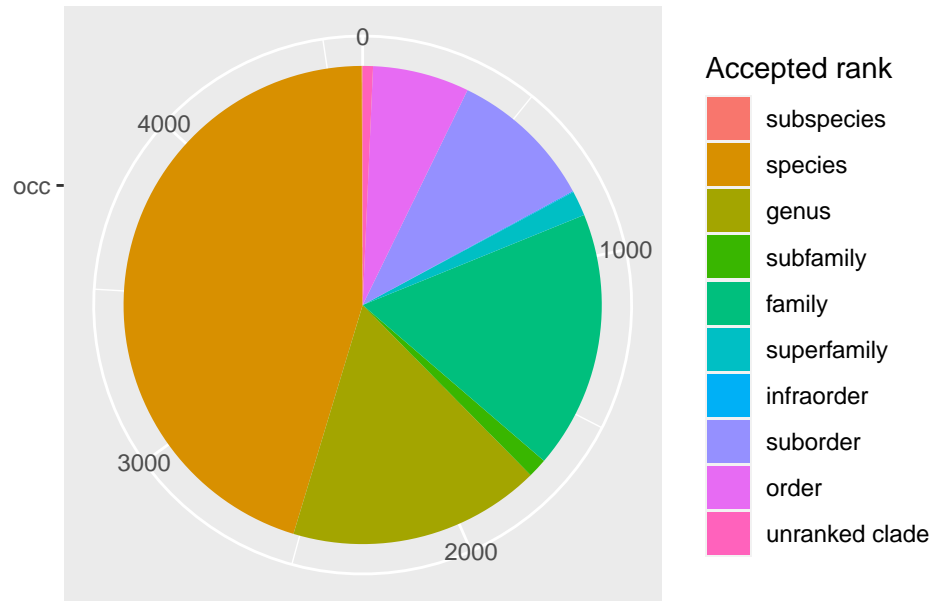
→ The distribution looks similar, with some noise due to higher variance with smaller sample.

## Repartition among accepted ranks

Occurrences are associated with an **accepted rank**, depending on the level of detail of their description and attribution to a given taxonomic position.

## Pie chart

### Repartition of occurrences among accepted ranks

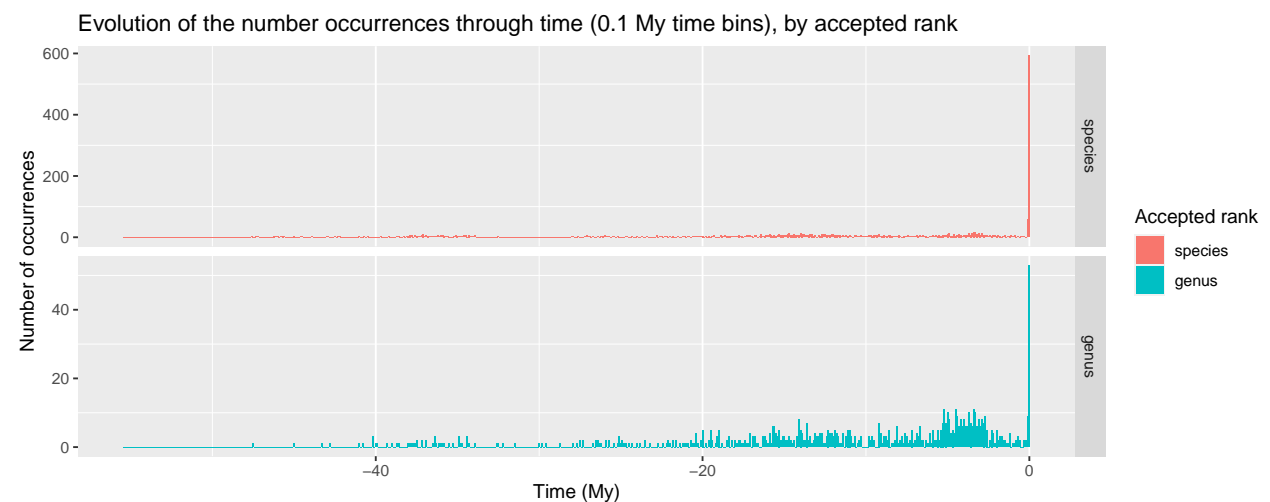


→ Half of the occurrences are identified at the level of the species and 1/3 at the genus or family.

Some clades are unranked:

##	Chaeomysticeti	Neoceti	Panphyseteroidea	Pelagiceti
##	26	2	1	1
##	Platanidelphidi	Squaloceti		
##	1	1		

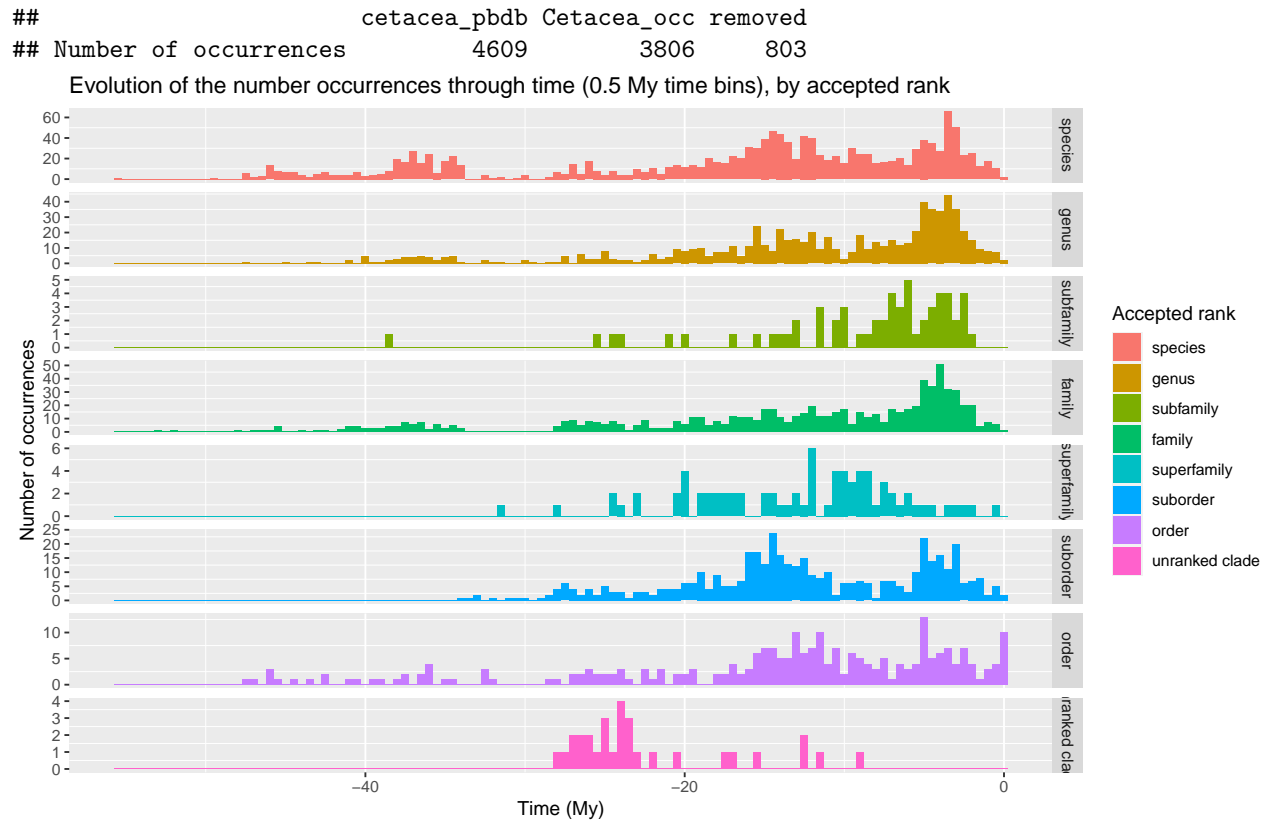
## Time repartition by rank



→ An apparent huge cluster of occurrences appears in recent times, with very precise dating = Artifact due to the “Pull of the Recent” effect ? Bones collected by ancient hominids ?

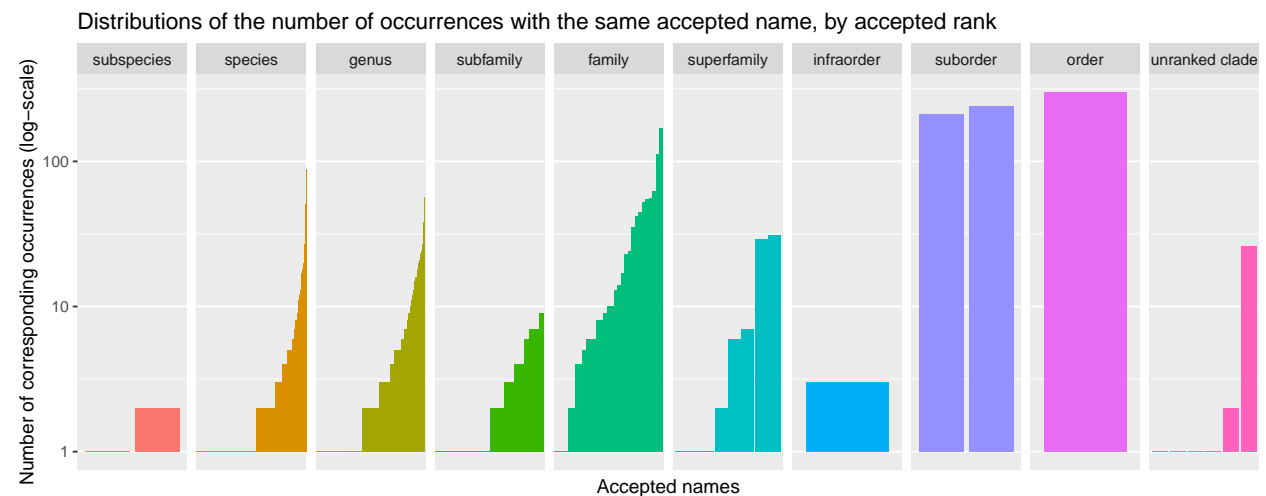
⇒ These occurrences seem to obey to a completely different process and over a very short period of time, so we decided to **remove all Late Pleistocene and Holocene occurrences** in order to avoid this bias.

### BIAS CORRECTION N°1: Remove all Late Pleistocene and Holocene occurrences



→ The distributions are much more homogeneous after the correction. Similar trends can be observed at each rank, with peaks at ~15My and ~5My.

### Redundancy of occurrences with the same accepted name



→ ~Half of species/genera/subfamilies have only one specimen by accepted name, but it could go up to ~50 within the same species and ~200 occurrences within the same suborder. **In our model all species are supposed to have the same abundance (identical sampling rates among branches), so those huge differences will have to be mitigated.** See later for a quantification of this discrepancy and a comparison with our model's assumptions

⇒ Our goal now will be to correct this abundance bias.

## Mitigate the abundance bias

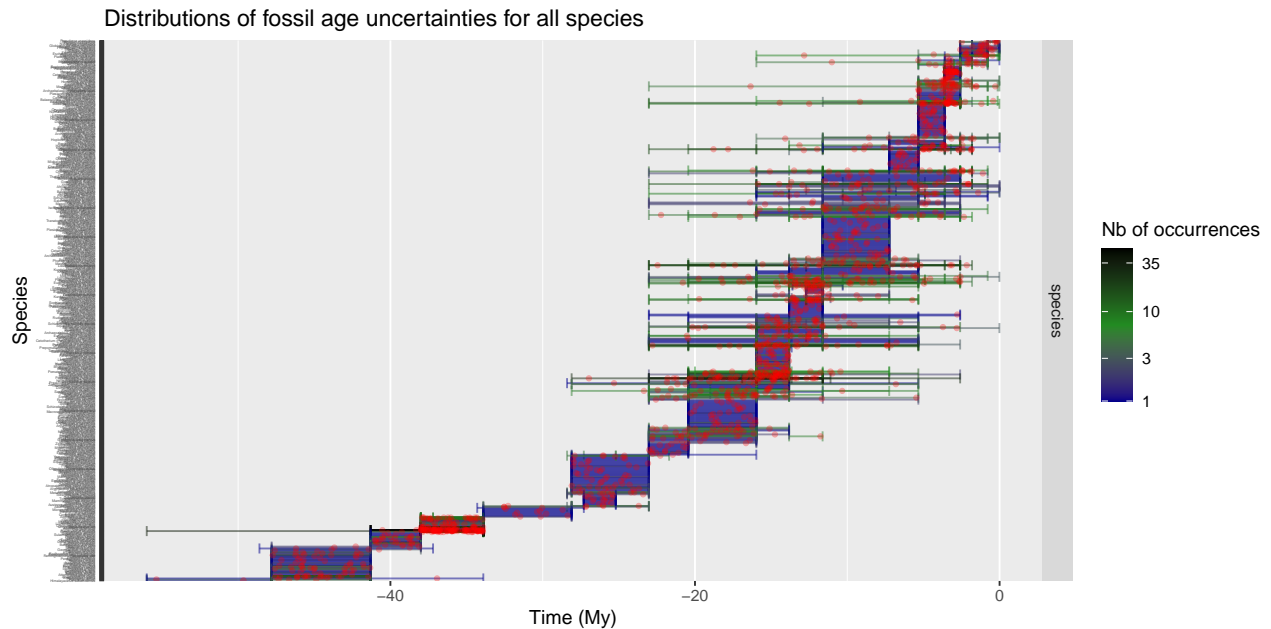
To quantify more precisely this bias, we will:

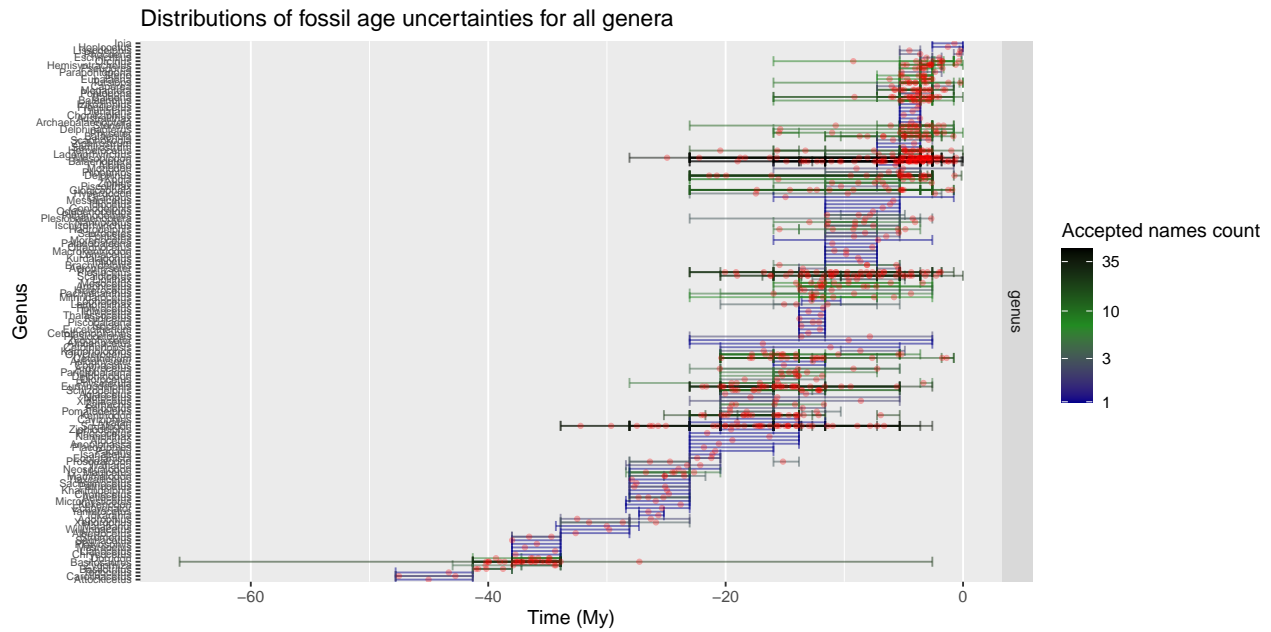
1. Count the **number of occurrences** for each taxon
2. Obtain the **stratigraphic range** of each taxa - ie. the time interval in which a taxon is observed - by combining the stratigraphic age uncertainties of its occurrences
3. Compute the **density of occurrences** = nb of occurrences / stratigraphic range, expressed in “number of occurrences by million years”

## Compute the occurrence density

### Count occurrences by accepted name

Count the number of occurrences with the same accepted name and plot their distribution.





→ Some occurrences have too much age uncertainty (unrelated to taxon longevity in extreme cases like the one spanning 60My), they risk to artificially increase species duration.

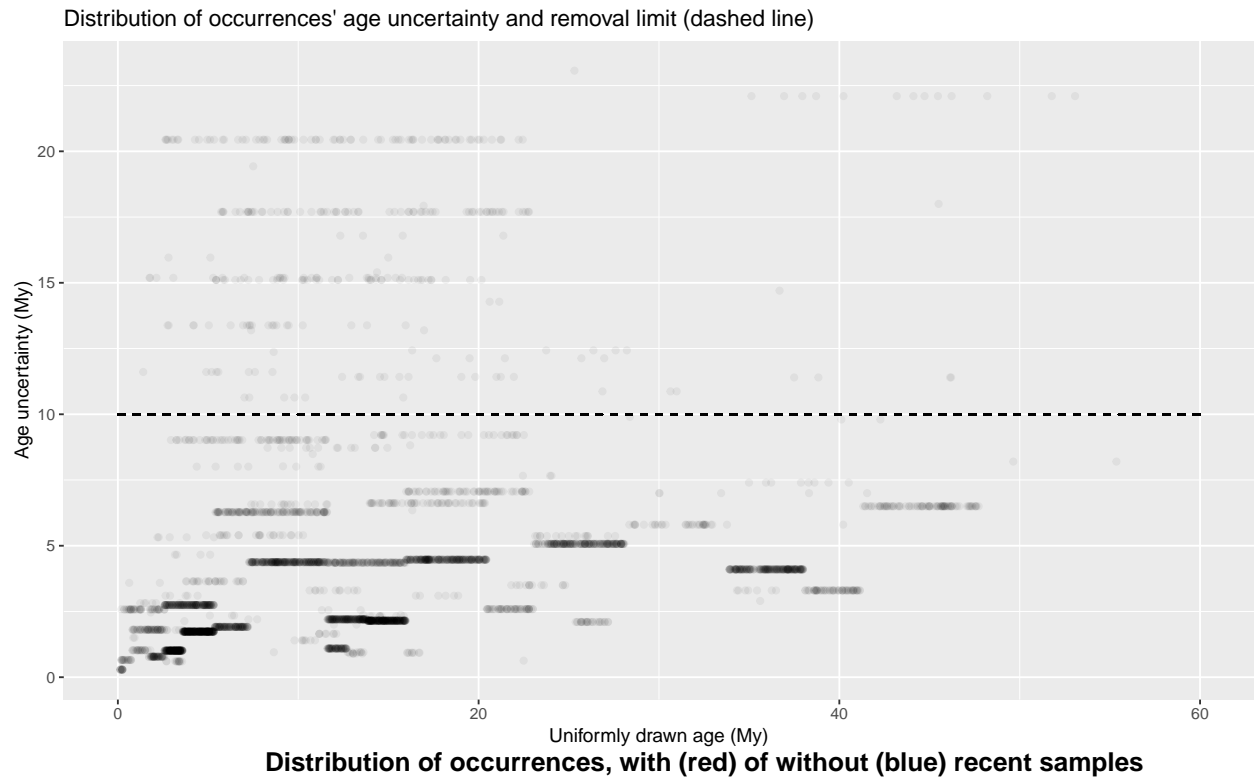
## BIAS CORRECTION N°2: Exclude occurrences with highly uncertain dating (> 10My)



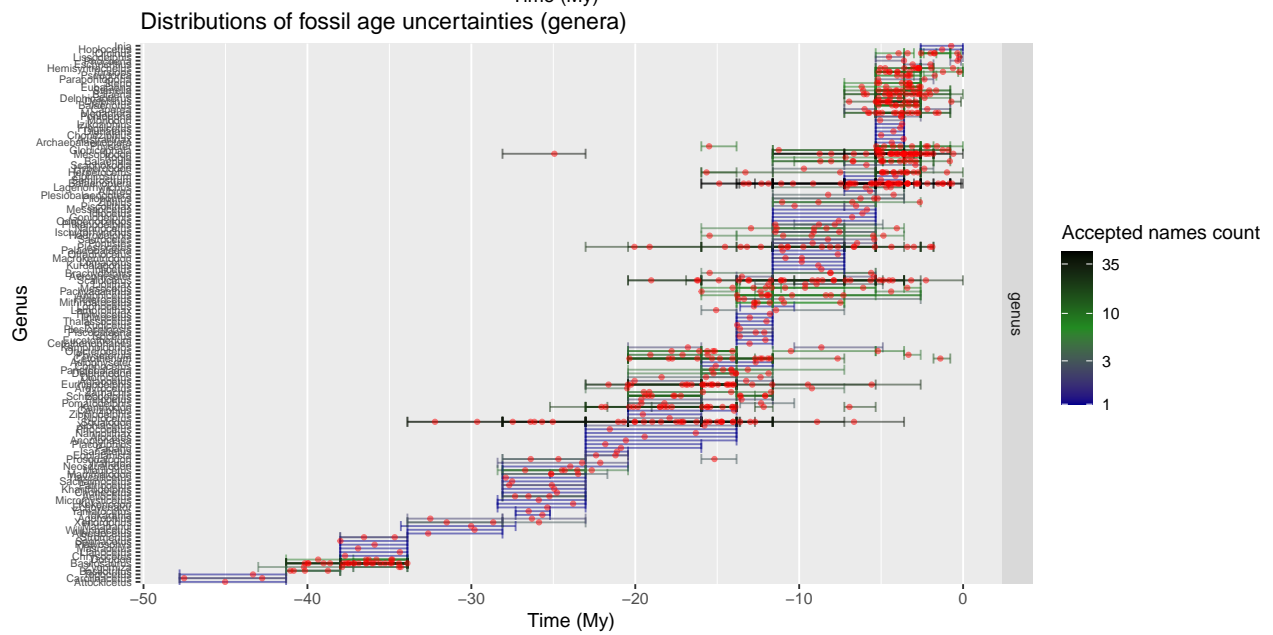
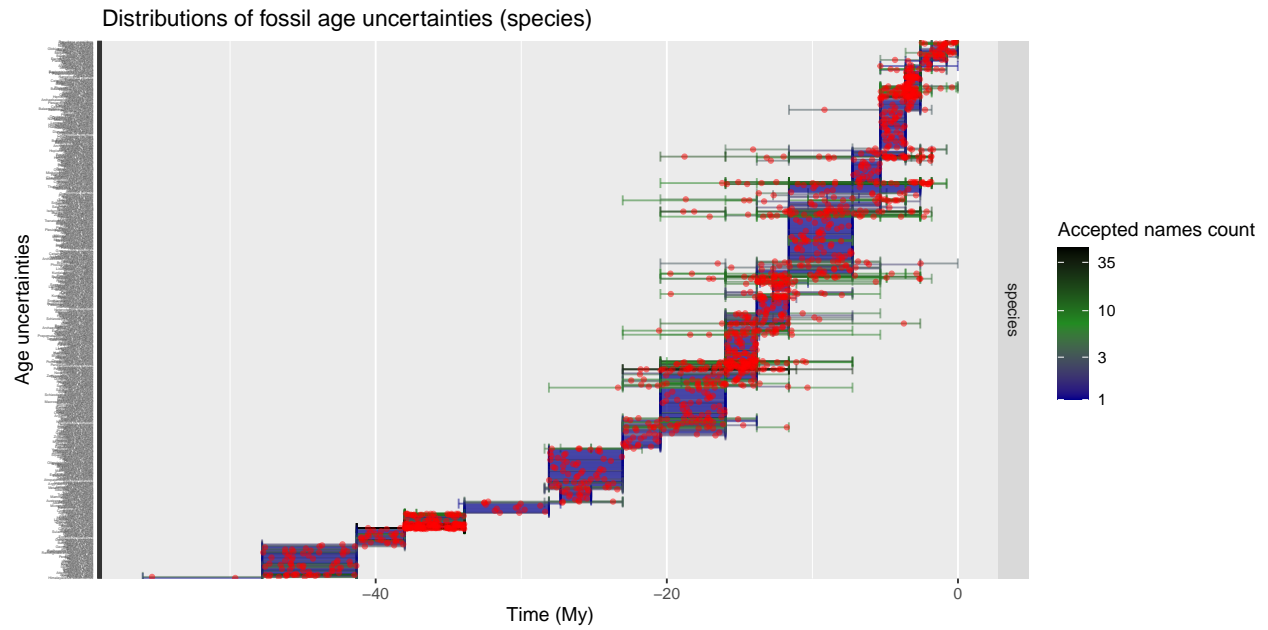
Most of occurrences show less than 10 My age uncertainty, let's keep only these ones.

```
##                               all_ranges smaller_10My removed
## Number of occurrences      3806      3504      302
```



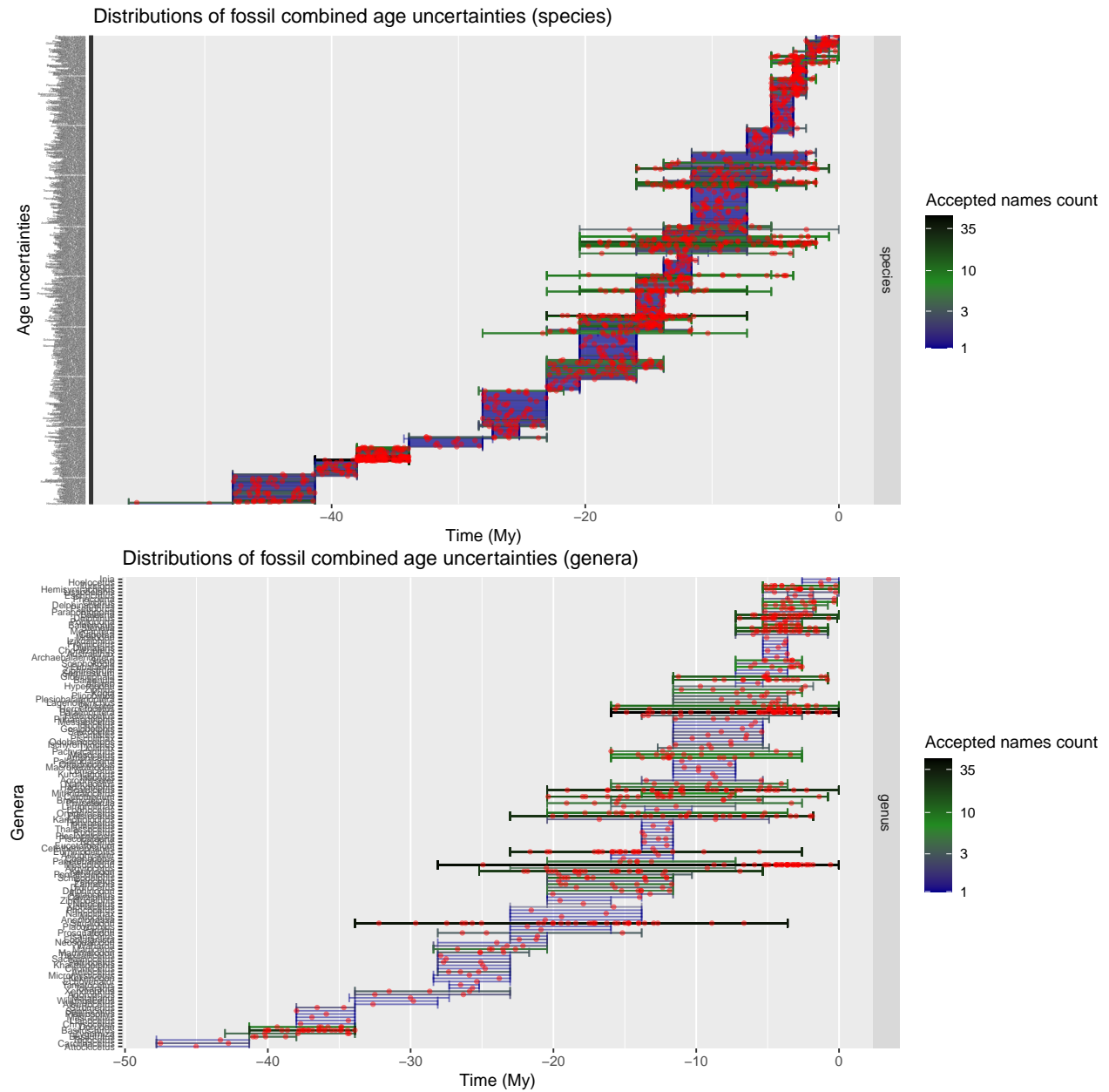


→ The removal of highly uncertain occurrences seems to be only a little biased, even if uncertainty globally increases with age.



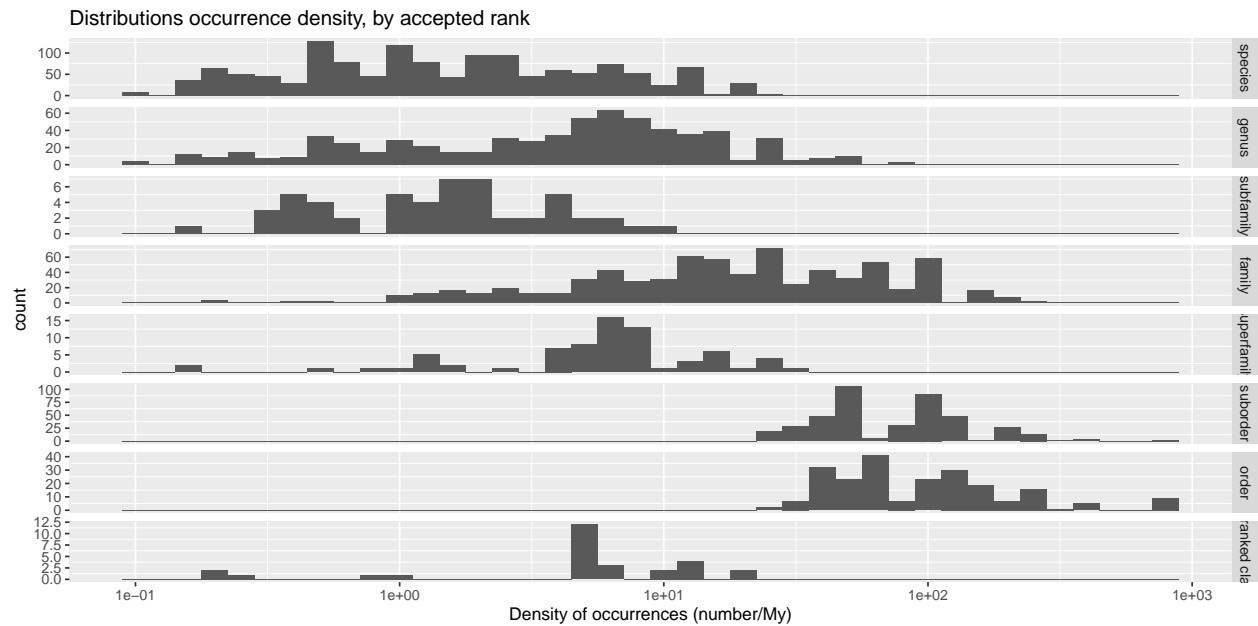
→ Some taxa have several occurrences with distinct age uncertainty intervals, **let's combine them into a unique stratigraphic range.**

Stratigraphic ranges = unique time interval combining age uncertainties of occurrences of the same taxon (excluding the most uncertain ones)



## Density distributions

Plot the density of occurrences (= number of occurrences by unit of time, in the stratigraphic interval of a taxon) for each taxonomic rank.

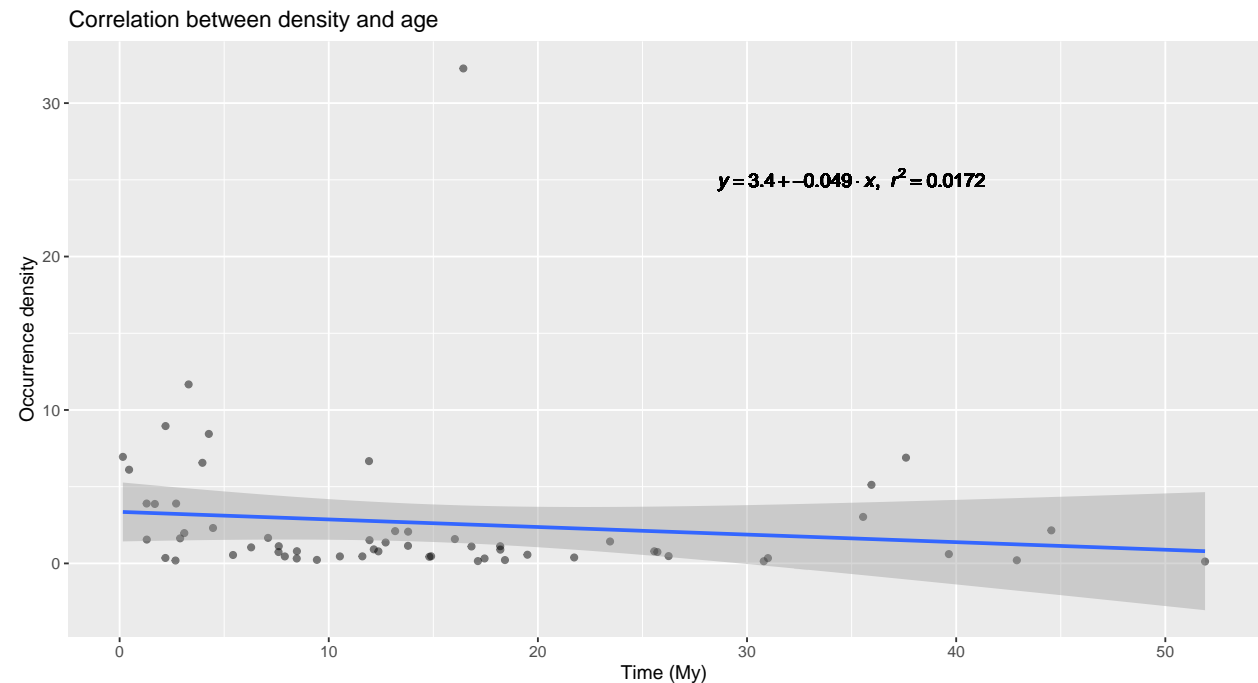


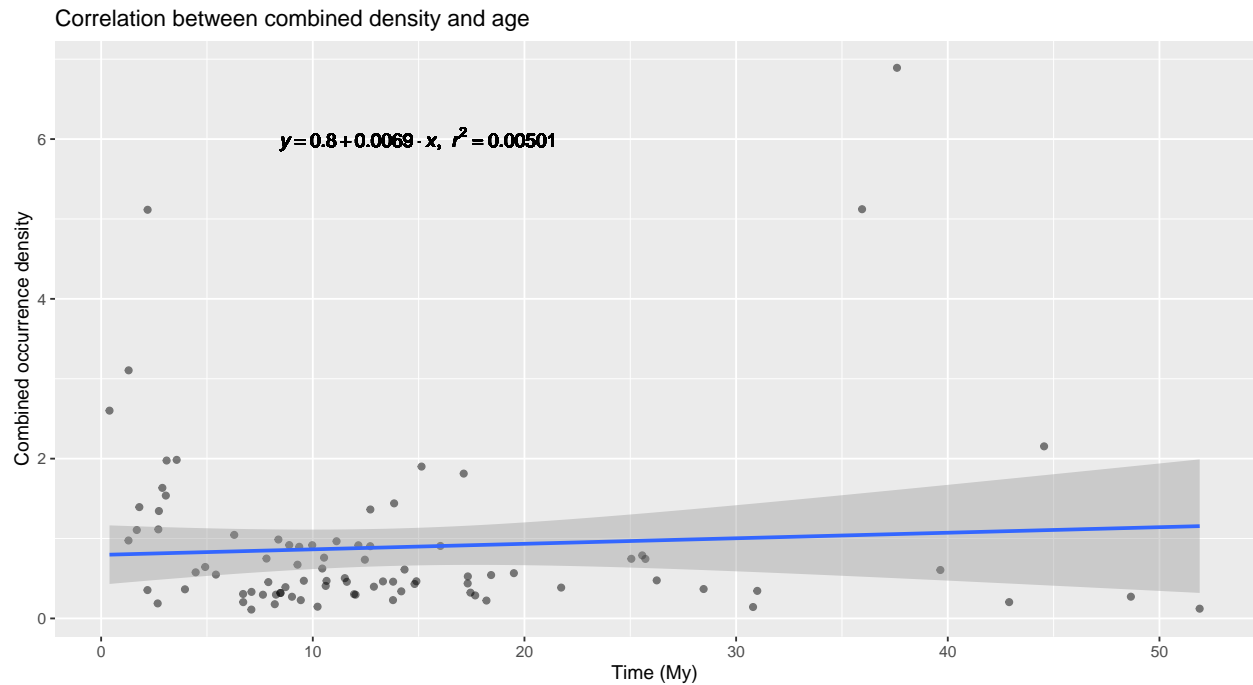
→ Density logically increases with taxa ranks, but the distributions for species and genera dramatically spread over two orders of magnitude.

### Correlation between occurrence density and age

If we want to correct species abundance differences based on the density of occurrences in the age interval, those factors should not depend on time in order to avoid penalizing periods with higher densities.

Let's compare the occurrence densities computed with the initial time ranges and the *combined densities* computed with the combined stratigraphic ranges.





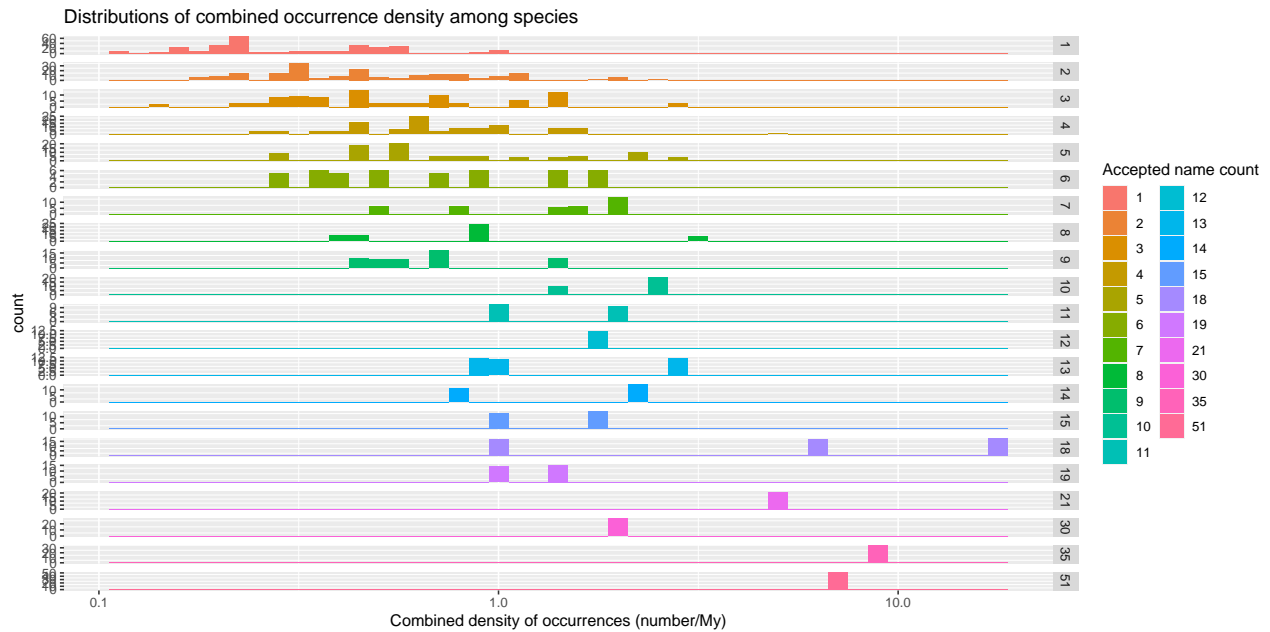
→ The density based on combined ranges seems less time-dependent than the density based on initial time ranges. We will therefore use the combined density for our corrections.

### BIAS CORRECTION N°3: Subsample occurrences by homogenizing the combined occurrence density

**Subsampling by homogenization** = reducing the discrepancy between high and low occurrence densities among taxa by subsampling preferentially the densest ones.

#### Compare densities by accepted name count (species only)

Let's focus now on the occurrences accepted at the species level because they are the one for which we can correct the abundance bias by subsampling the most concentrated combined intervals.

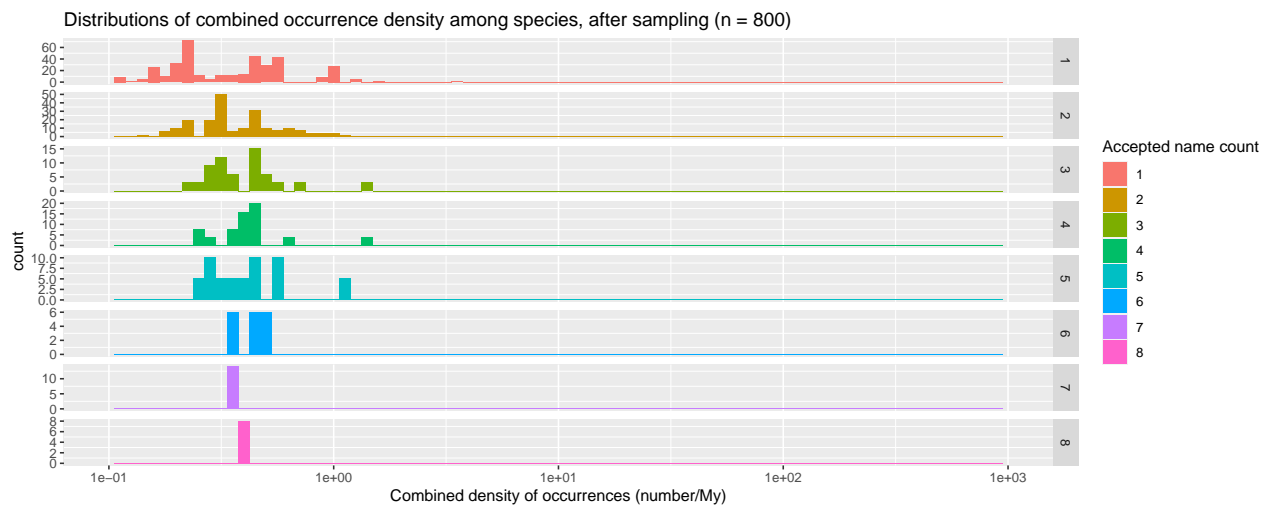


→ There is a huge span of densities driven by the number of occurrences for the same species that we can reduce by subsampling the most concentrated intervals.

### Impact of correcting subsampling on density distributions (species only)

## Warning: Removed 2 rows containing non-finite values (stat\_bin).

## Warning: Removed 16 rows containing missing values (geom\_bar).

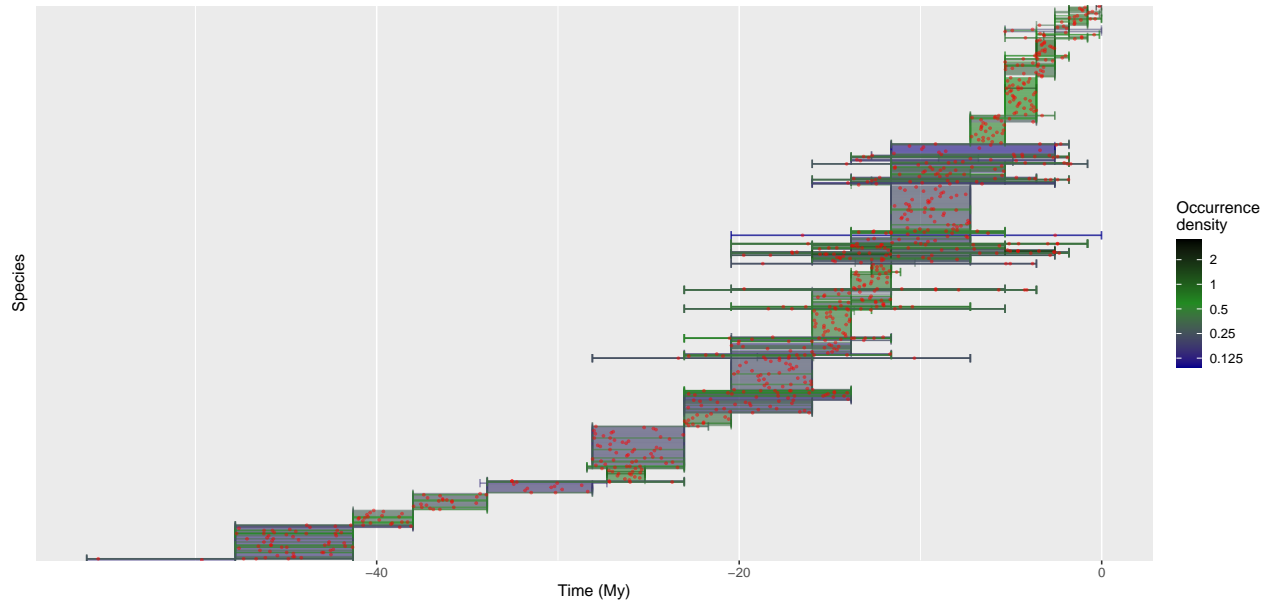


→ Subsampling by homogenization successfully reduces the density span from 2 to 1 order of magnitude.

### Impact of subsampling on occurrences repartition (species only)

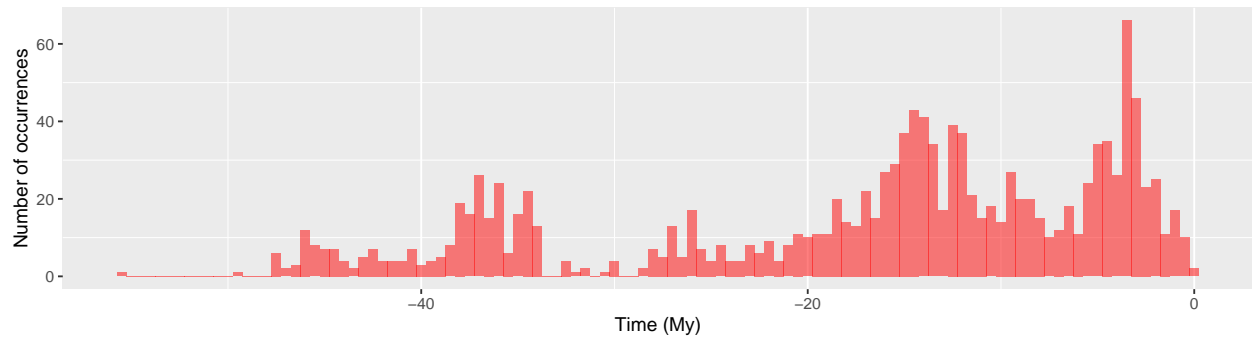
See what our distributions look like after subsampling:

Distributions of species fossil age uncertainties, after subsampling by homogenization (n = 800)

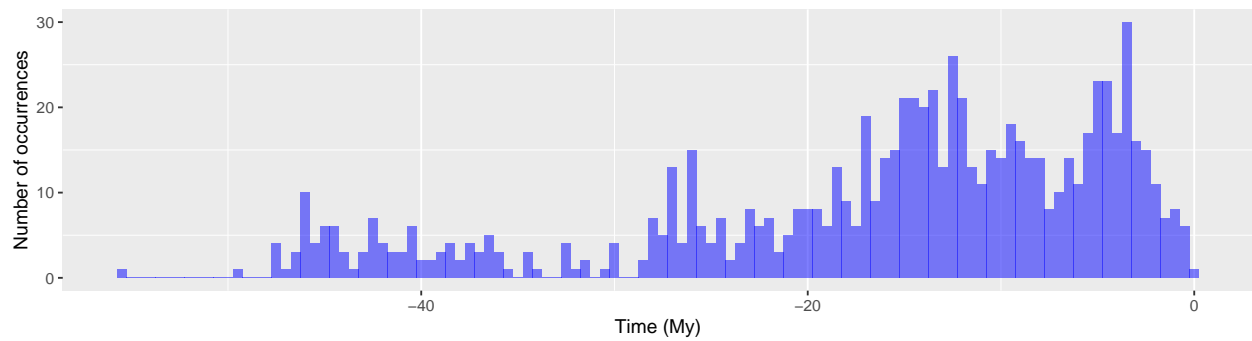


→ Some highly dense cluster became much more similar to the others.

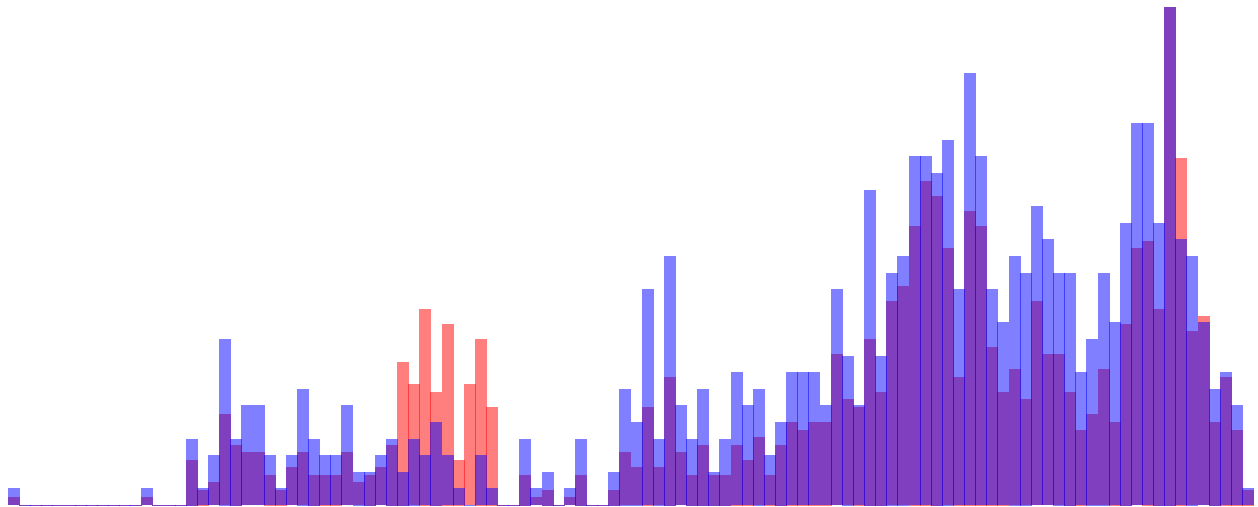
Evolution of the species occurrence number through time (n = 1335)



Evolution of the species occurrence number through time, after subsampling by homogenization (n = 800)



If we superpose these 2 plots:

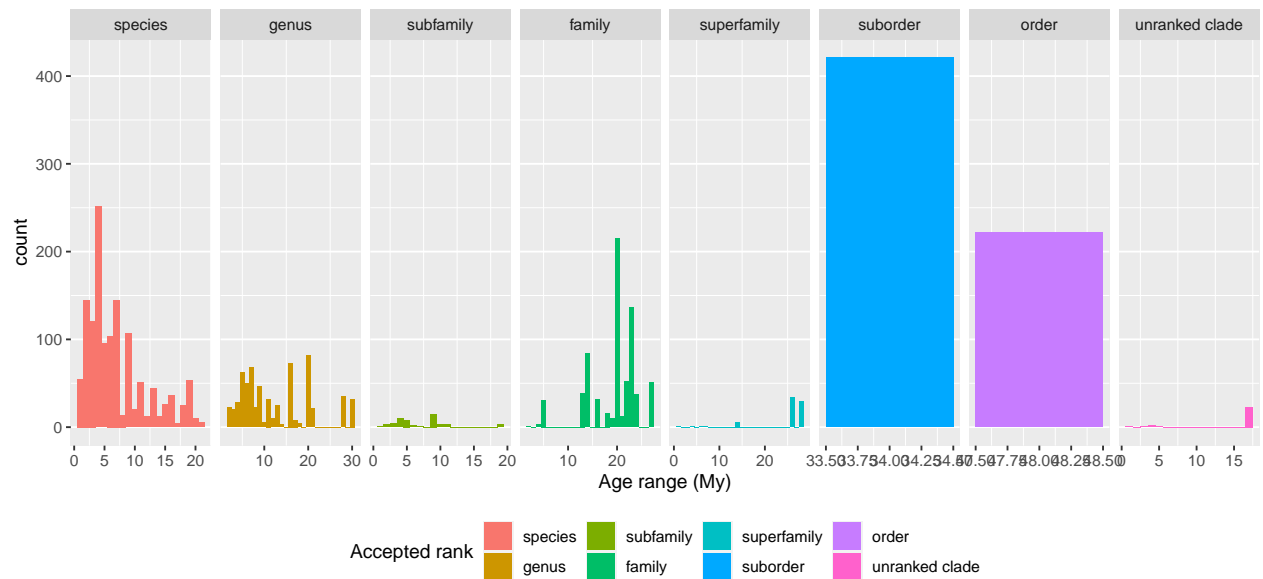


→ We get the new species occurrence distribution after subsampling, that could be used for doing inference with the occurrence birth-death model.

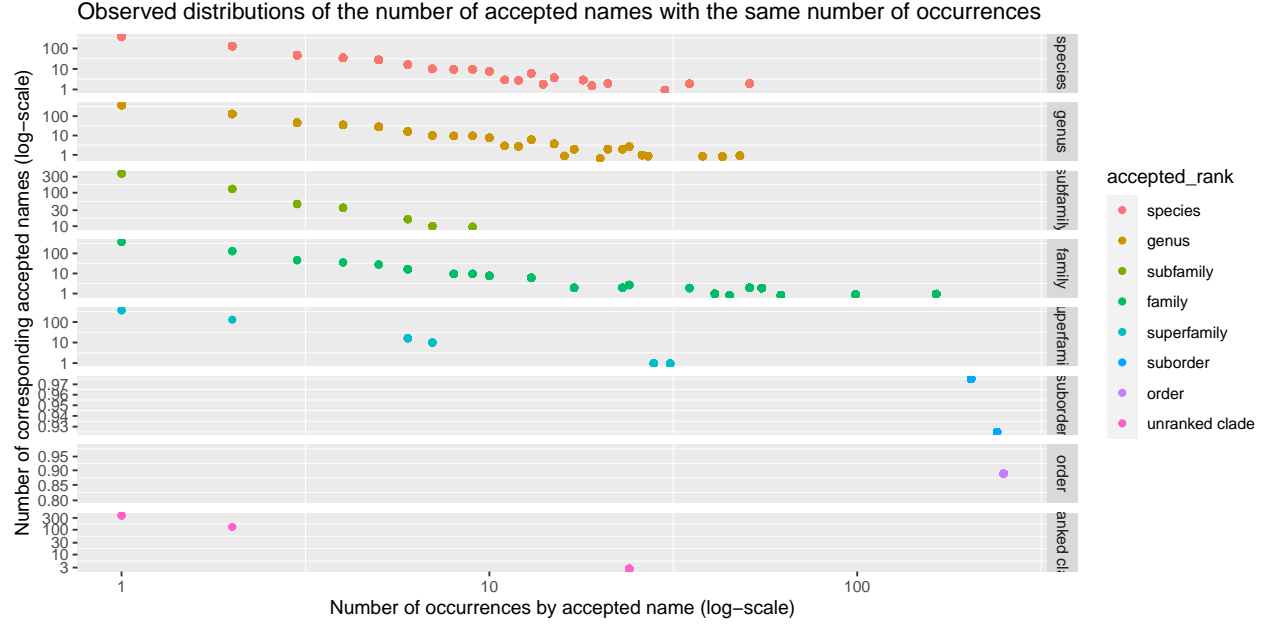
### Compare with a Poisson sampling process

In order to check if the data fit our assumptions of constant-fossilization-rate Poisson sampling we compare the observed occurrences distributions with the expected ones. Specifically, we will look at the number of taxa represented by 1, 2, 3, ... occurrences and the one that we would expect for a given distribution of stratigraphic ranges (seen here as a proxy for species duration).

Observed distributions of stratigraphic ranges







In a Poisson process with occurrence sampling rate  $\omega$  and for a given time interval of length  $t$ , the probability of observing  $N_t = k$  occurrences is given by the Poisson distribution of mean of parameter  $\omega \times t$ :

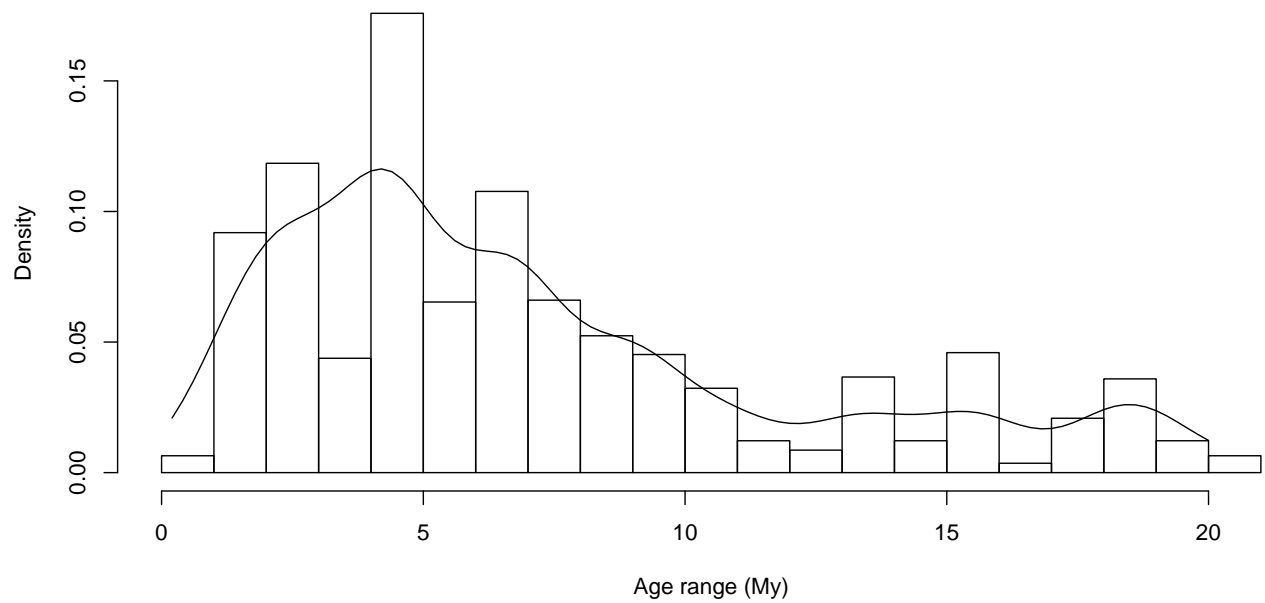
$$\mathbb{P}(N_t = k) = e^{-\omega t} \frac{(\omega t)^k}{k!}$$

So in order to have the absolute probability of observing  $N_0 = n$  occurrences we have to integrate over the full distribution of age uncertainties  $t$ , called  $f(t)$ :

$$\mathbb{P}(N_0 = n) = \int_t P(N_t = n) f(t) dt = \int_t e^{-\omega t} \frac{(\omega t)^n}{n!} f(t) dt$$

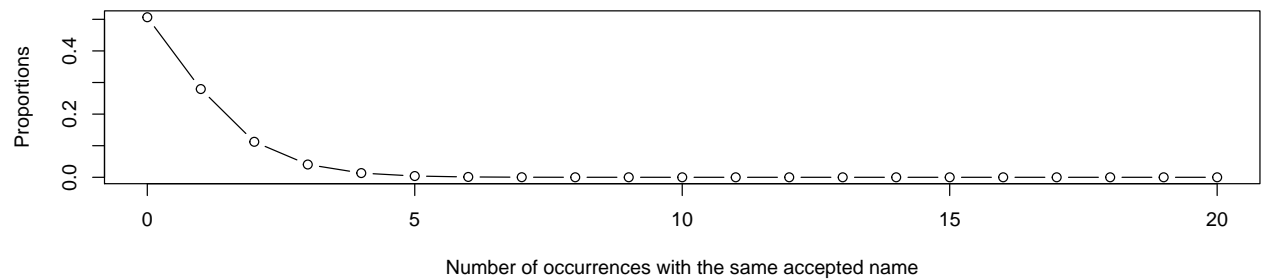
First, approximate this distribution:

### Density approximation of the empirical stratigraphic range distribution

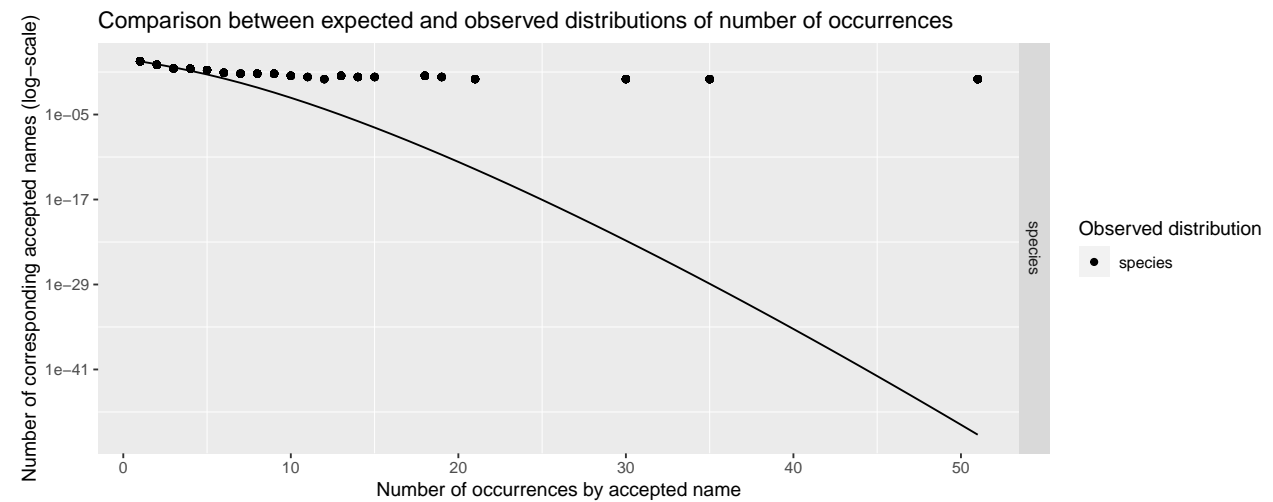


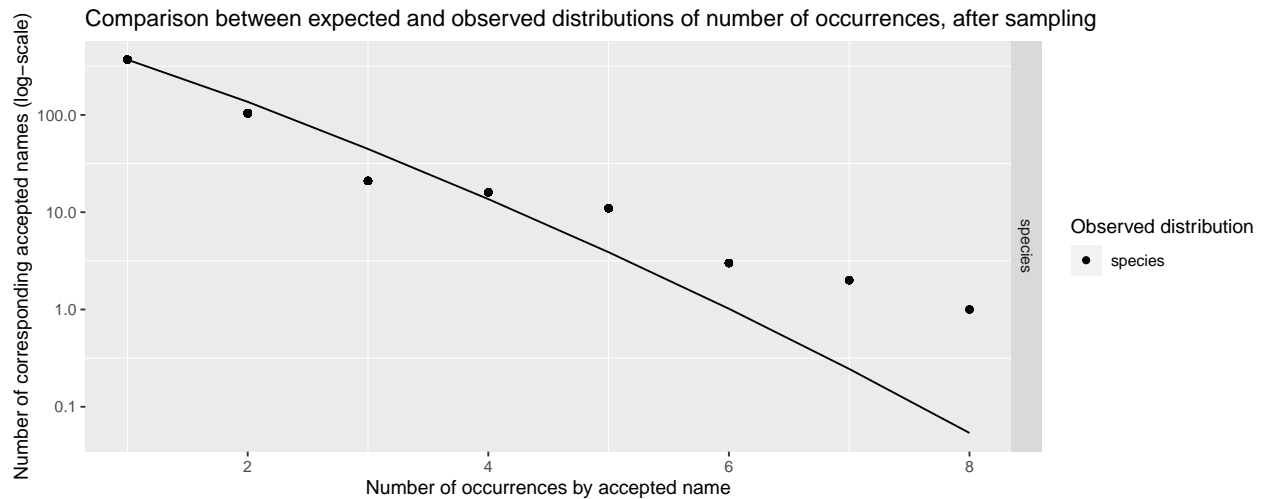
Then integrate and plot the expected distribution for a given omega:

### Expected distribution of the number of accepted names with the same number of occurrences



Finally, try to find an  $\omega$  value that approximately fits the first points (the least affected by oversampling biases) and check if the other points follow the expected curve:





⇒ Initial observations really do not fit the expectations, **species with more than 5 occurrences must remain very rare !** But our subsampling seems to correct most of this bias.

However, this method requires to make several arbitrary choices that may introduce new biases so we will instead subsample at other levels (palaeontological collection, geological formation). In each case only one occurrence will be sampled for the similarly identified, a process we will refer to as **aggregating** these occurrences according to the chosen factor.

## BIAS CORRECTION N°3bis: Subsample occurrences by aggregating similarly identified occurrences in each collection

**Subsampling by aggregation** = aggregate all the occurrences of a given taxon according to a given criterion (collection, geological formation, country, geological plate).

### Wrong track: Aggregate similarly identified occurrences in each collection

In order to reduce the abundance bias, we may keep only one occurrence for each collection:

##	Cetacea_occ	Cetacea_occ_aggreg	removed
## Number of occurrences	3806	3558	248
## Number of occurrences (species only)	1435	1362	73

→ Not enough occurrences are removed to make a sufficient difference. If we look at the collection with the highest number of occurrences:

##		
##	Aprixokogia kelloggi	Balaena
##	1	2
##	Balaenoptera acutorostrata	Balaenopteridae
##	2	2
##	Balaenula	Bohaskaia monodontoides
##	2	1
##	Cetotheriinae	Delphinapterus
##	1	3
##	Delphinidae	Delphinus
##	2	2
##	Globicephala	Gricetoides aurorae
##	2	1

##	Herpetocetinae	Herpetocetus sendaicus
##	1	1
##	Herpetocetus transatlanticus	Kogia breviceps
##	1	1
##	Kogiidae	Kogiinae
##	1	2
##	Kogiopsis floridana	Lagenorhynchus
##	1	2
##	Lagenorhynchus harmatuki	Megaptera
##	1	2
##	Mesoplodon longirostris	Monodon
##	2	1
##	Ninoziphius platyrostris	Orycterocetus
##	3	1
##	Physeter macrocephalus	Physeteridae
##	1	1
##	Physeterinae	Physeterula dubusi
##	2	1
##	Plesiocetus	Pliopontos littoralis
##	1	1
##	Pontoporia	Pontoporiidae
##	1	2
##	Pseudorca	Scaldicetus
##	2	1
##	Stenella	Stenella rayi
##	2	1
##	Tursiops	Ziphius cavirostris
##	2	2

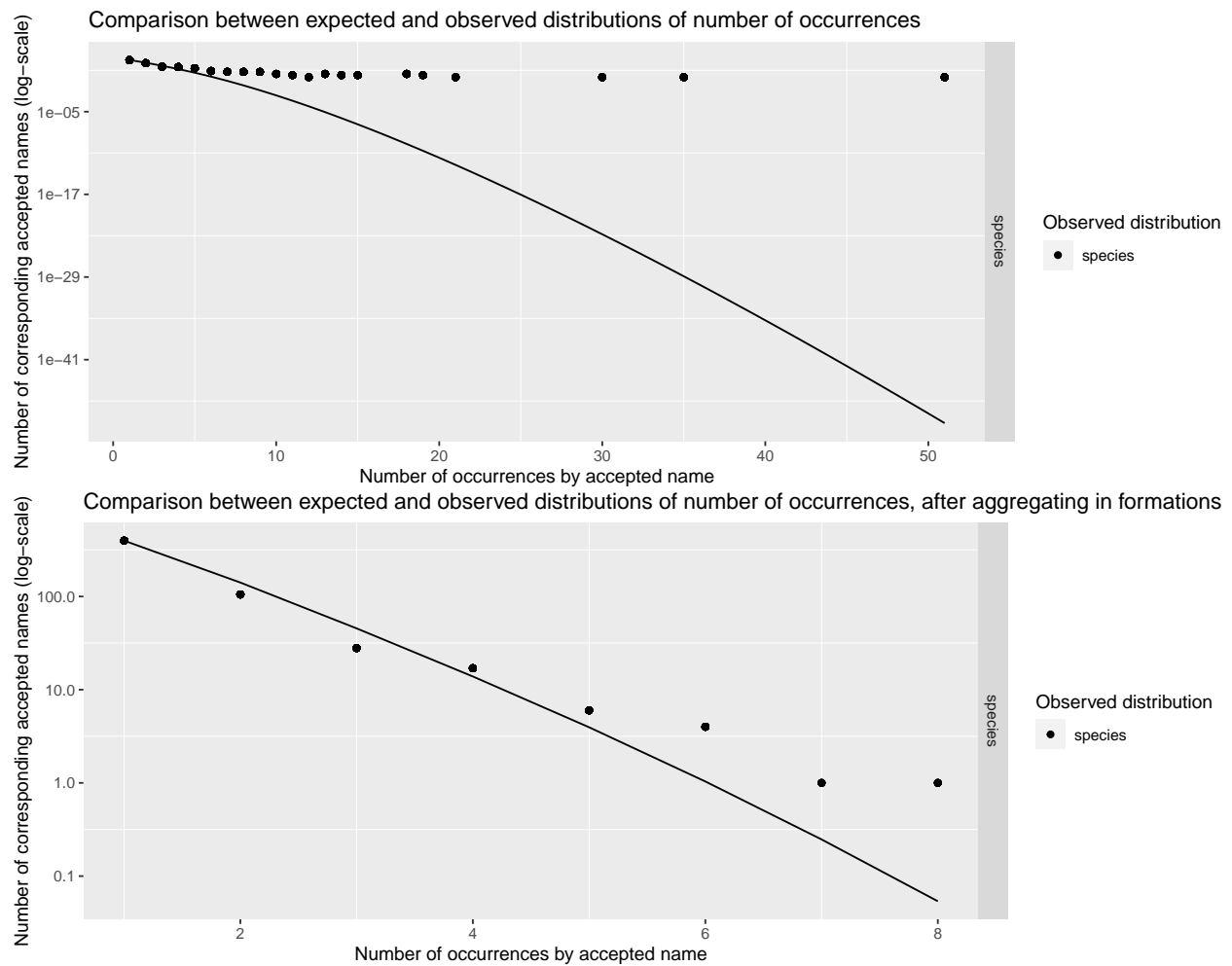
→ There are very few redundancies among the accepted names in collections so aggregating those won't reduce the abundance bias.

Instead, we may try to **aggregate occurrences with the same accepted name at the level of the geological formation** (ie subsample only one for each).

## Aggregate similarly identified occurrences in each formation

##	Cetacea_occ	Cetacea_occ_aggreg	removed
## Number of occurrences	3806	1983	1823
## Number of occurrences (species only)	1435	828	607

→ In that case the sub-sampling is big enough to hope correcting our bias.



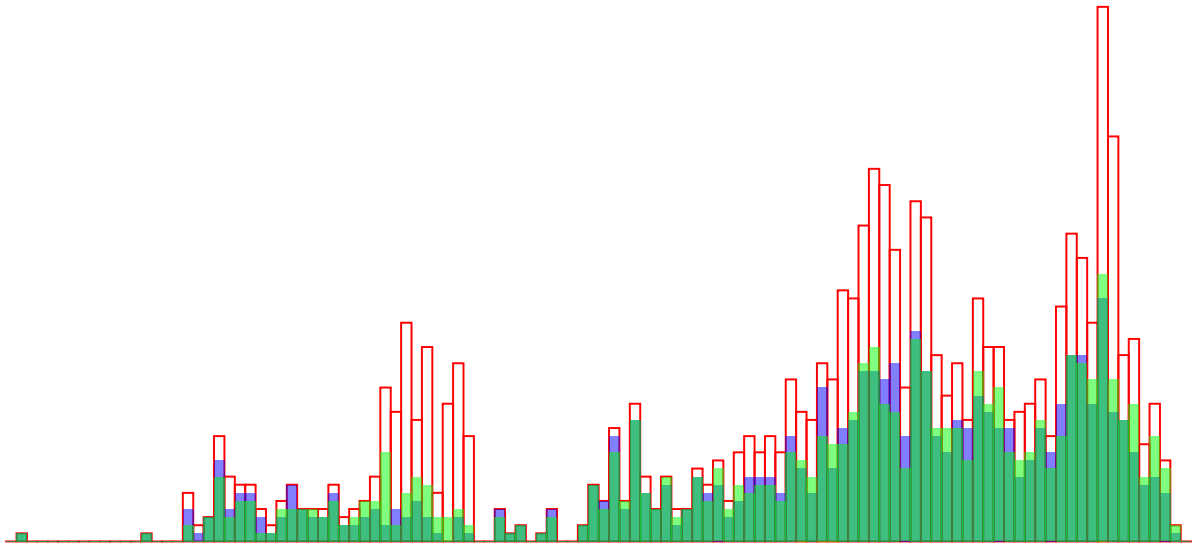
→ The initial bias is mostly corrected.

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

Initial occurrences distribution (red) and comparison after sub-sampling (blue) or aggregating in formations (green)



→ Comparing with the initial occurrences distribution and with the distribution after our first sub-sampling it appears that both methods lead to very similar distributions. This comforts us about the robustness of those approaches.

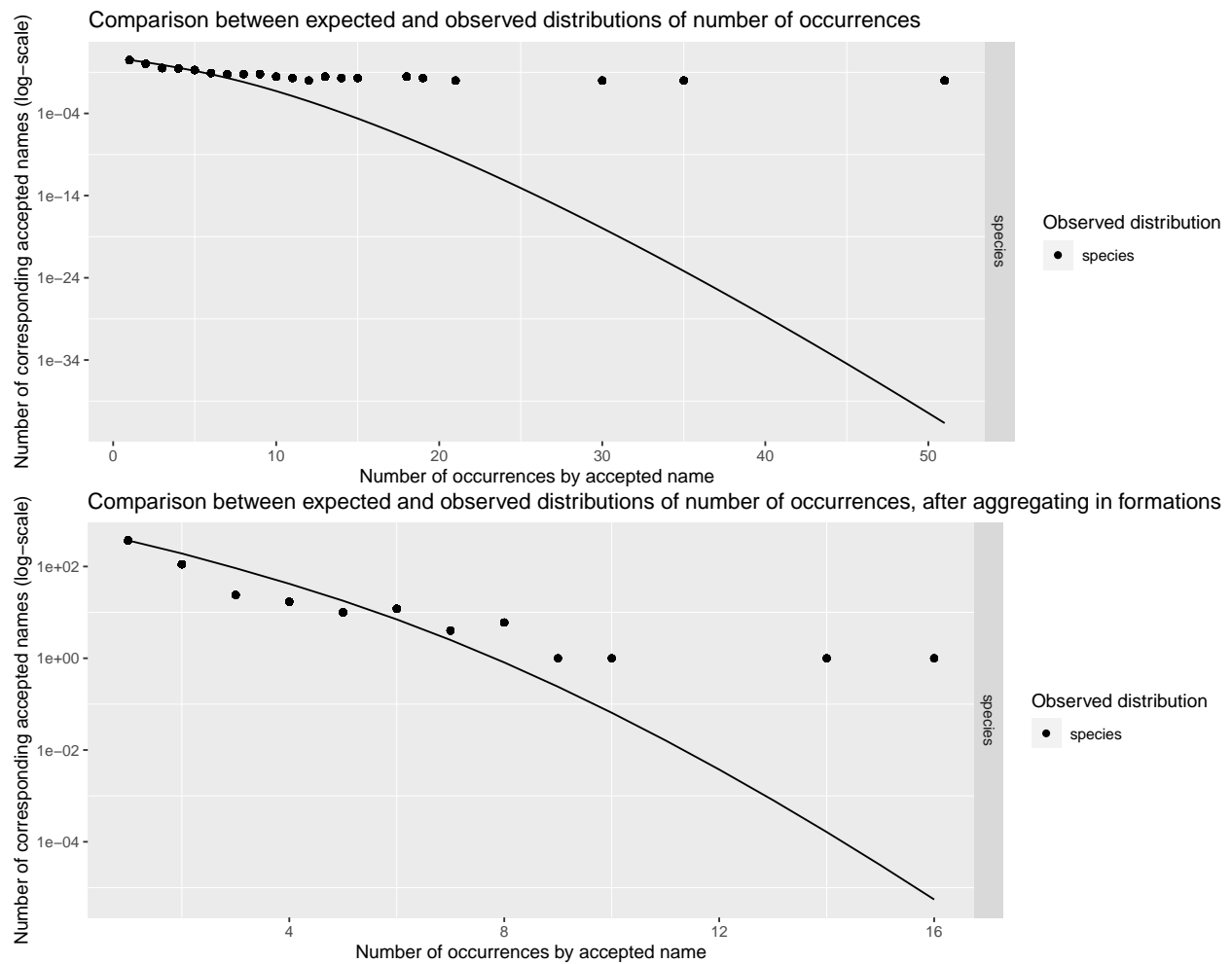
However, we have to take into account the occurrences that do not have any indicated geological formation to subsample them separately.

### Wrong track: Aggregate occurrences without formation by country + early interval

To approximate geological formation we chose to proceed to the aggregation based on the combination of the country and the early stratigraphic interval.

##	Cetacea_occ	Cetacea_occ_aggreg	removed
## Number of occurrences	3806	2642	1164
## Number of occurrences (species only)	1435	981	454

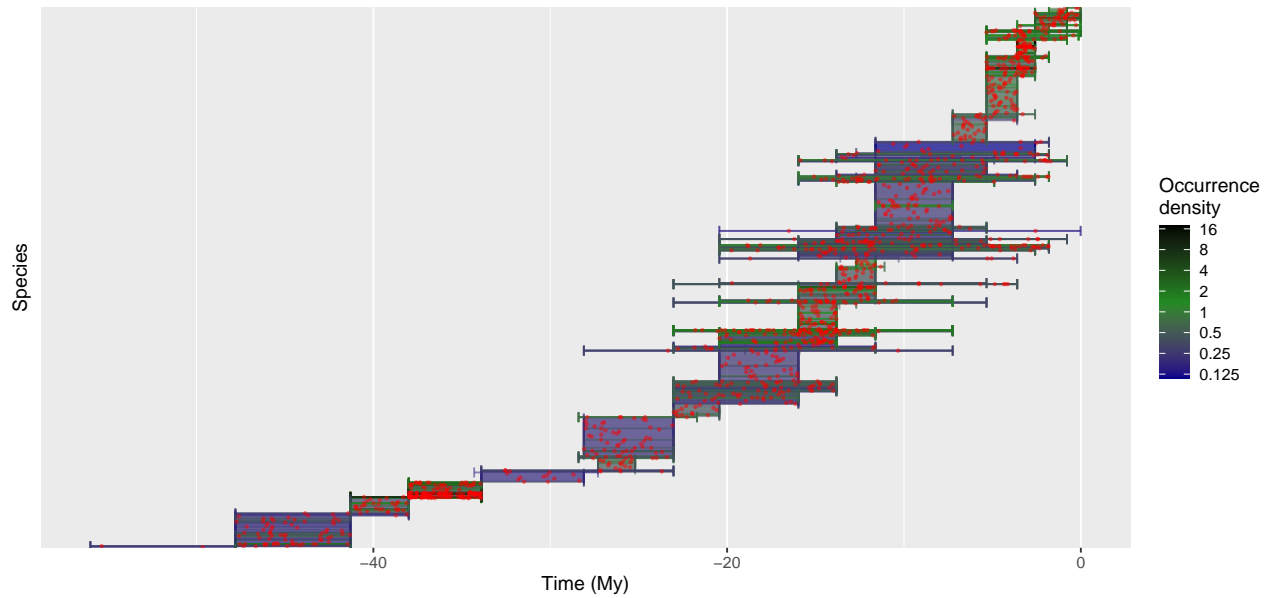
More occurrences remain after aggregating with this new method. Let's compare again with the theoretical distribution:



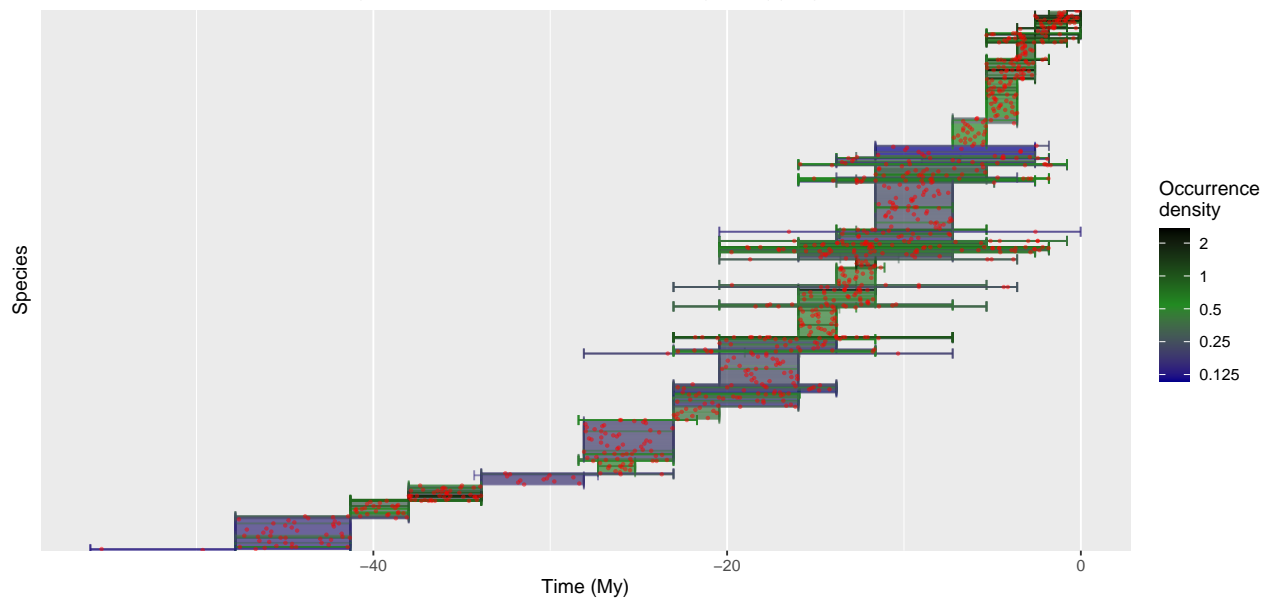
→ The correspondence is still good, except for two taxa:

```
##
##   Scaldicetus grandis Schizodelphis sulcatus
##               14                16
```

Distributions of species fossil age uncertainties, before subsampling (n = 1335)



Distributions of species fossil age uncertainties, after subsampling by aggregation (n = 800)



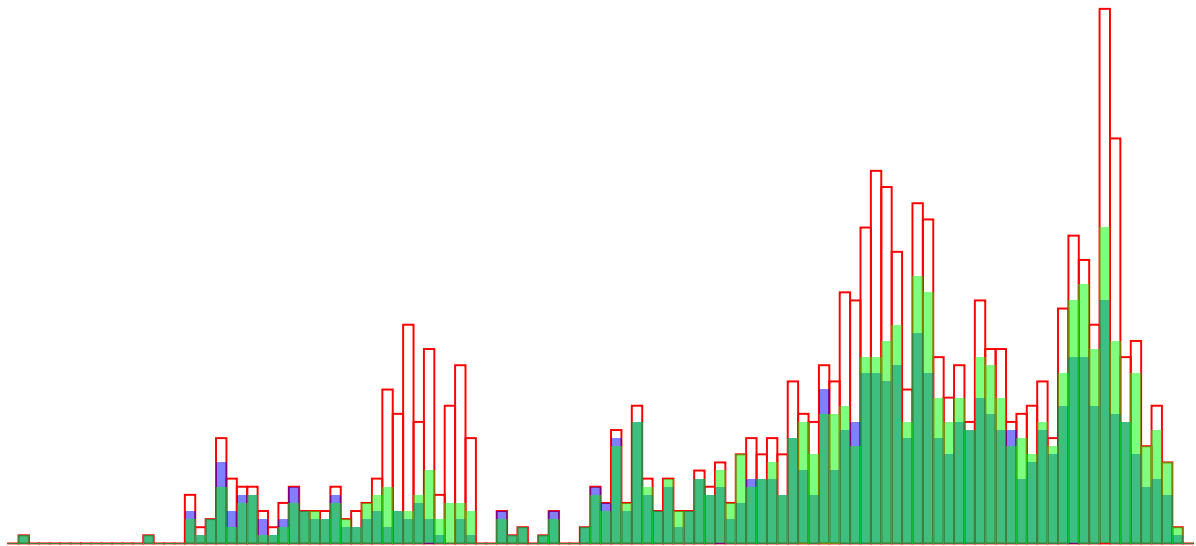
## Warning: Removed 2 rows containing missing values (geom\_bar).

## Warning: Removed 2 rows containing missing values (geom\_bar).

## Warning: Removed 2 rows containing missing values (geom\_bar).



Initial occurrences distribution (red) and comparison after sub-sampling (blue) or aggregating in formations (green)



## Saving 10 x 5 in image

## Warning: Removed 2 rows containing missing values (geom\_bar).

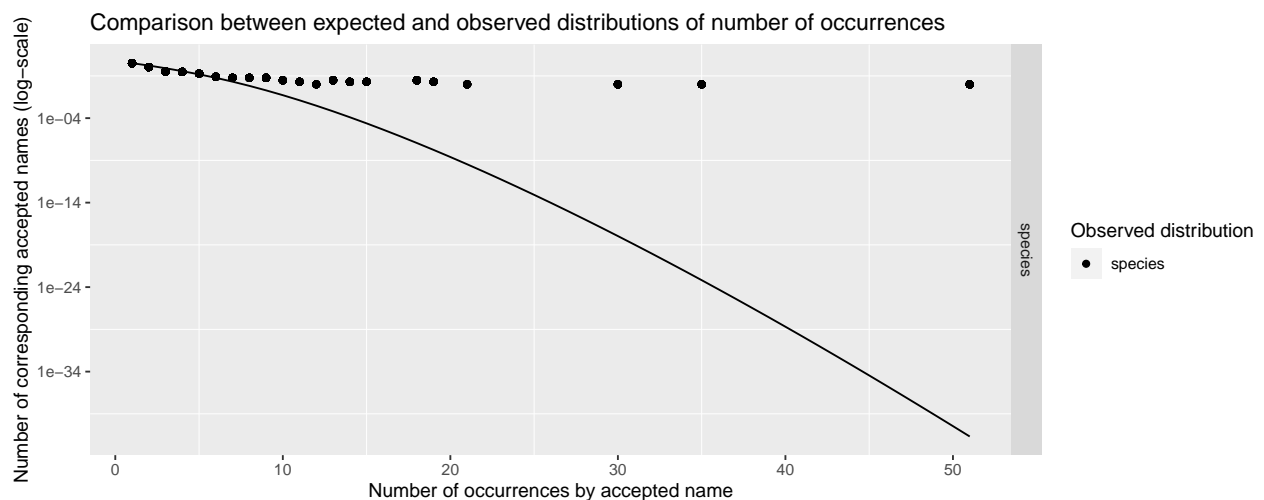
→ Comparing with the initial occurrences distribution and with the distribution after our first sub-sampling it appears that both methods lead to very similar distributions. This comforts us about the robustness of those approaches.

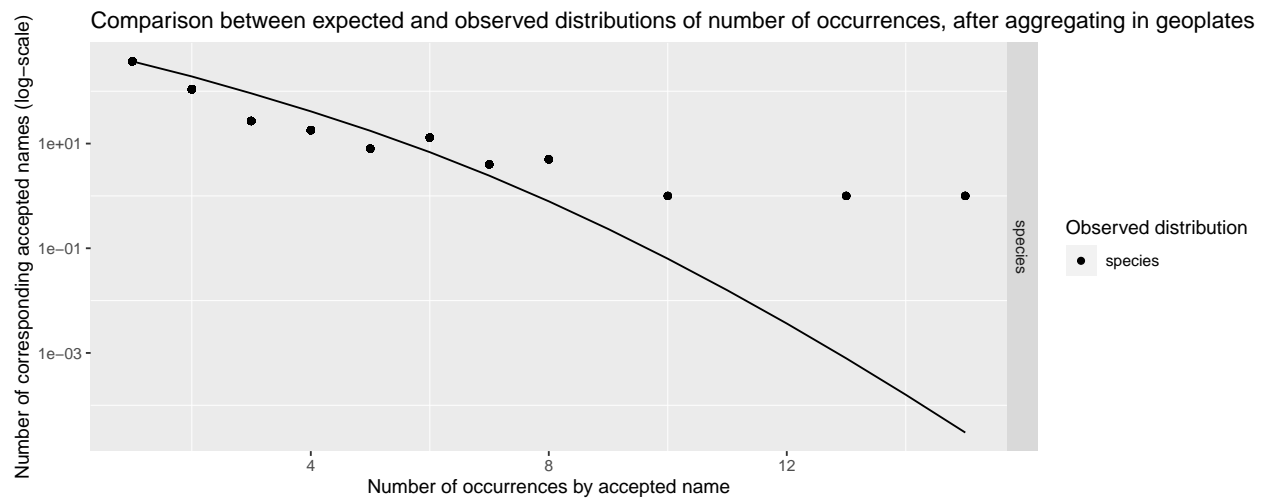
Replacing countries by geological plates seems to make more sens from a palaeontological perspective, so let's try it.

## Aggregate occurrences without formation by geoplate + early interval

	Cetacea_occ	Cetacea_occ_aggreg	removed
## Number of occurrences	3806	2605	1201
## Number of occurrences (species only)	1435	967	468

More occurrences remain after aggregating with this new method. Let's compare again with the theoretical distribution:

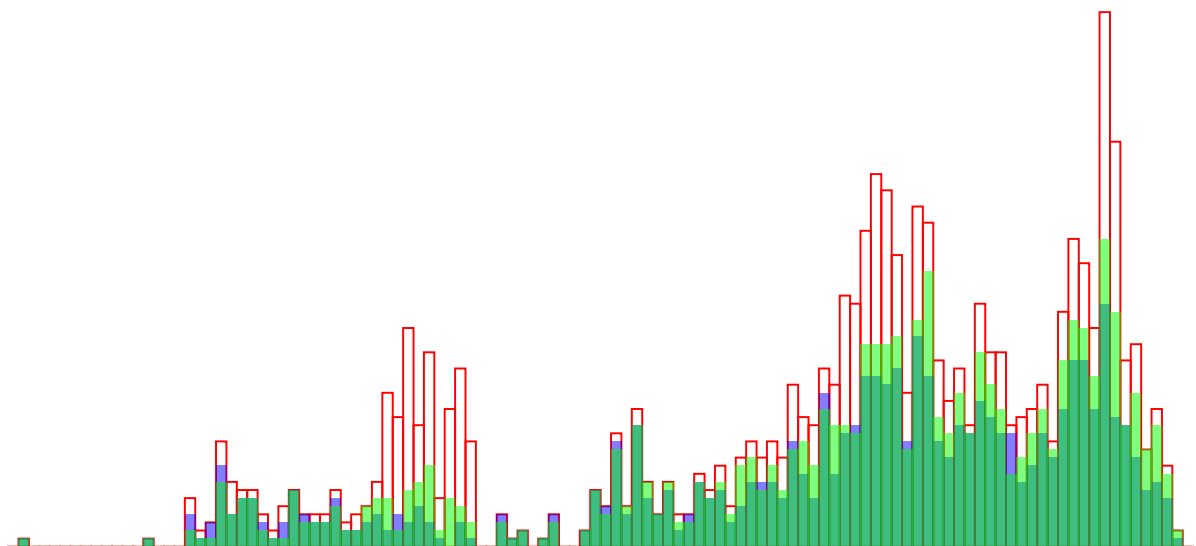




→ The correspondence is still good, except for two taxa:

```
##
##   Scaldicetus grandis Schizodelphis sulcatus
##               13                      15
## Warning: Removed 2 rows containing missing values (geom_bar).
## Warning: Removed 2 rows containing missing values (geom_bar).
## Warning: Removed 2 rows containing missing values (geom_bar).
```

Initial occurrences distribution (red) and comparison after sub-sampling (blue) or aggregating in geoplates (green)

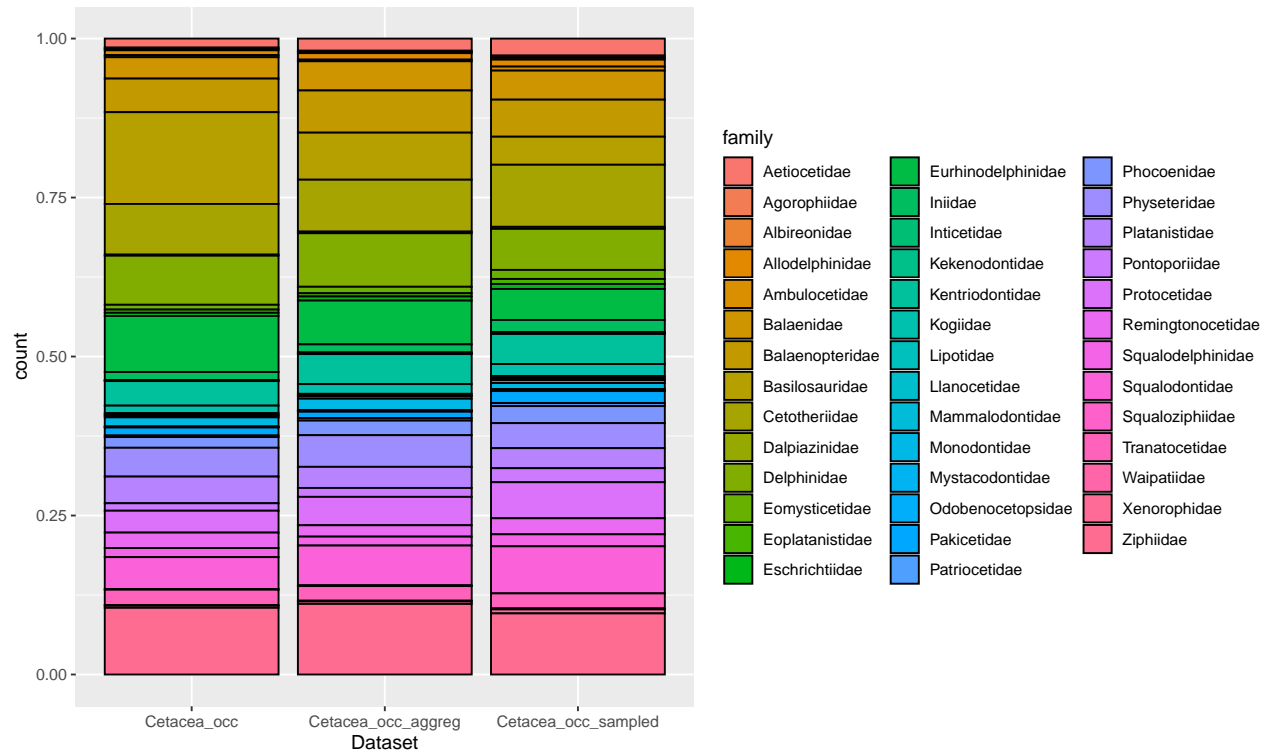


→ Delimiting by geological plates (+ age) instead of countries (+ age) leads to similar distributions, so we will keep it.

## Check that the sampling methods do not introduce biases in the repartition between Odontoceti and Mysticeti

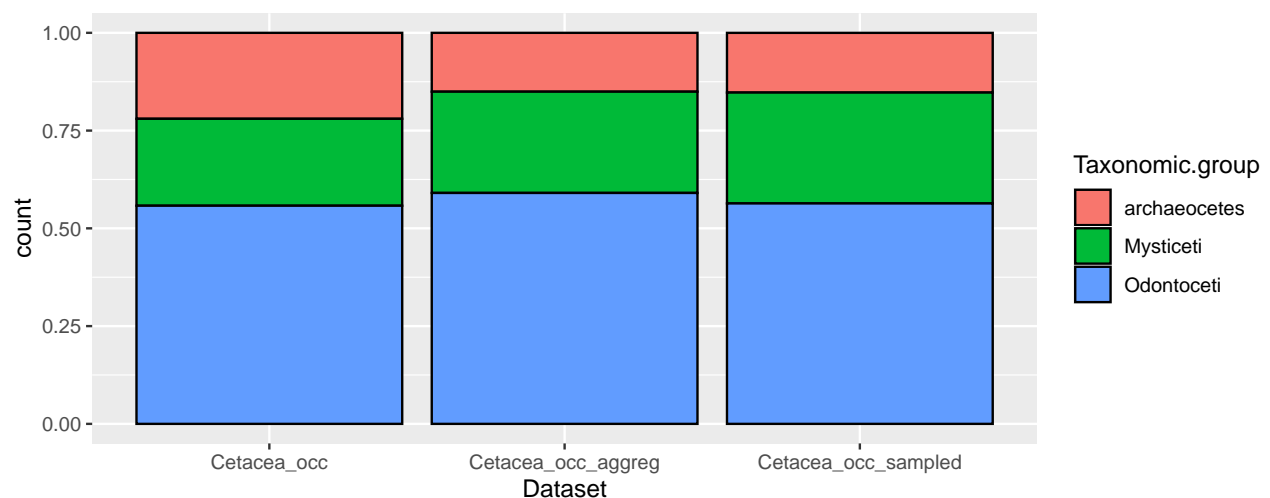
Mystecetes are usually larger than odontocetes, and size is associated with a wider geographic range so since we are subsampling occurrences according to geological formation we may be biasing our data towards more widespread species, therefore towards mystecetes.

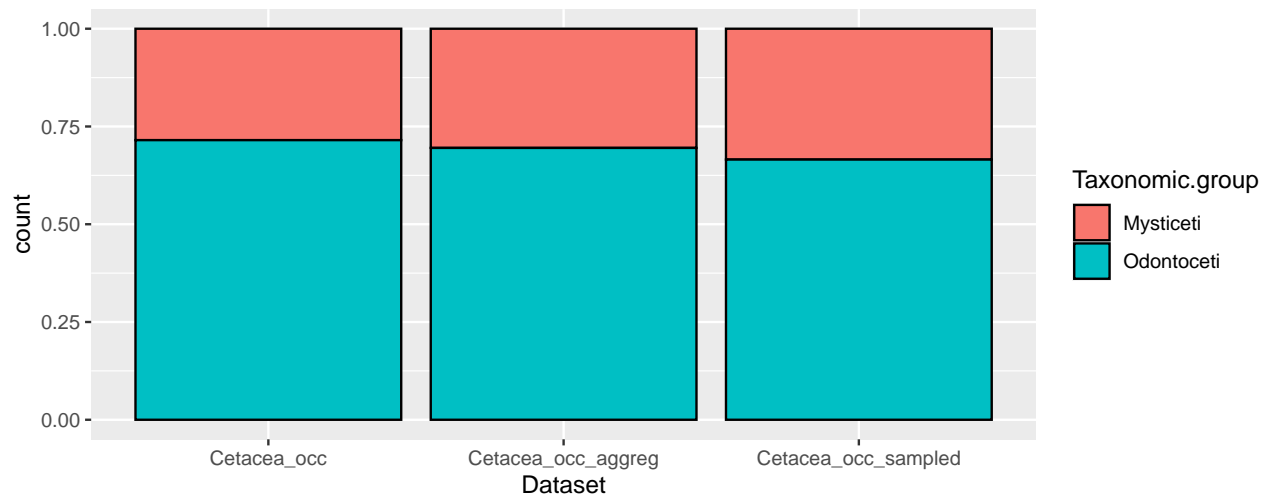
Look at the families first:



→ The family distributions vary a bit after subsampling, but because of the limited number of species by family the fact that we corrected the oversampling of some species could have a disproportionate effect.

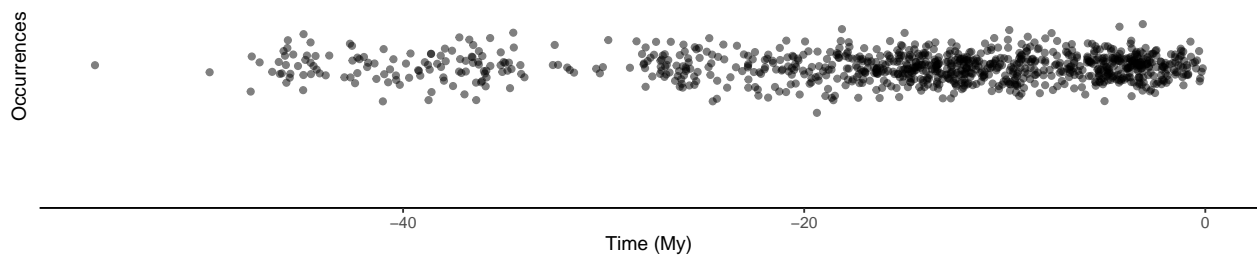
Let's look rather at a higher taxonomic rank, by importing the topology of cetacean families (from Marx et al. 2016):





→ There is a smaller proportion of archeocete occurrences after either sampling, because a huge cluster is subsampled around 35 My ago, but this effect is expected. However, for the Mysticeti vs. Odontoceti there is no huge apparent bias, especially with the aggregating method.

Repartition of 4609 recorded occurrences through time

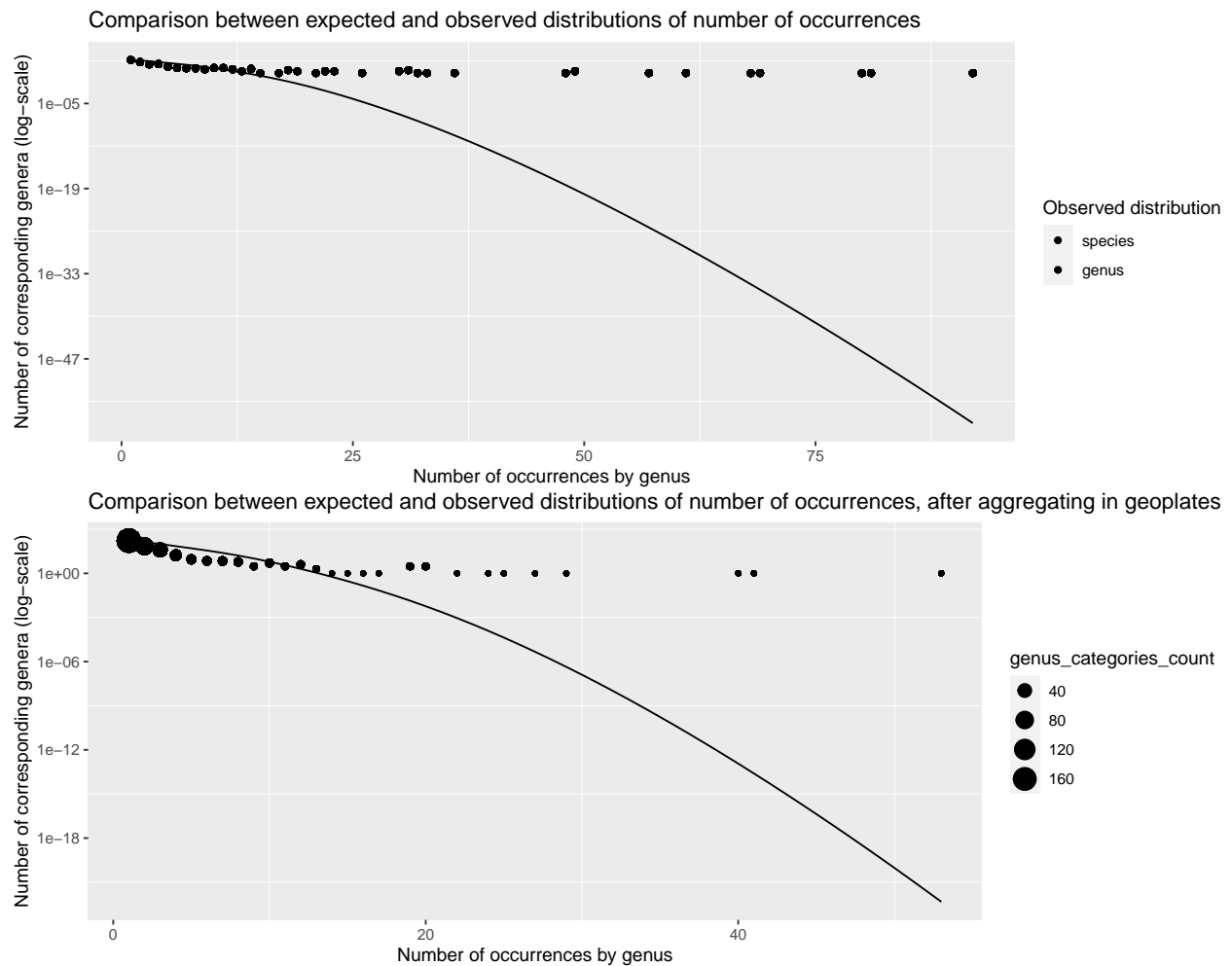


## Faster genus-level analysis

Apply the same transformations at the generic level to perform the analysis faster.

```
##                               Cetacea_occ
## Number of occurrences          3806
## Number of occurrences (species and genera only) 2160
##                               Cetacea_occ_aggreg_gen removed
## Number of occurrences          1836    1970
## Number of occurrences (species and genera only) 1300    860
```

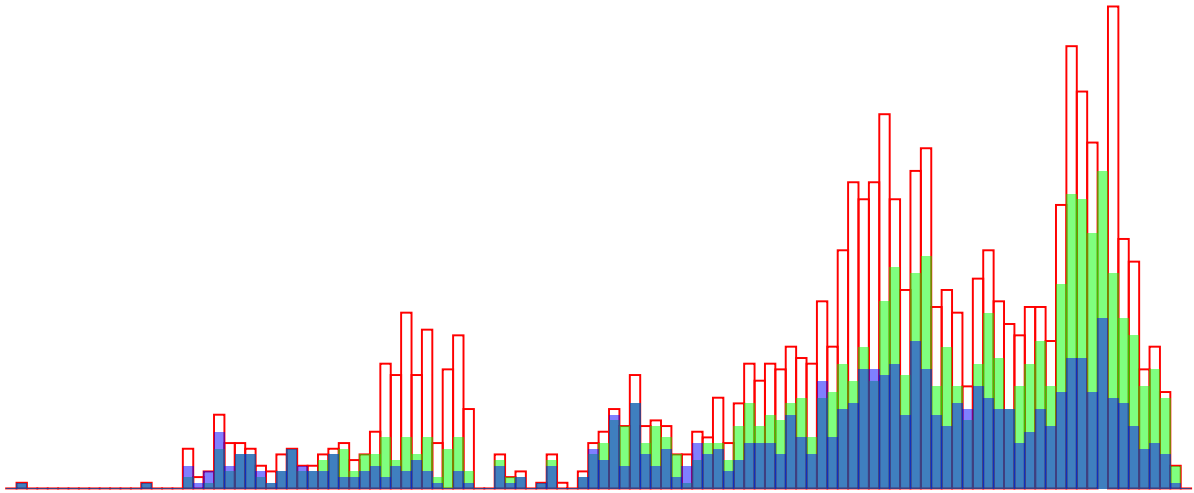
More occurrences remain after aggregating with this new method. Let's compare again with the theoretical distribution:



→ The correspondence is much less improved than with species aggregation because some genera have a lot of occurrences due to their high number of species:

```
##
## Balaenoptera   Kentriodon   Mesoplodon   Scaldicetus   Squalodon
##           53           27           29           41           40
## Warning: Removed 2 rows containing missing values (geom_bar).
## Warning: Removed 2 rows containing missing values (geom_bar).
## Warning: Removed 3 rows containing missing values (geom_bar).
```

Initial occurrences distribution (red) and comparison after sub-sampling (blue) or aggregating in geoplates (green)



→ Delimiting by geological plates (+ age) instead of countries (+ age) leads to similar distributions, with a bit more occurrences.

## Conclusions

Several biases affect the cetacean fossil record:

- BIAS 1: Very recent samples may have been dated with a more precise method and contain much more fossils, so they should be removed or treated separately
- BIAS 2: Some occurrences show a highly uncertain dating ( $> 10\text{My}$ ), they are excluded when combining the uncertainty intervals into a unique stratigraphic range per taxon
- BIAS 3: Some taxa and geological periods seems to have been oversampled, due to variation in species abundance, geological factors or human unequal sampling,
  1. It seems possible to adequately reduce the abundance bias by subsampling the richest taxa, based on a metrics of occurrence density (number of occurrences / stratigraphic range). Using combined stratigraphic ranges, instead of age uncertainty intervals, appears to be more robust
  2. An alternative sampling method consisting in aggregating the occurrences of the same taxa found in the same geological formation (or the same geoplate + early stratigraphic interval if the geological formation is not indicated) seems to work equally well, and is more paleontologically relevant