Analysis of Malaria API Variability in the Brazilian Amazon

Daniel Villela, Raquel Lana

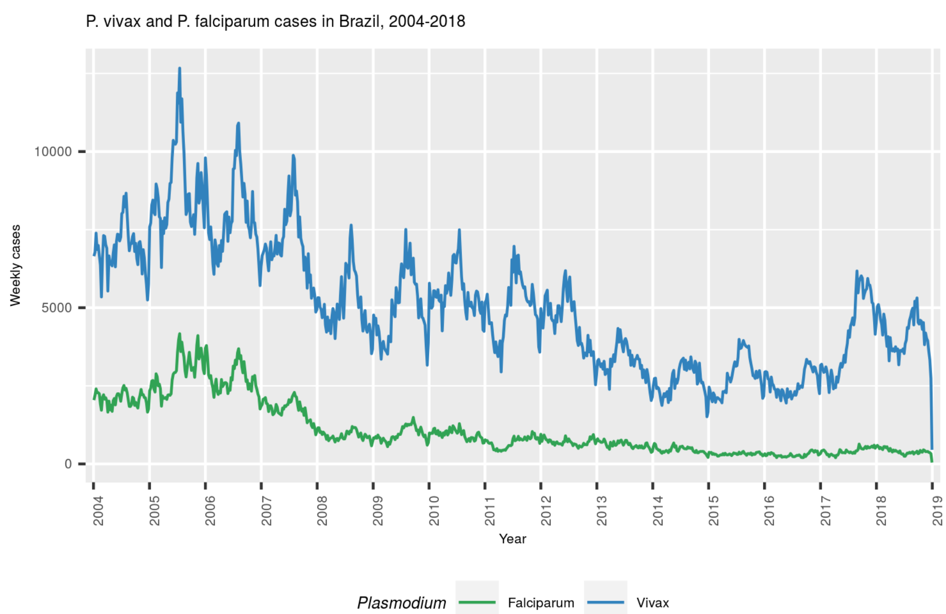
May 10, 2019

Analysis of the coefficient of variation of API in the Amazon using SIVEP data with several political adminstritive levels:

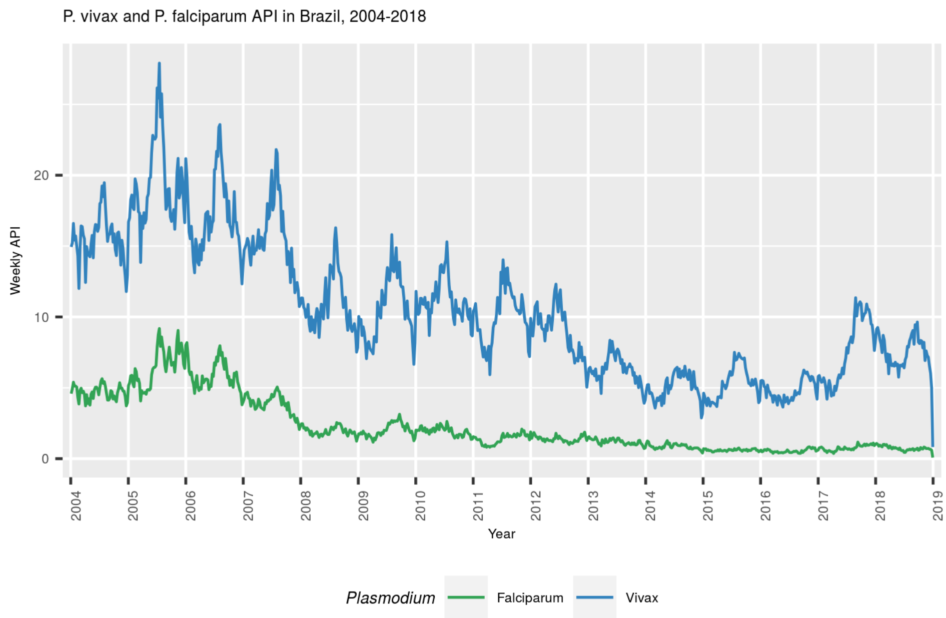
Municipality, state and Amazon Basin

First, an update on API and Cases time series up to the end of 2018

- Cases

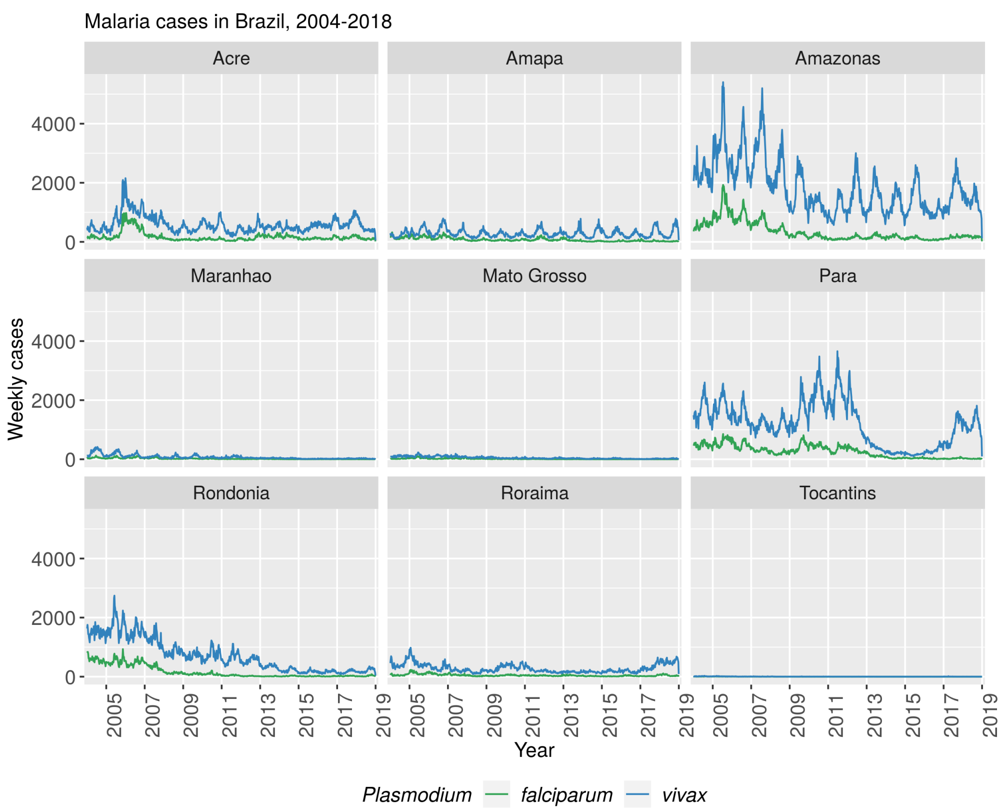


- API (computed on weekly number of cases, but scaled to yearly values)

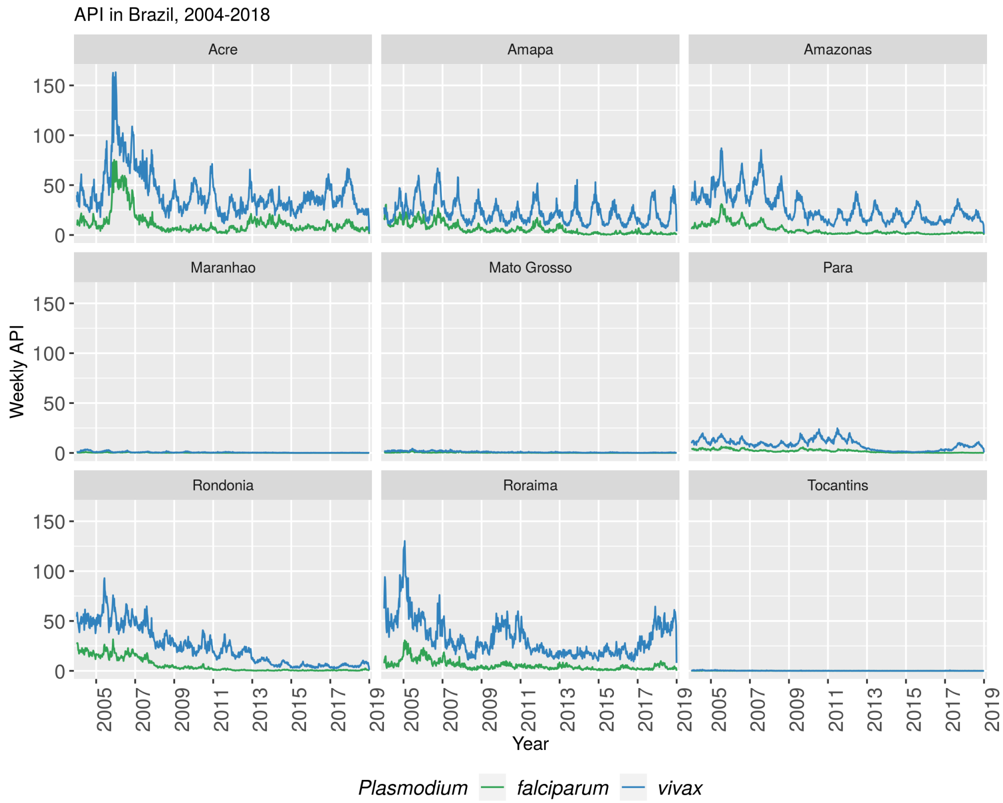


Comparing the states

- Cases



- API



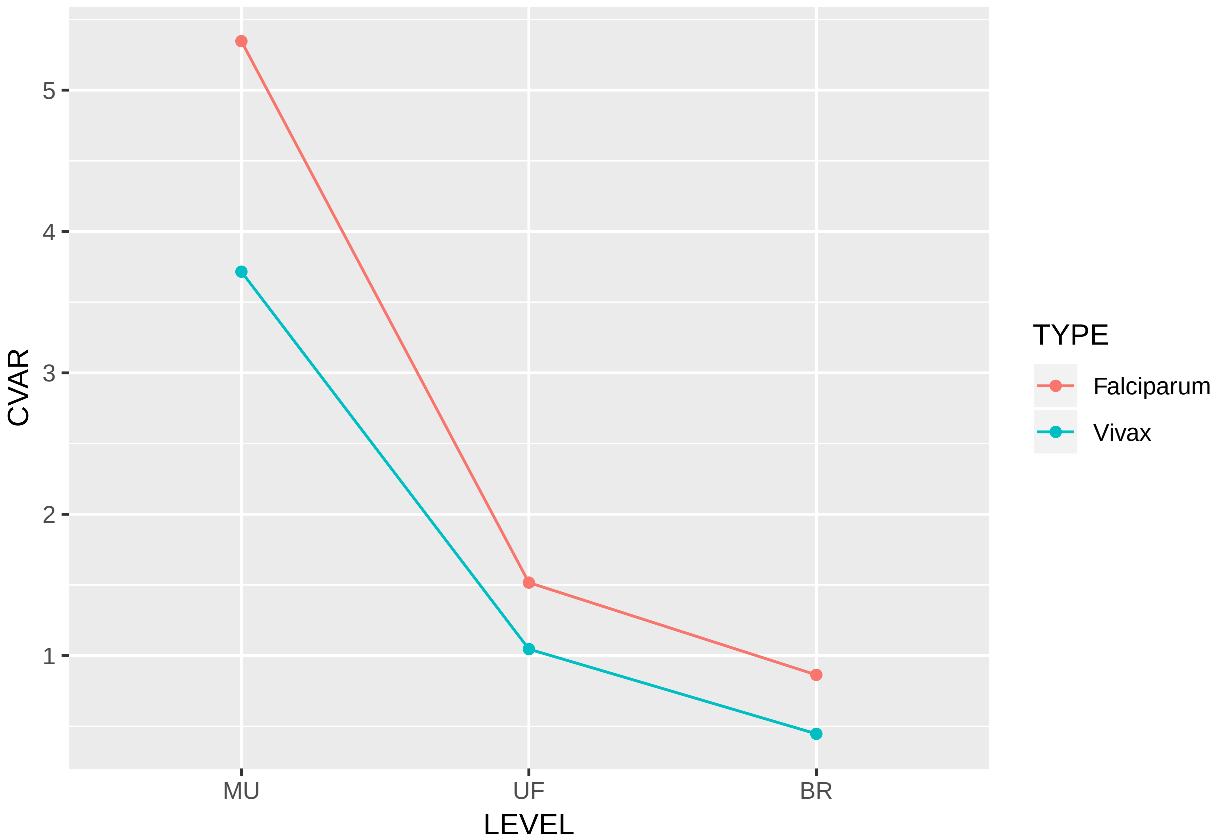
Coefficient of variation

Since we obtained the API for each unit (municipality, state or Amazon), we can compute several summarising indices such as:
- mean over all municipalities
- mean over all municipalities of a given state
- mean over all municipalities in a given month of the year
- standard deviation over all municipalities
- standard deviation over all municipalities of a given state
- standard deviation over all municipalities in a given month of the year

This allows us to obtain the coefficient of variation which is the ratio between the standard deviation and the mean.

The coeficient of variation declines when we analyse APIs in the municipality level, state level and amazon level, in this order, which shows large variability of the API in the municipality level.
For falciparum, the standard deviation is about 4 times the mean.

```
ggplot(aes(x=LEVEL, y=CVAR, color=TYPE, group=TYPE), data=dNotix) +
  geom_point() + geom_line()
```

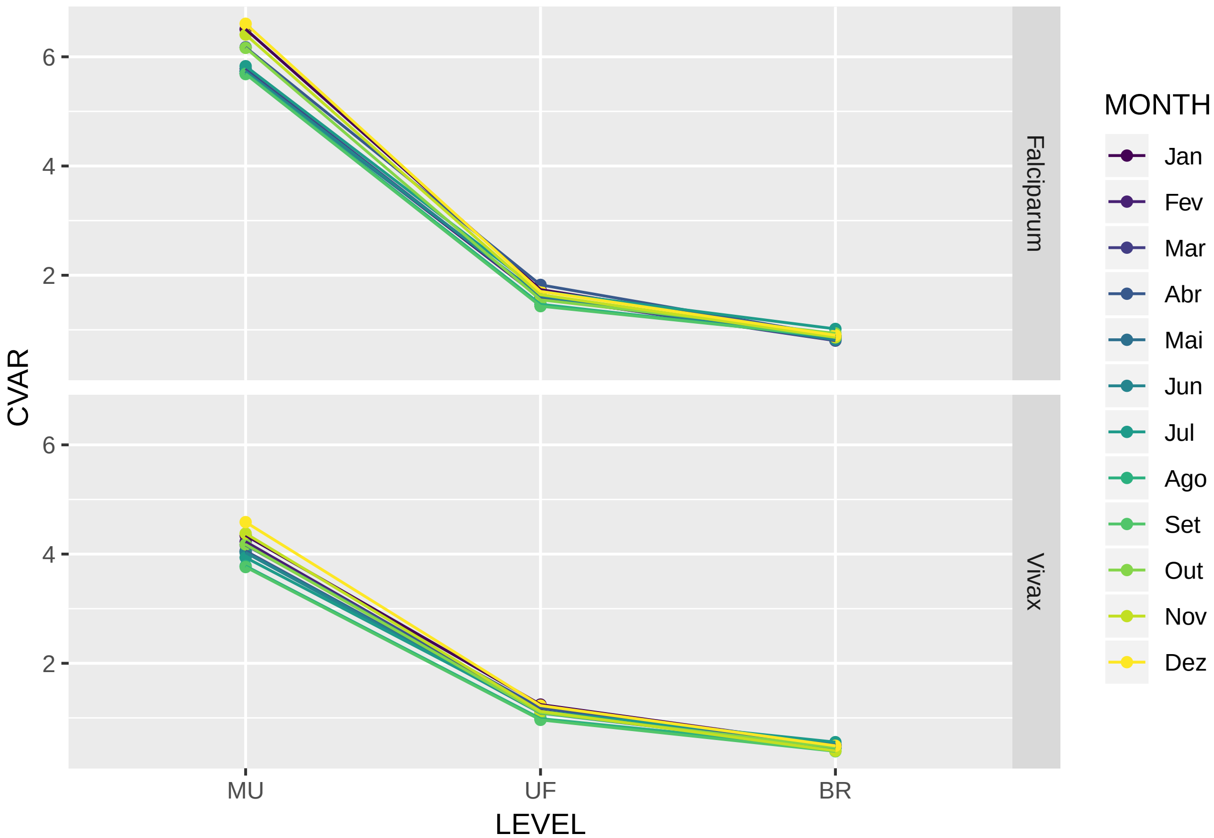


Is this variability the same over all months of the year?

The previous analysis was using the API in the years from 2004 (yearly values).
Here we analyze monthly values of API (scaled by 12 to consider what would be the value assuming this value constant).
Therefore, API of january, API of february etc.
This is to obtain the coefficient of variation separated by month.

Basically we observe the same effect as in the previous one.
It seems that the summer months exhibit greater variation.

```
ggplot(aes(x=LEVEL, y=CVAR, color=MONTH, group=MONTH),
       data=dNoti.mx) + facet_grid(TYPE ~ .) +
  geom_point() + geom_line()
```

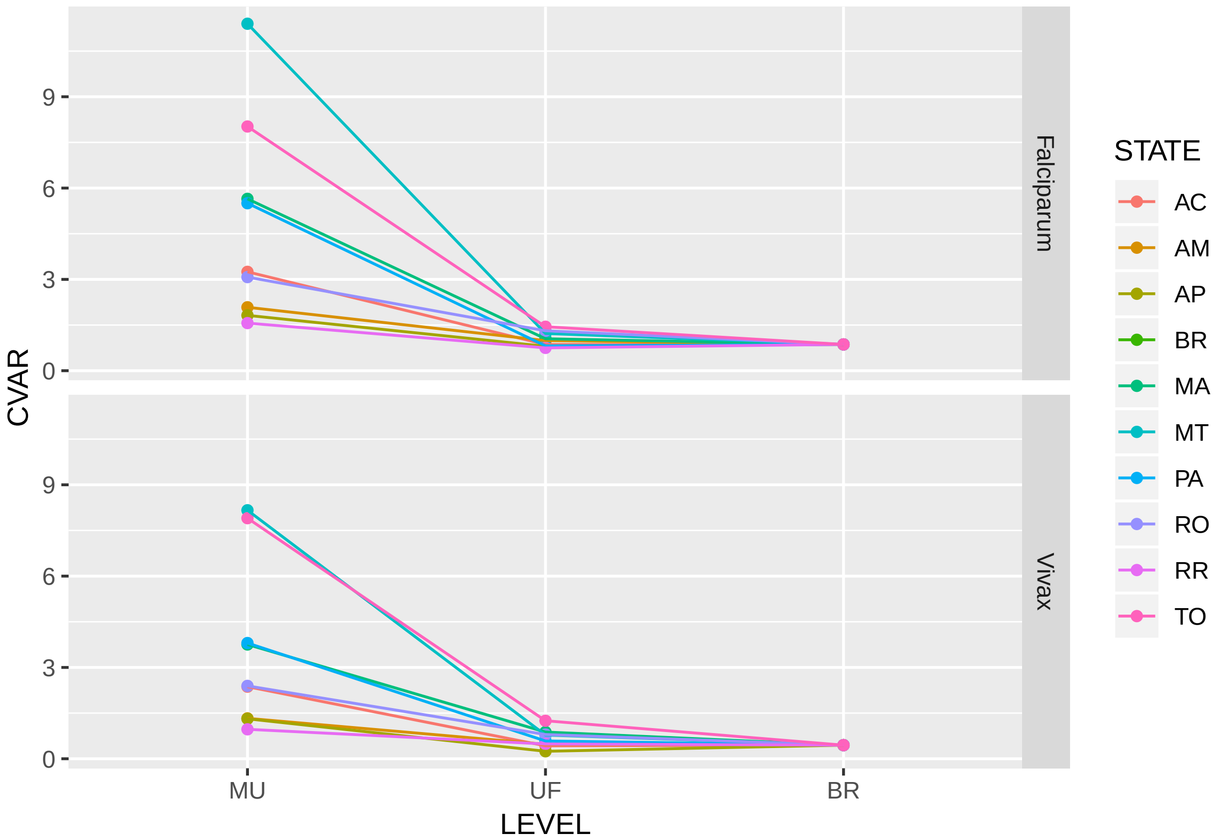


STATE

Now were are going to analyze the coefficient of variation but separating into groups of states.

We see about the same effect but it is useful to see which states have great variability in the municipality level.

```
ggplot(aes(x=LEVEL, y=CVAR, color=STATE, group=STATE),
       data=dNoti.sx) + facet_grid(TYPE ~ .) +
  geom_point() + geom_line()
```



Levels of Population sizes

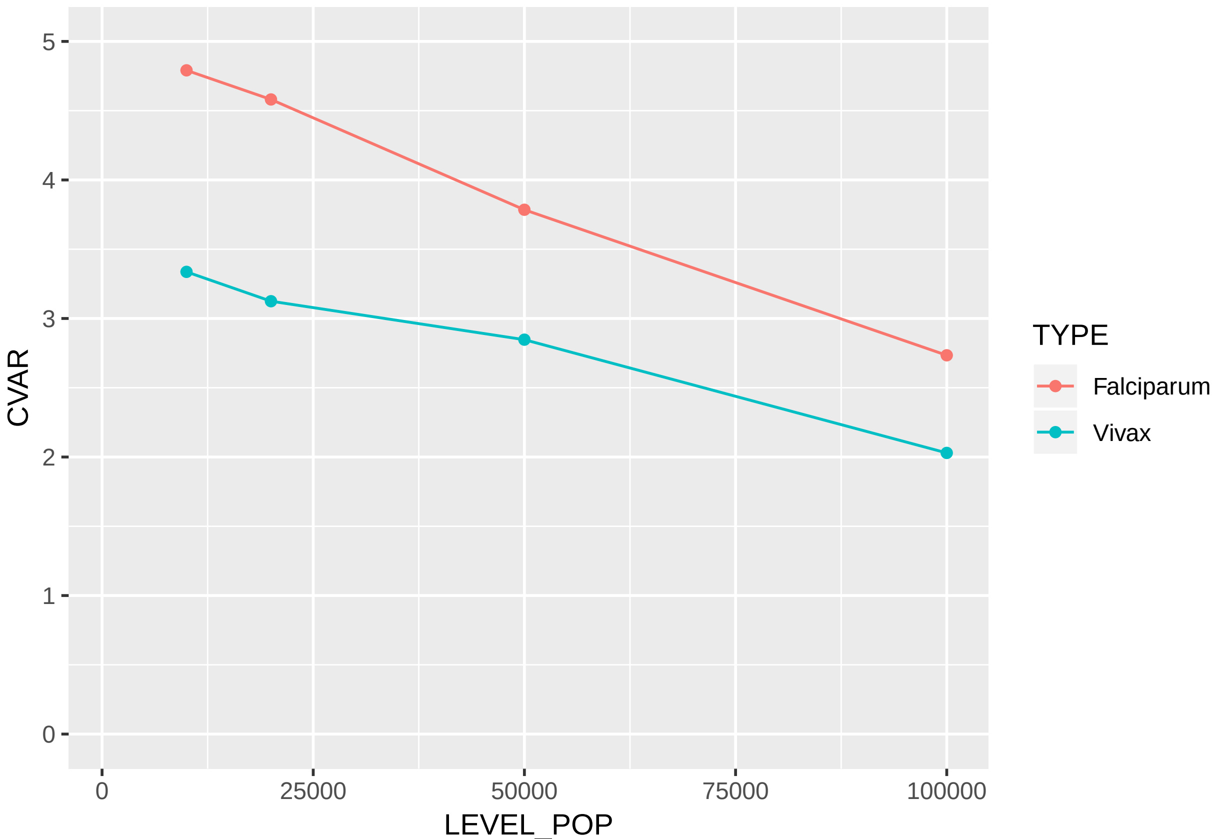
Since we had defined different population levels:
Another option of levels of population sizes is by the limits in each stratum

```
above 100000  ~ 100000,
50000 - 100000 ~ 50000,
20000 - 50000 ~ 20000,
10000 - 2000 ~ 10000,
1000 - 10000 ~ 1000
```

We analyzed the coefficient of variation in these levels.
This makes possible to separate municipalities into different groups, but puts together state and municipality (problem solved ahead).

Now it is interesting that we observe the coefficient of variation to be approximately constant, but we still have some municipalities together with some states and the BR level.

```
options(scipen = 10000) # to be numeric
ggplot(aes(x=LEVEL_POP, y=CVAR, color=TYPE, group=TYPE), #color=MONTH, group=MONTH),
       data=dNoti.lx) + #facet_grid(TYPE ~ .) +
  geom_point() + geom_line() + ylim(c(0,5))
```

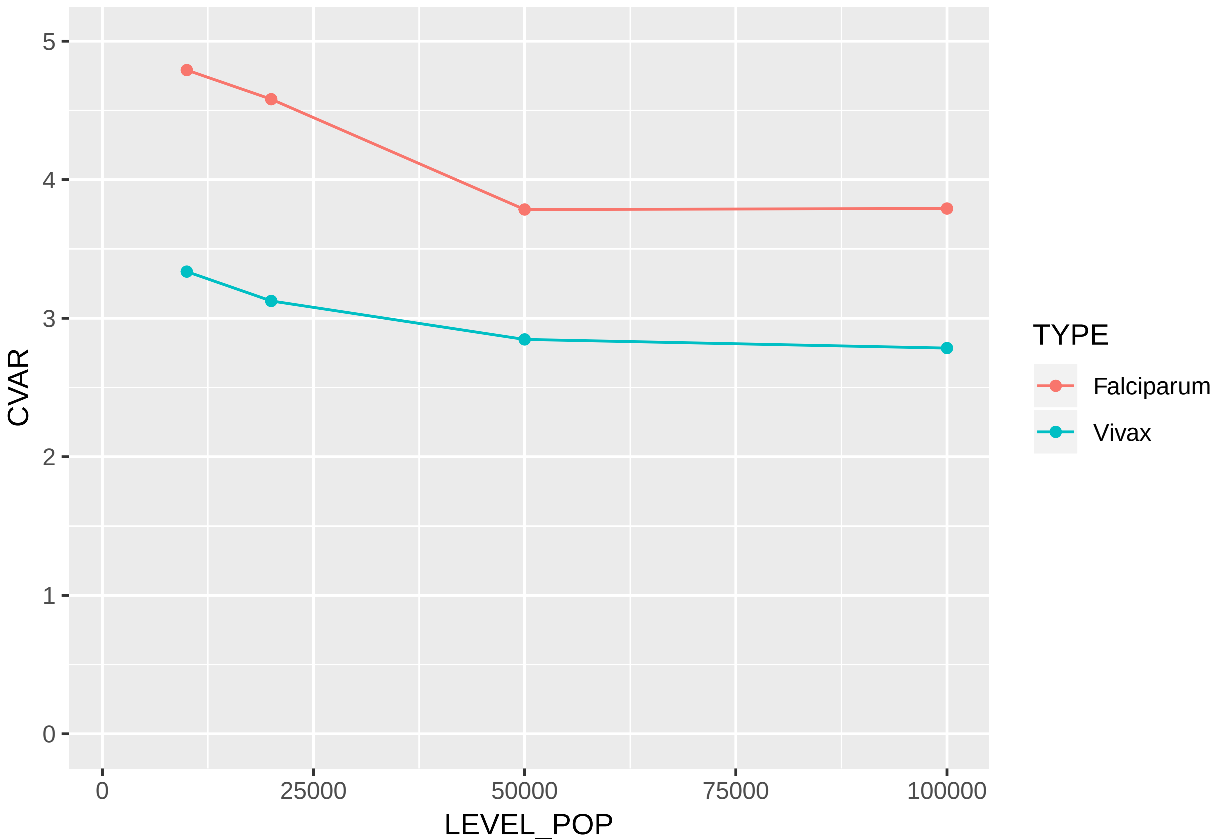


So we excluded the UF and BR levels.

And we observe an interesting property: for all municipalities the coefficient of variation is constant. We observe the same variability in groups of municipality of 1000 inhabitants or groups of 100000 persons.

Very interesting. Even when we try to make homogeneous strata of municipalities, we still see great variability of API and pretty much constant! Standard deviation is greater than three times the mean for *P. falciparum* malaria and little less than three times the mean for vivax malaria.

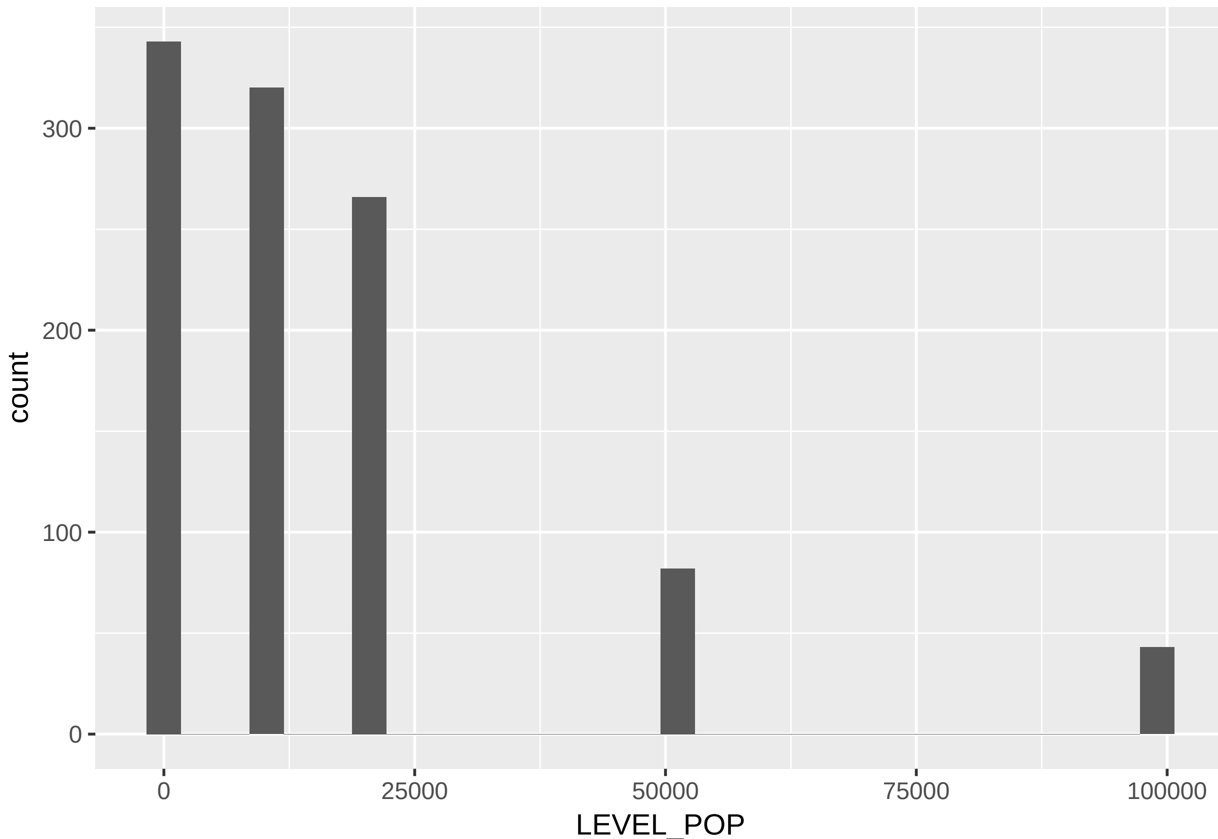
```
ggplot(aes(x=LEVEL_POP, y=CVAR, color=TYPE, group=TYPE), #color=MONTH, group=MONTH),
       data=dNoti.lx) + #facet_grid(TYPE ~ .) +
  geom_point() + geom_line() + ylim(c(0,5))
```



Just to check the population structure of municipalities.

First, the counts in each stratum of municipalities.

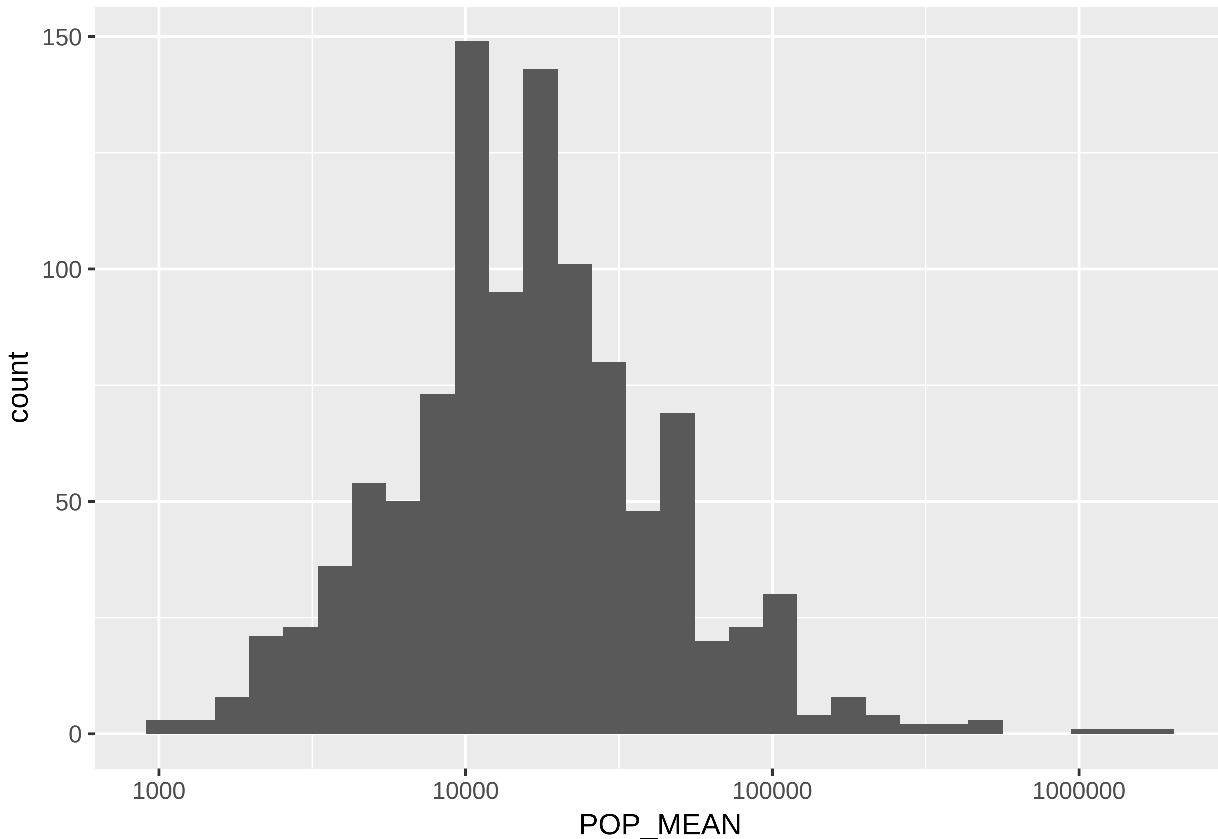
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Second, the histogram of population sizes given by the mean over the years in the studied period of time. A lognormal distribution appears to fit quite nicely to describe the population in municipalities of the Brazilian Amazon.

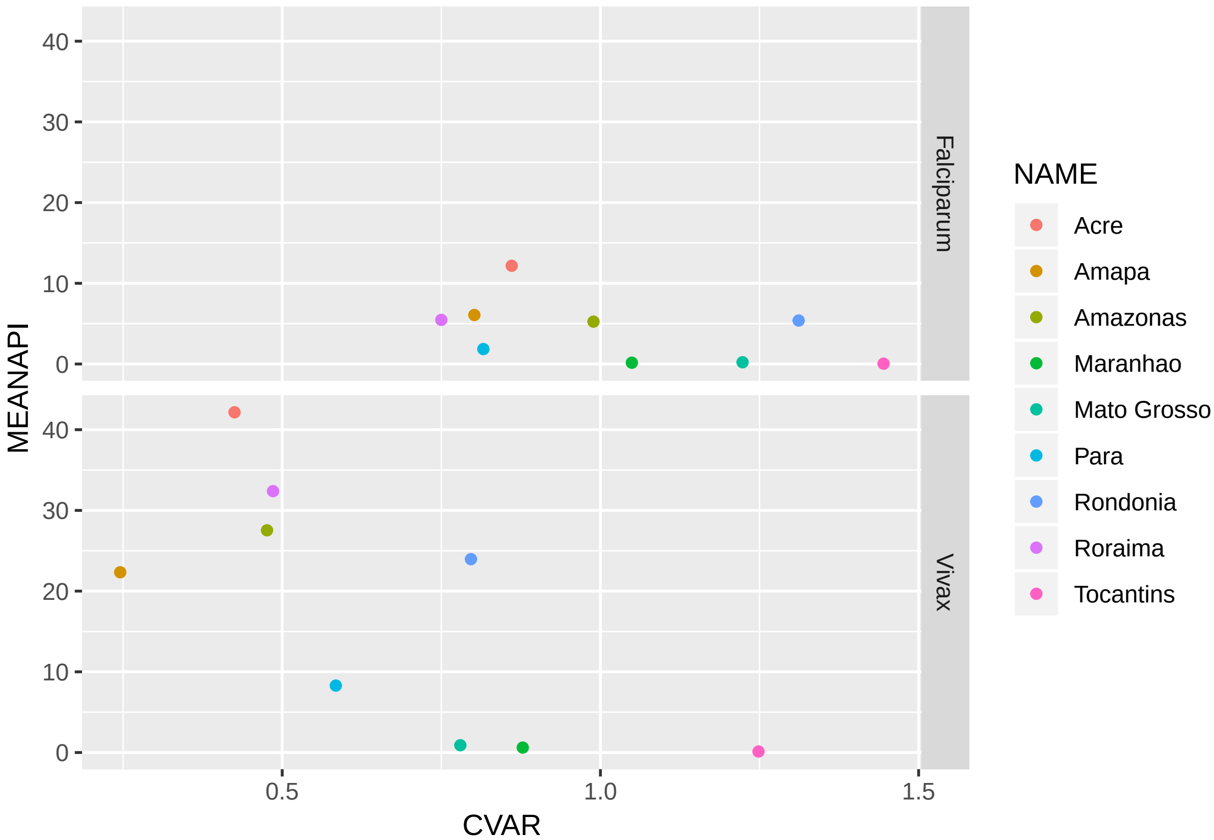
```
## Scale for 'x' is already present. Adding another scale for 'x', which
## will replace the existing scale.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

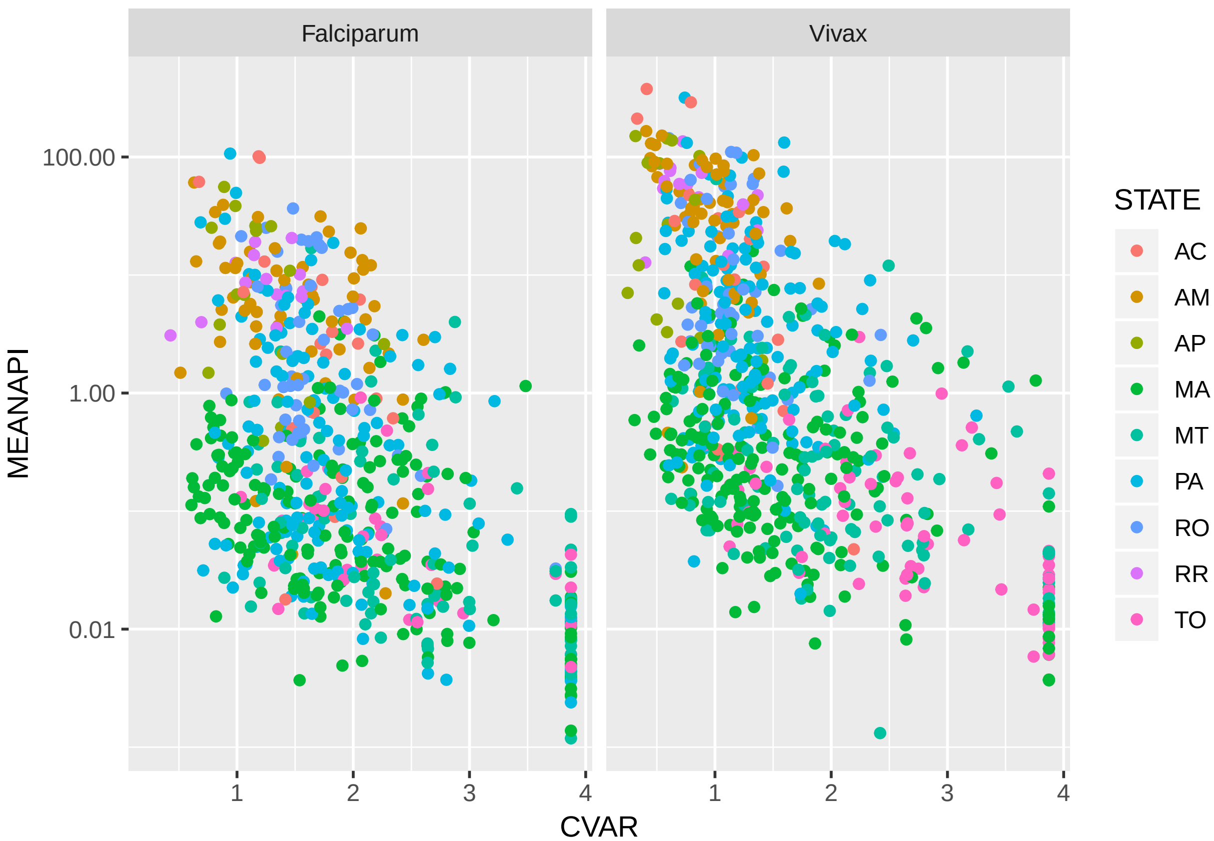


Analysis of Mean vs Coefficient of Variation

We proceed to evaluate how the mean API and coefficient of variation varied at state and municipality level over time.

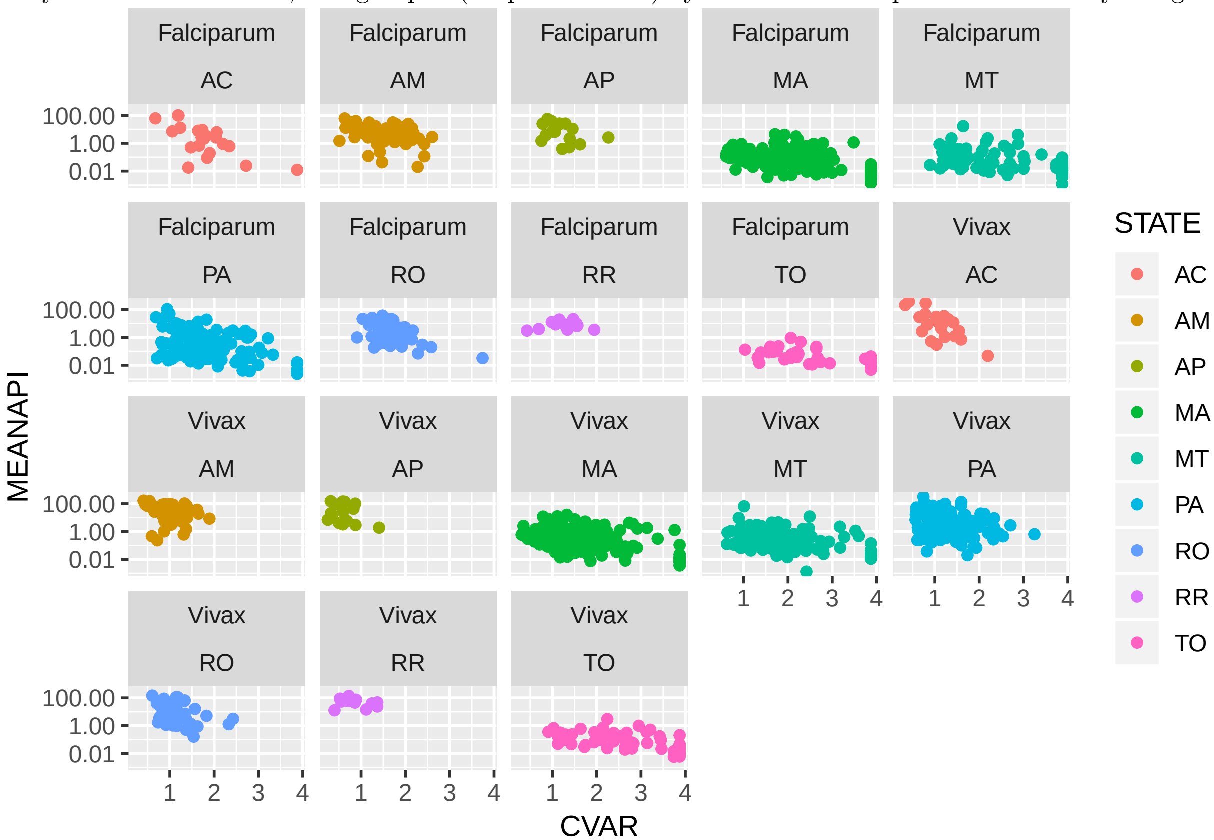


MEANAPI
CVAR
Falciparum
Vivax
NAME
Acre
Amapa
Amazonas
Maranhao
Mato Grosso
Para
Rondonia
Roraima
Tocantins

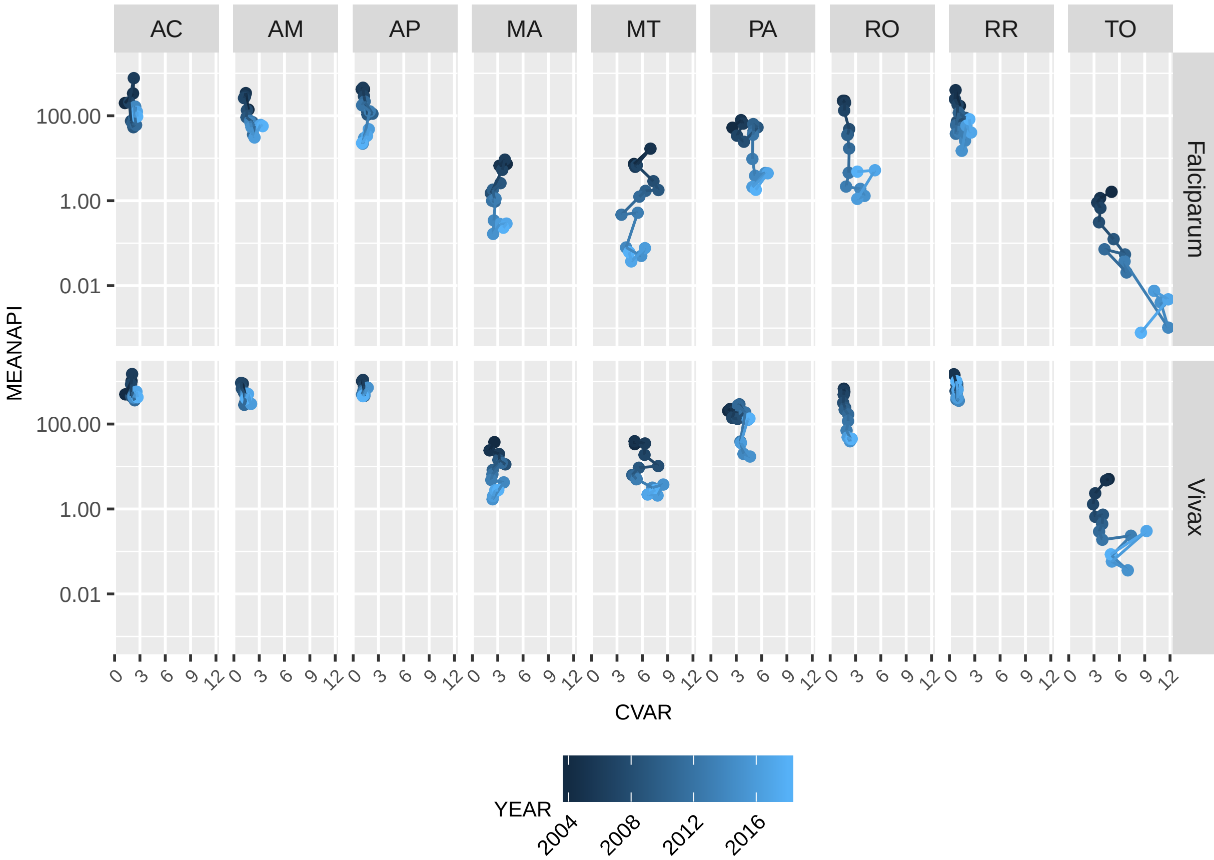


Falciparum
Vivax
MEANAPI
CVAR
STATE
AC
AM
AP
MA
MT
PA
RO
RR
TO

Here we are investigating only municipalities, hence we calculated API and CV for all municipalities, across all years in the database, and grouped (shape and color) by state. Different patterns are hardly recognized.



So we separated the point "clouds" by states and *falciparum/vivax*. Again not sure about states TO, RO, RR - we have to verify.



Now for each year we obtain the API for state and CV of API across all municipalities.
We connect some dots, in this case connecting by the sequence of years.
Now this plot is interesting to observe the patterns when we get to more recent years.

Potential endemic number

Let us define the concept of the potential for an epidemic, which will tell us some expected quantity (risk, number of new cases) based on the previous number of cases. This will be defined as the product of the number of cases and a function of the population size and the number of cases.
Hence, for $C_{i,y}$ cases for municipality i in year y we have for potential $P_{i,y}$: $P_{i,y} = C_{i,y} f(C_{i,j}, N_{i,y})$.
We should use a function that is decreasing with $C_{i,y}$ and increasing with $N_{i,y}$.
An intuition here is that the potential number behaves close to an expected new flow of cases or a next expected API from $C_{i,y}$ cases and a population of $N_{i,y}$.

From a simple susceptible-infected model, we have an upper bound on expected number of new cases given by $C_{i,y}(N_{i,y} - C_{i,y})$.

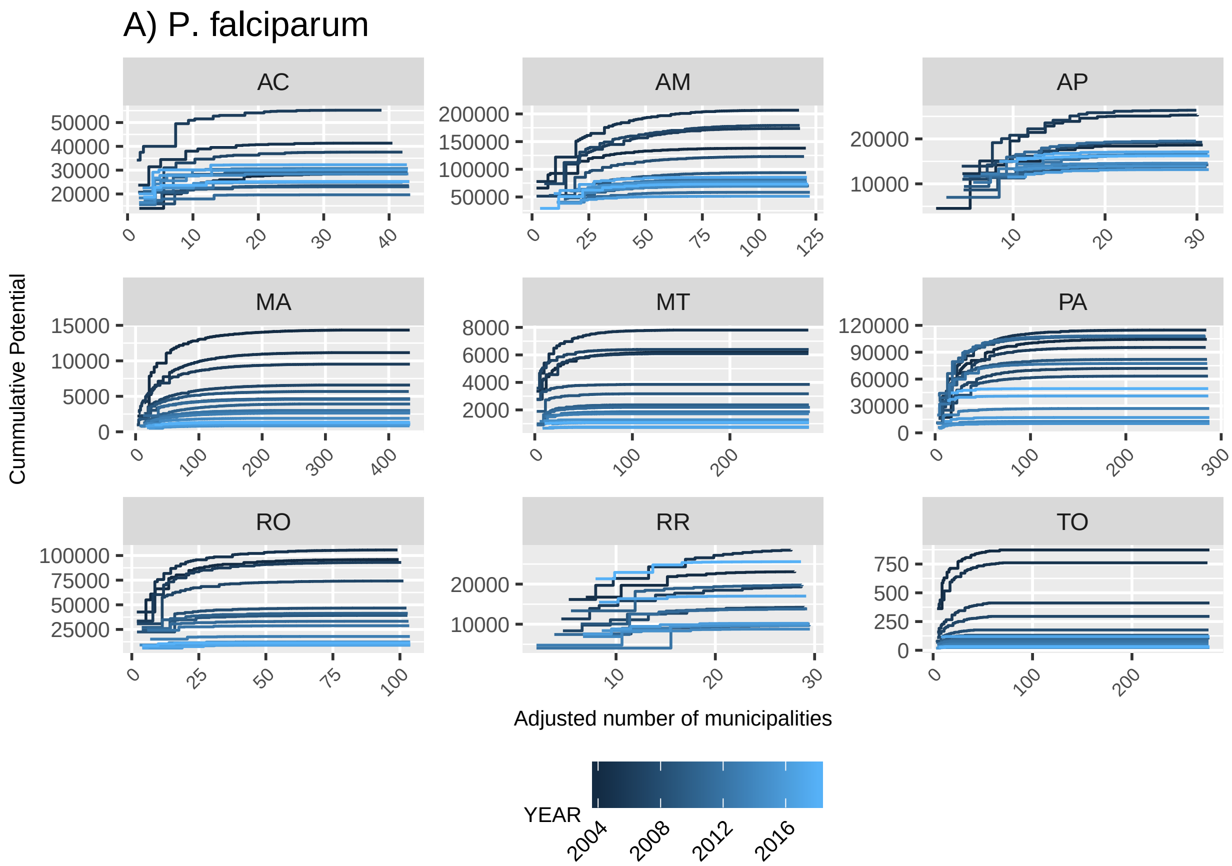
Following this approach, when divided by $N_{i,y}$, we have $P_{i,y} = C_{i,y}(1 - C_{i,y}/N_{i,y})$.

In this case, we can think about this potential as a product between the number of cases $C_{i,y}$ and a normalized number of susceptibles, given by $1 - C_{i,y}/N_{i,y}$, scaled by the population size as the fraction of susceptibles.

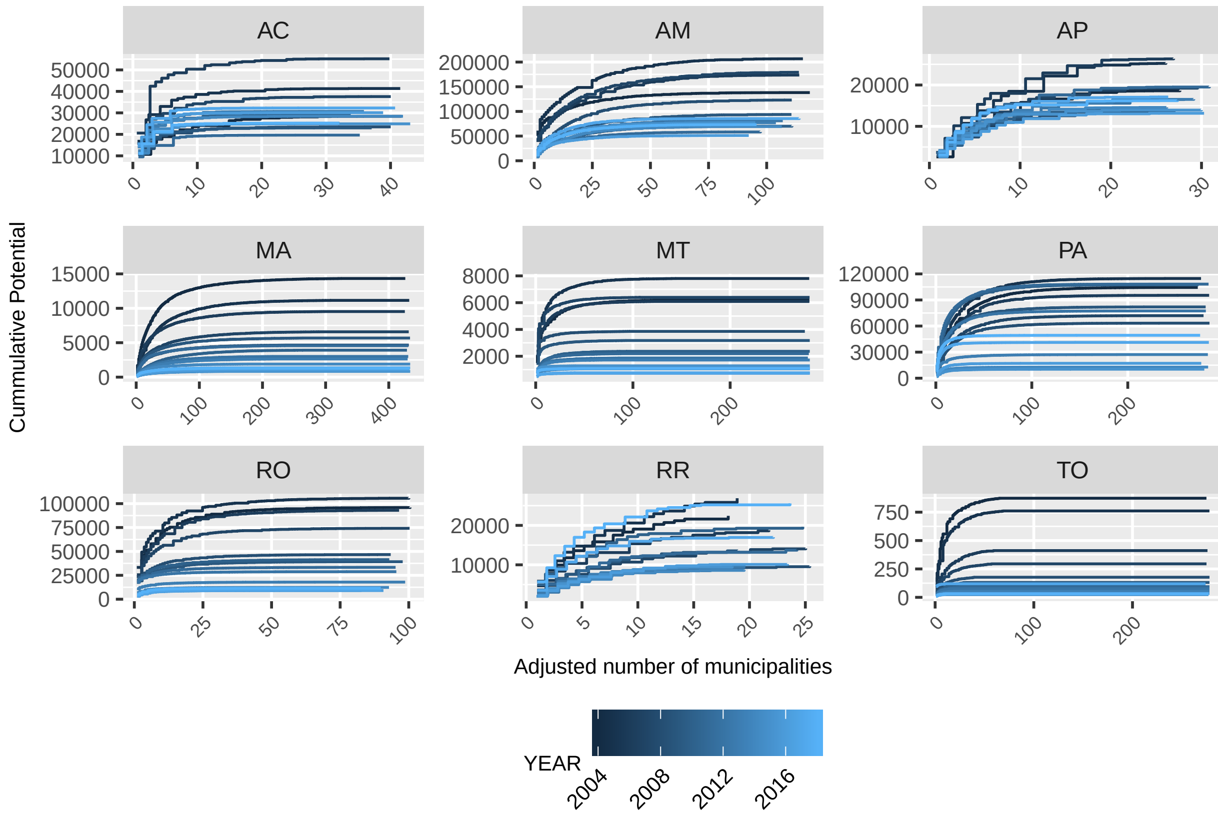
The idea is that for tackling interventions we can sort municipalities by their potentials.
When we sum up all potentials over all municipalities for given years, we can study how this evolves over time.

It will be useful to sort municipalities by descending potential values. We also compute cummulative potential, $\sum_{i=1}^{j} P_{i,y}$, and cummulative number of normalized susceptibles, say $J_{j,y} = \sum_{i=1}^{j}(1 - C_{i,y}/N_{i,y})$. The latter can also be viewed as an ajusted number of municipalities, say for a given state and/or Brazil. When cases are close to zero, the contribution of a municipality equals 1.

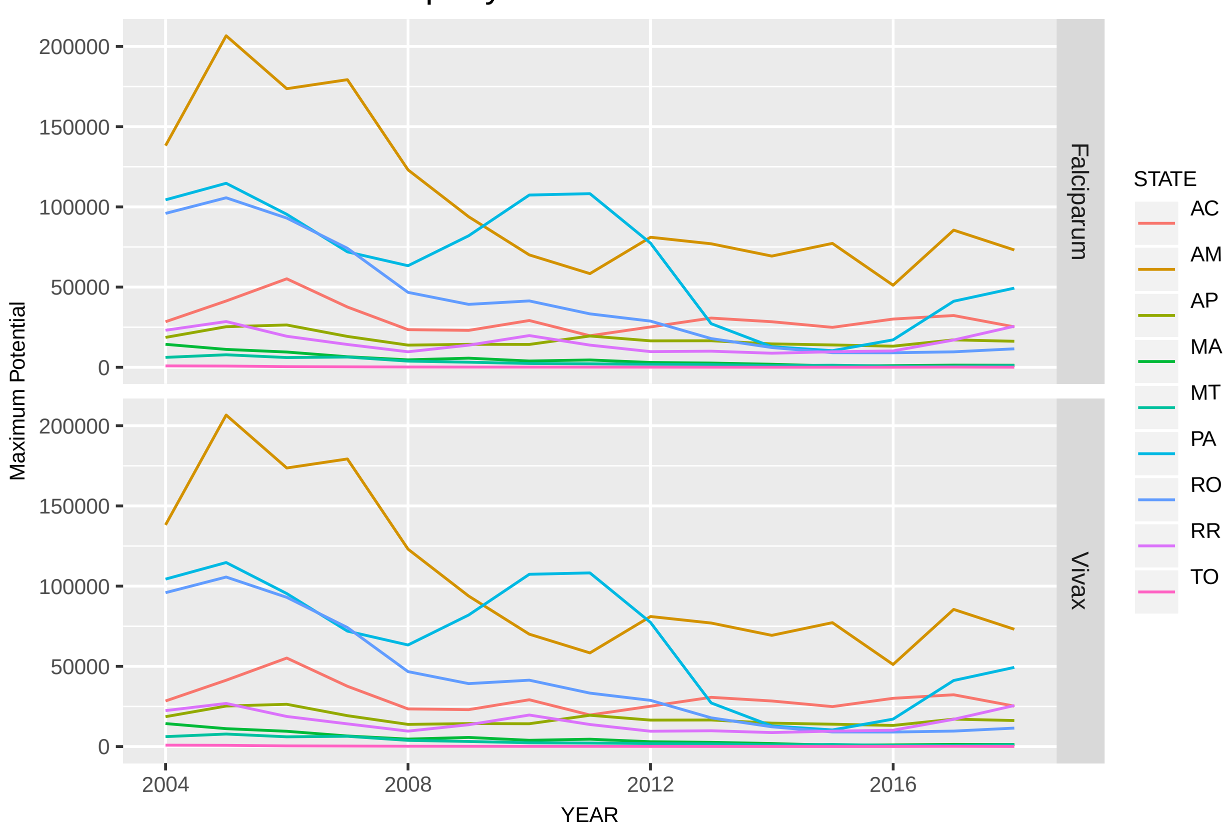
Therefore we will show some plots in which we show the potentials as a function of the cummulative pool of susceptibles as well as plot for the cummulative potential and normalized cummulative potential.



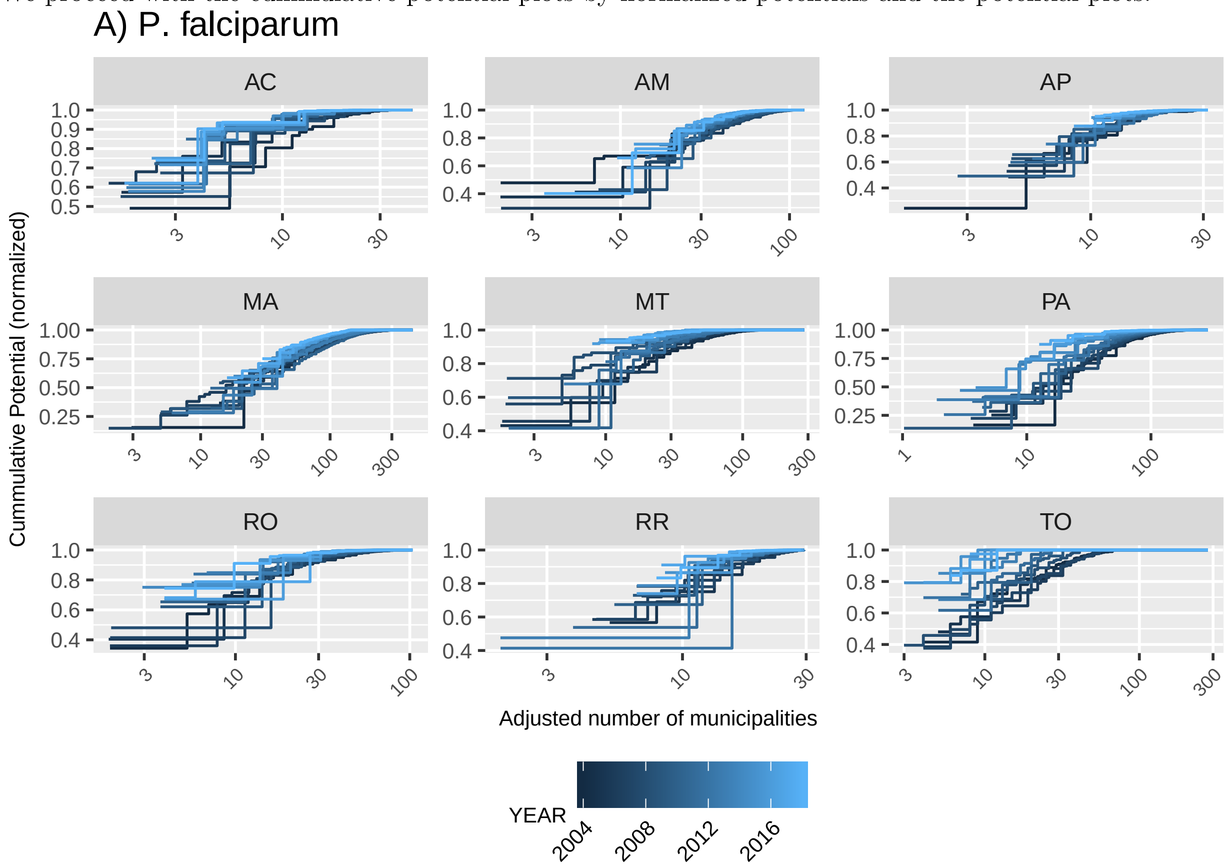
B. P. vivax



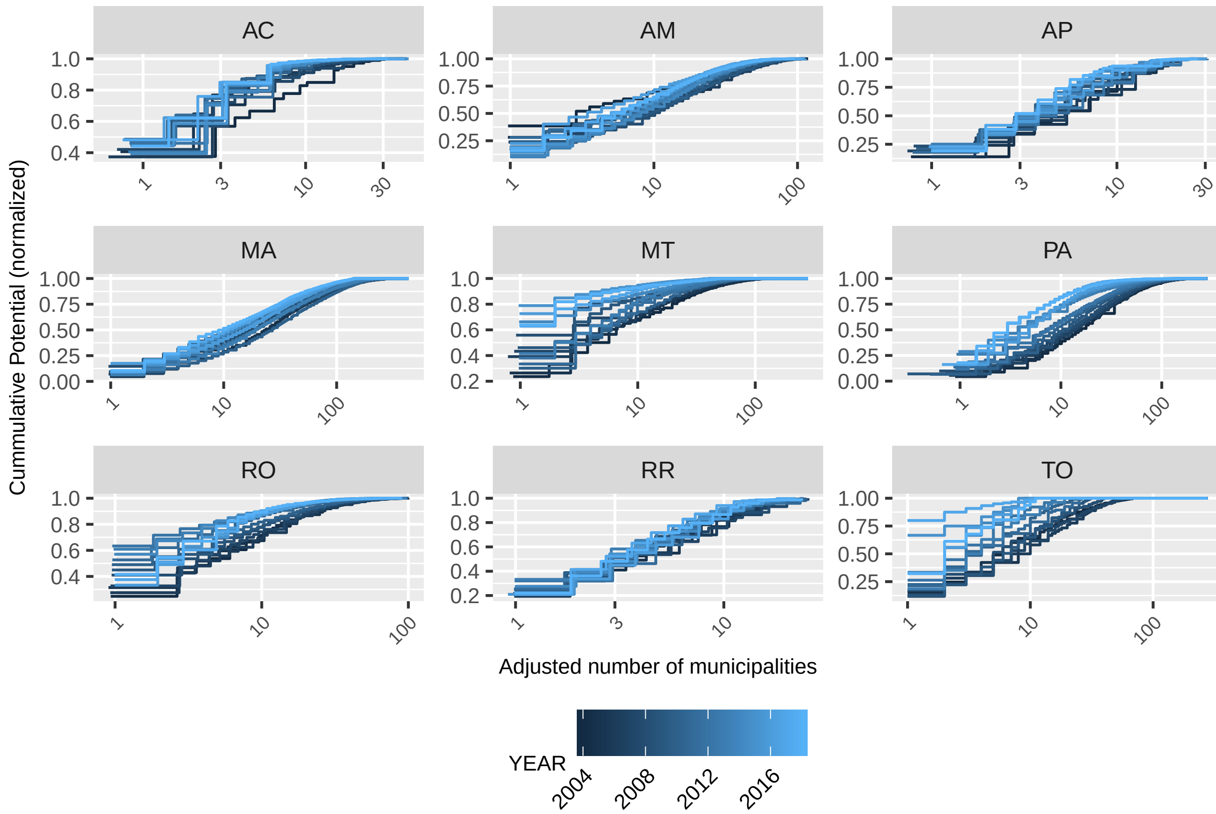
Maximum Potential per year



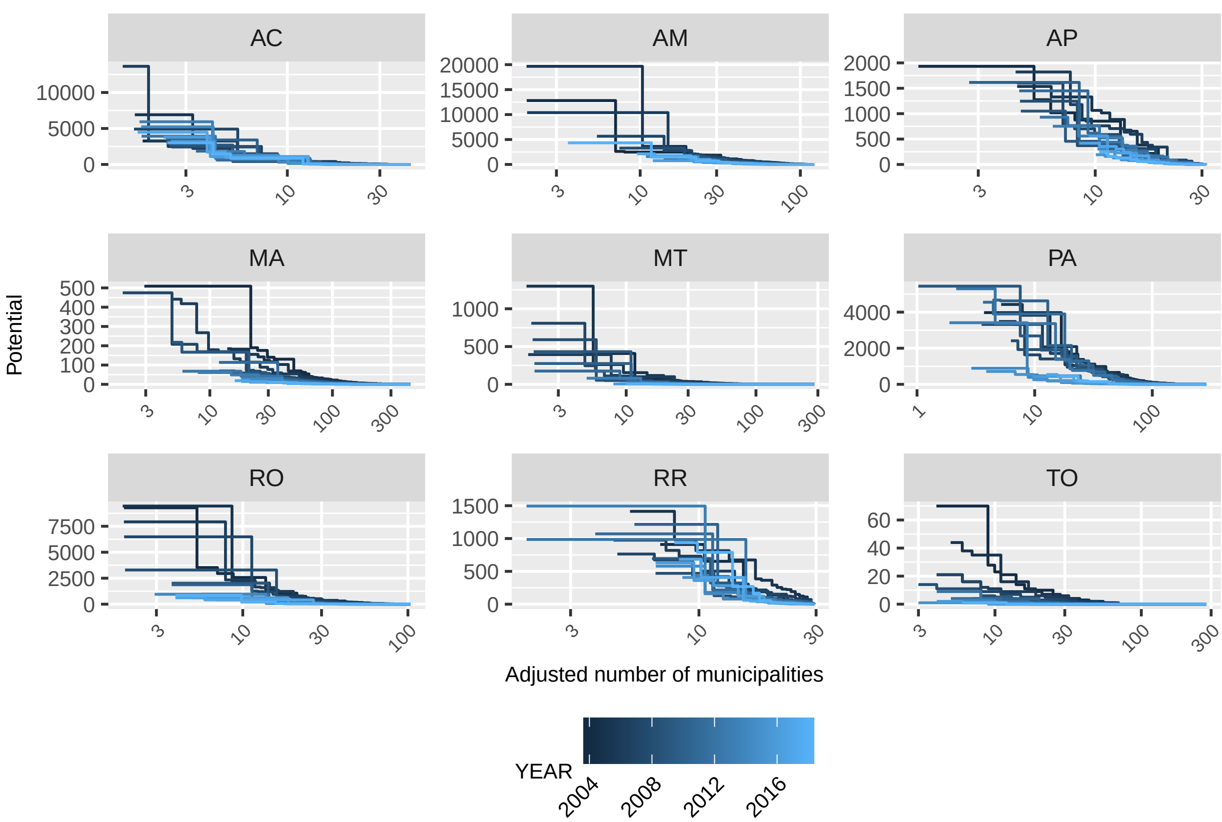
When we inspect the potential for states over time we see that in recent years, numbers stay more or less constant, despite an initial decrease in early years.
We proceed with the cummulative potential plots by normalized potentials and the potential plots.

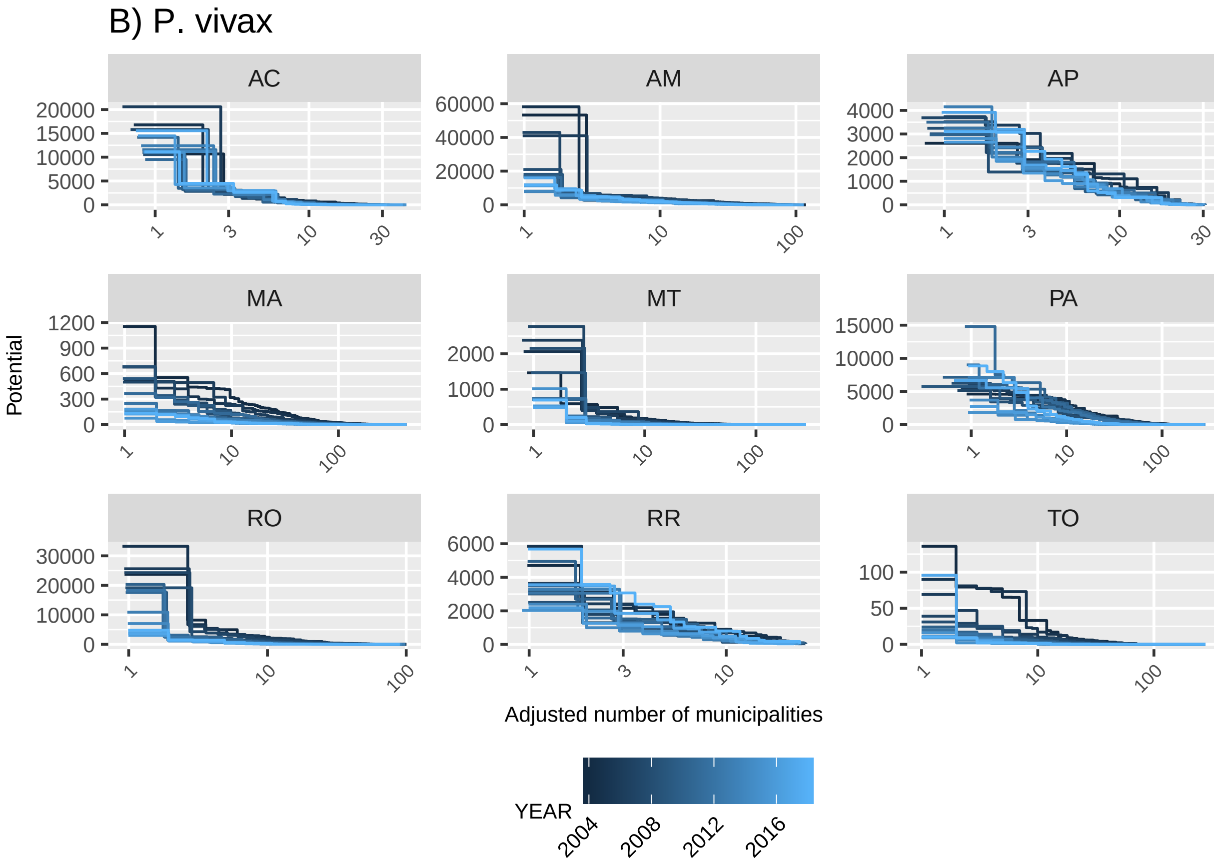


B) P. vivax



A) P. falciparum

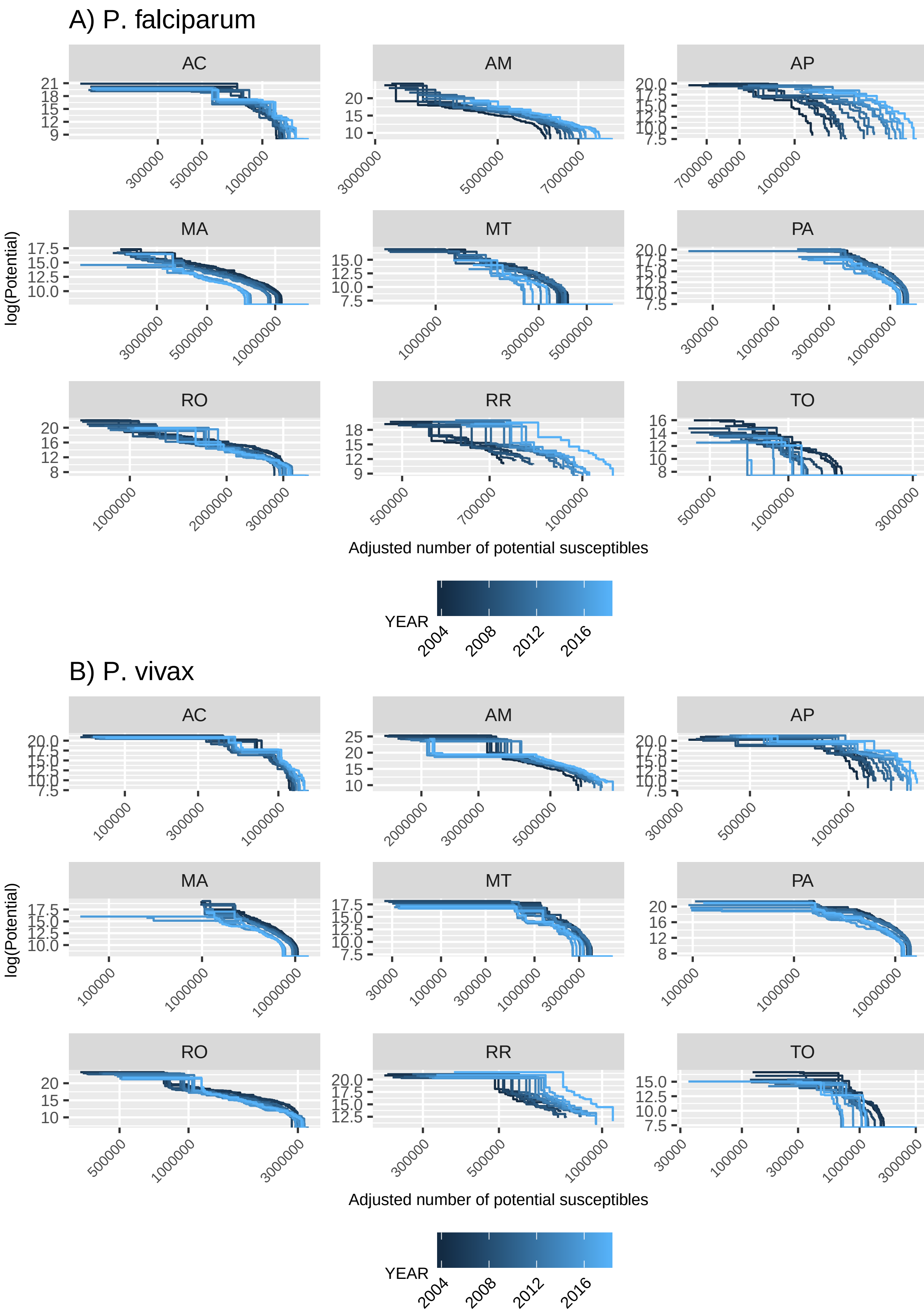




Potential related to the next expected number of cases

Here we change the function to have $P_{i,y} = C_{i,y}(N_{i,y} - C_{i,y})$. Now the term $(N_{i,y} - C_{i,y})$ will be the adjusted number of susceptibles. The potential in this case is related to the expected number of new cases and we apply a logarithm for convenience, $log(P_{i,y})$.

We again order municipalities by descending potentials.


A) P. falciparum
AC
AM
AP
MA
MT
PA
RO
RR
TO
log(Potential)
Adjusted number of potential susceptibles
YEAR
2004
2008
2012
2016
B) P. vivax
AC
AM
AP
MA
MT
PA
RO
RR
TO
log(Potential)
Adjusted number of potential susceptibles
YEAR
2004
2008
2012
2016