Statistical Analysis and Charts

P. Vassiliadis, 2022-04-30

1 Introduction

We performed a statistical analysis for the measured properties of the 195 projects, relating properties like duration or activity to θ-synchronicity, lags and α-attainment. For each project, we used the following measures for statistical analysis: (a) *SchemaUpdatePeriod (SUP), TotalActivity, TablesAtStart, AttrsAtStart, ProjectUpdatePeriod (PUP), PrjCommits, PrjFileUpds, TotalTableDelta*, (b) *5% and 10% synchronicity* (in both number of absolute months, and as a percentage over the PUP), (c) *50%, 75%, 80%, and 100% attainments as percentage of PUP*, and also, (d) the lags of time and source code with respect to the schema as the Boolean flags *TimeLagsSchemaAlways*? and *SrcLagsSchemaAlwaysFlag*? and the measures *AvgTimeLag, AvgSrcLag, TimeLagsSchemaOccurrences, SrcLagsSchemaOccurrences, TimeLagsSchemaPctPUP, SrcLagsSchemaPctPUP*.

We also use *Taxon* as a labeled, nominal class of values per project, as well as the bucketizing of 10% synchronicity (*TenPctSyncClass*), 75%-attainment (*SeventyfiveAttainClass*), Time Lag (*TimeLagPctClass*) and Source Lag (*SrcLagPctClass*) as intervals of 10% range (i.e., [0%-10%), [10% - 20%), . . .]).

2 Tests for Normality

We tested all these attributes for obeying a normal distribution via a Shapiro-Wilk test, and quite expectedly, we found no sign of normality. *The largest p-value we obtained was in the area of 0.007* and the vast majority of the p-values were below 10^{-10}. Software (and schema) evolution is a human activity, and thus normality is not normally expected.

3 Correlations

Kendall's rank test, returning a value called tau-value is the most powerful test for correlations. If the p-value is small and tau quite large (close to 1.0), this means that two attributes are strongly correlated. If the tau value is close to -1.0, then the two attributes are anti-correlated.

3.1 Intra-family correlations

In terms of correlations, the introduced metrics are fairly correlated per family. We performed Kendall tau tests to test correlation within the members of each family.

- Concerning synchronicities, *5% synchronicity is correlated to 10% synchronicity at 0.67.* Being strongly correlated safely allows us to use only 10% synchronicity in further tests.
- *All the attainments are strongly correlated with each other*, except for 50% and 100% attainment that have .50 correlation. *The best representative of the family is 75%-attainment* with Kendall tau values of 0.83 for 50%-, 0.94 for 80%- and 0.60 for 100%-attainment. Therefore, in the rest of our deliberations, *we will focus to 10% synchronicity and 75%-attainment as representatives of their respective families*. We also tested the *Life Percentage of Schema Advance* Over: (i) *Time* and (ii) *Source*: their Kendall value is 0.75 indicating a very strong correlation.

- *SrcLag* and *TimeLag*: their Kendall tau value is 0.75, which means that they are very correlated, too; therefore, either of them could be used as a representative (see next too, though)
- *PrjCommits* and *PrjFileUpds*: their Kendall tau value is 0.70, which means that they are very correlated, too; therefore, either of them could be used as a representative
- *TablesAtStart* and *AttributesAtStart*: their Kendall tau value is 0.80, which means that they are very correlated, too; therefore, either of them could be used as a representative

3.2 Inter-family correlations

We also tested measures of different families for their correlation.

10% synchronicity and *75%- attainment* are correlated with -0.19 Kendal tau value, meaning that they are very slightly anti-correlated and have to be studied independently.

However, *we found a very strong anti-correlation of 75%-attainment with both lags:* specifically, the Kendall tau values for *the correlation of 75%-attainment with Life Percentage of Schema Advance Over Time is -0.73 and -0.75 for its correlation with Life Percentage of Schema Advance Over Source.* Consequently, in the rest of our deliberations, we can use 75% attainment as a fairly good representative of the lag measures. The strong anticorrelation of 75% attainment with both lags means that the later you attain the 75% milestone, the less in advance you are from time or source code evolution. This, a posteriori, sounds reasonable: if a schema is late to achieve the milestone, this means it has been collecting schema evolution progress "regularly", and as such, it is aligned with the progress of time and source evolution. In the exactly inverse reading, *observing a schema evolving regularly over time or source indicates that the probability of sharply amassing schema change and then freezing is small!*

3.3 Measures of interest, discretization and outlier removal

Based on all the above *we can focus our study on 10% synchronicity and 75% attainment*.

10% Synchronicity	#Prjs
00%-10%	12
10%-20%	17
20%-30%	17
30%-40%	30
40%-50%	23
50%-60%	21
60%-70%	21
70%-80%	11
80%-90%	19
90%-100%	8
100%-100%	16
Grand Total	**195**
Avg group size:	17.73

75% attainment	#Prjs
00%-00%	48
00%-10%	22
10%-20%	25
20%-30%	16
30%-40%	10
40%-50%	10
50%-60%	9
60%-70%	13
70%-80%	15
80%-90%	14
90%-100%	7
100%-100%	6
Grand Total	**195**
Avg group size:	16.25

To facilitate subsequent analyses, we have quantized the two measures in intervals of 10% range; we also report the size of each such class.

We have performed all our tests on both (a) the full data set and (b) the data set where outliers have been removed. To remove outliers, we have z-scored all the measures of interest, specifically, *TotalActivity, SUP, TablesAtStart, PrjFileUpds, PUP, TotalTableDelta* (remember that z-score measures distance from the mean and has standard deviations as units). Each time one of these measures was studied, we removed the projects outside the range of [-2 … 2] of z-scores for the respective attribute (thus, the removal takes out projects more than 2 standard deviations away from the mean). Naturally, Taxon cannot be part of z-scoring and has no outliers.

Attribute	***#outliers***
Taxon	No outliers
TotalActivity	3 outliers
SUP	9 outliers
TablesAtStart	6 outliers
PrjFileUpds	7 outliers
PUP	11 outliers
TotalTableDelta	2 outliers

We found no significant differences in our tests, due to the fact that (i) the number of outliers is small and (ii) our statistical tests are non-parametric, and do not measure the mean. However, quite expectedly, removing outliers removes variability from the data set, with values becoming more homogeneous, and thus the differences of different groups diminish. This results in a small increase of the p-value of most tests; however, no remarkable difference takes place effectively.

p-value of Kruskal-Wallis over …

	10% Sync Class		*75% Attain. Class*	
	full dataset	*outliers removed*	*full dataset*	*outliers removed*
Taxon	0.00282		0.00645	
TotalActivity	0.00087	0.00258	0.00482	0.00555
TotalTableDelta	0.00586	0.00841	0.03579	0.03193
SUP	0.00101	0.00393	0.07026	0.13350
TablesAtStart	0.00725	0.06441	0.30380	0.26460
PrjFileUpds	0.00481	0.01293	1.98E-07	8.41E-07
PUP	0.00057	0.00156	0.00188	0.00644

We elaborate on the production and significance of the abovementioned results in the next sections.

4 The role of Taxon over Synchronicity, Lags and Attainment

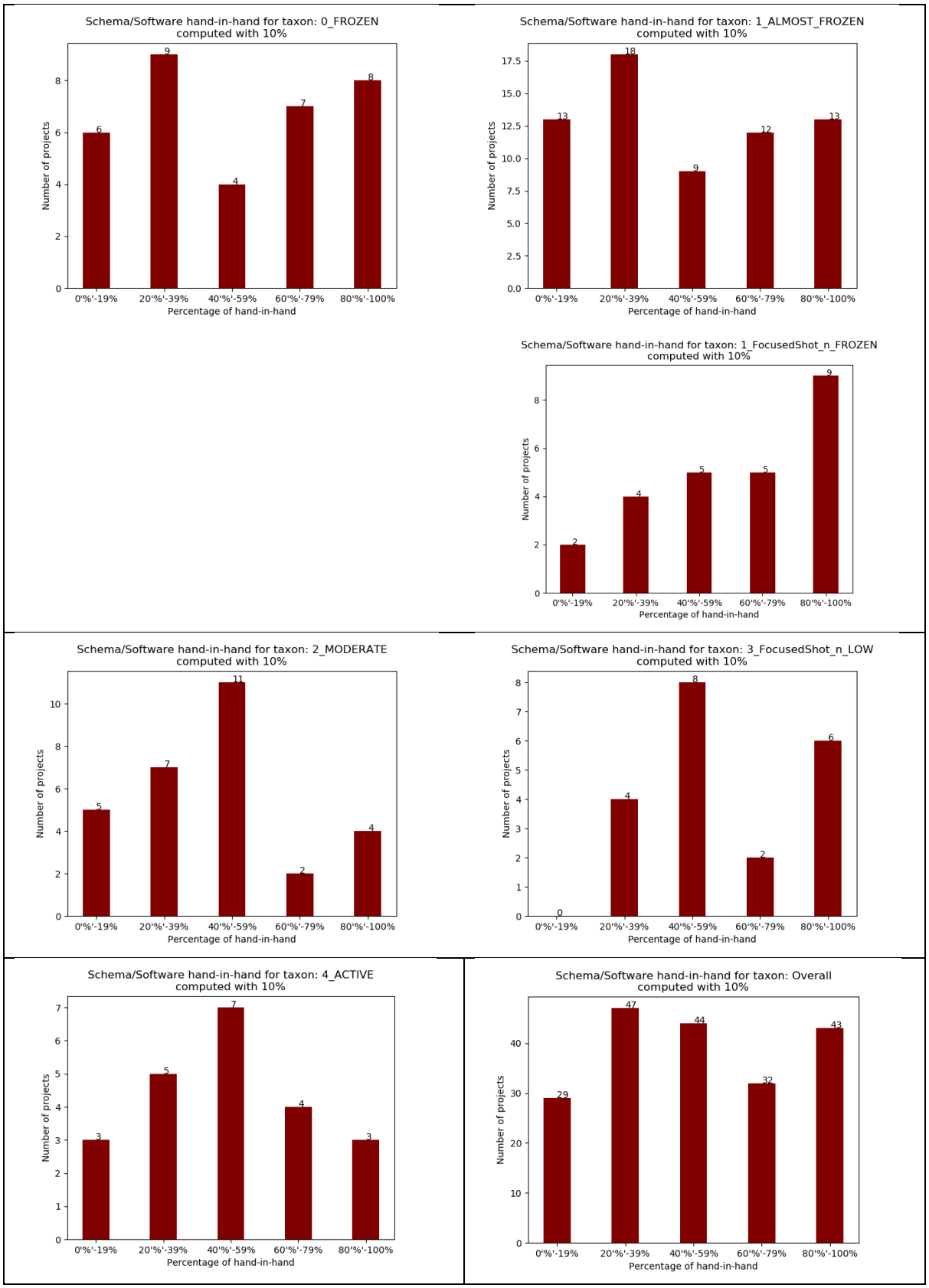


Figure 1 Breakdown of projects per value range for the 10%-synchronous co-evolution

4.1 Testing synchronicity

The bottom-right histogram depicting the entire breakdown overall all the 195 projects, is signifying quite emphatically that all kinds of behavior towards the co-evolution are present. *Different taxa have different distributions, however*. For the taxa where the schema is (almost) frozen, we can say that there is a polarity taking place: either the project evolves while the schema is not (low values of the 10%-synchronicity), or both of them are pretty much frozen (higher values). For the Moderate and Active taxa, where the schema evolves on a more regular basis (albeit at different volumes of change), there is almost a bell-curve formed. For the rather few projects of the two "focused-shot" taxa, there is a bias towards the right-hand side of the histogram, meaning that in several cases, there is a significant percentage of co-evolution, which in term implies that these projects are inclined towards a more "focused-shot" evolution for the surrounding code too, and not only for their schema.

The violin plot demonstrates the situation in a single chart.

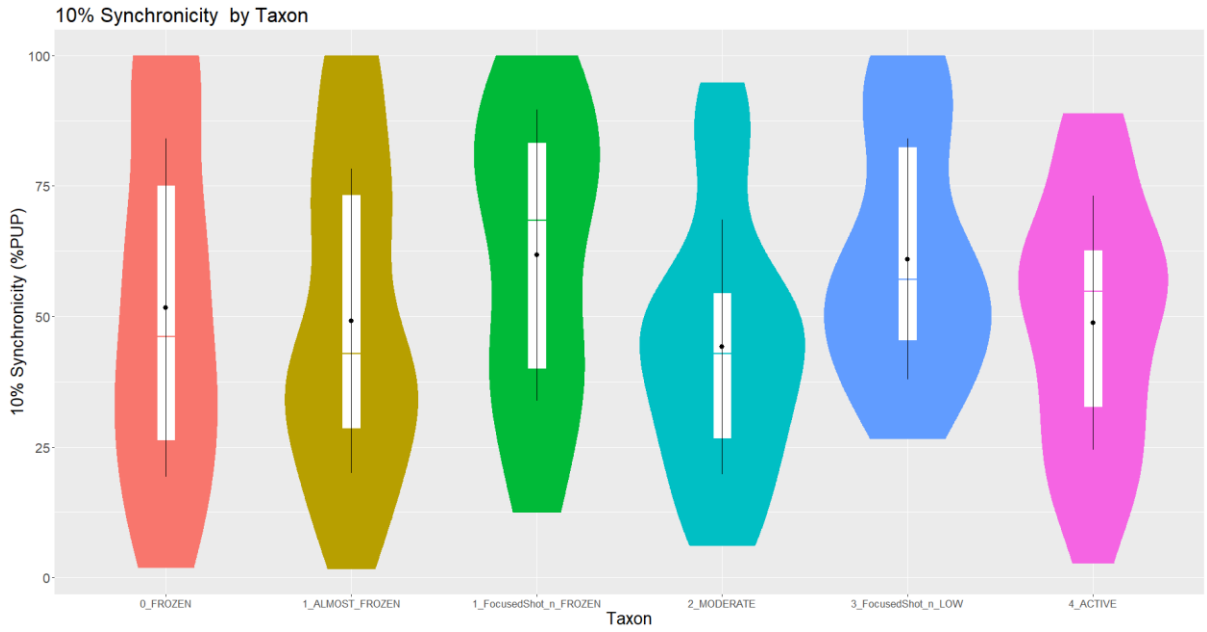


Figure 2 Violin charts of 10% Synchronicity per Taxon

Reading violin plots: the shape of the violin shows the concentration of the population of each group (larger concentrations produce "fatter" surface). The rectangle in the middle shows the IQR. The horizonal line in the rectangle is the median value. The dot is the mean value. The vertical line is the range of one standard deviation below and above the mean.

4.2 Testing Attainment

Attainment is a very different story than synchronicity. Almost half the projects reach their 80% attainment (i.e., 80% of the changes performed to their schema) in the first 20% of their life. So, when we study the behavior of the different taxa, the main problem is whether they behave similarly with respect to the treatment of the [0% - 20%] interval or not.

Practically, the hotter the taxon is, the later the attainment of cumulative change happens. We will use 75% attainment as the main measure in our deliberations. The Active taxon has a bell-curve concerning the attainment of 75% of cumulative change, meaning that for most projects, it happens around mid-life (both mean and median are close to 0.5). On the other hand, the three frozen taxa

are strongly biased towards completing their 75% cumulative change below 20% of project life. In the middle, *Moderate* and *FocusedShotAndLow* have slightly higher medians, around 0.3 of PUP. Special note to the two "FocusedShot" taxa: both have a long IQR because of a bipolar bias towards both low and high values of 75% attainment. This is explained by the existence of both long and short projects (where the focused shot nature of the schema evolution also applies to the entire project too).

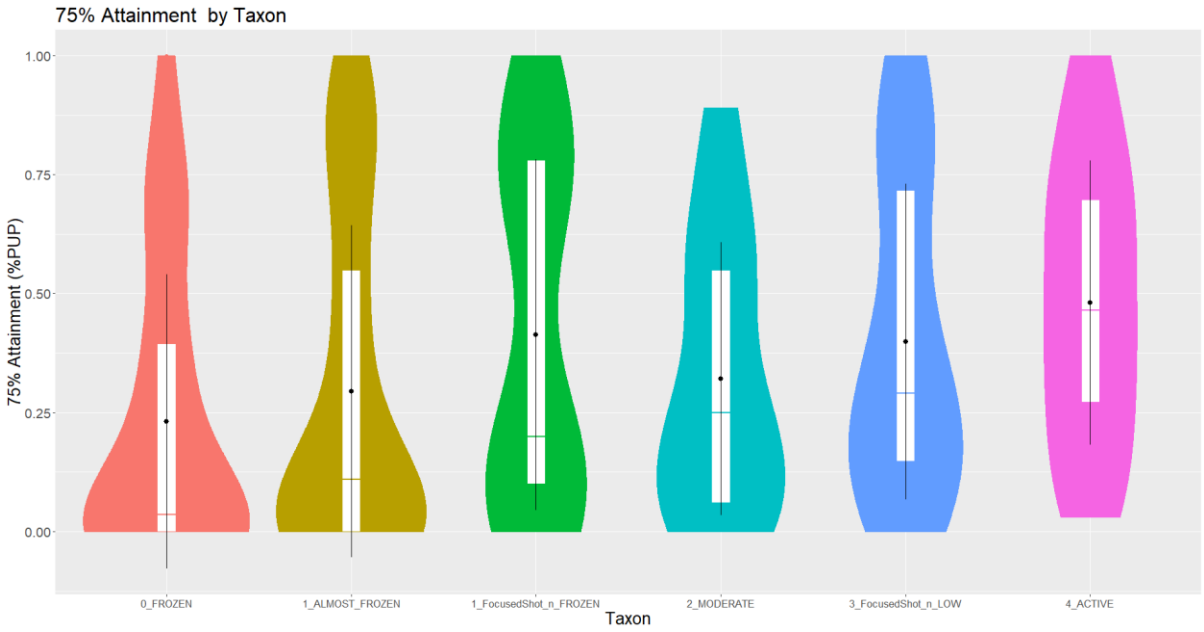


Figure 3 Violin charts of 75% Attainment per Taxon

4.3 Statistically testing synchronicity and attainment

We performed a Kruskal-Wallis test of the 10% Synchronicity class (i.e., the discretized values of 10% synchronicity in 10% intervals) against Taxon, with an α level of 0.05. *The p-value produced was a 0.00282*, i.e., one order of magnitude lower than the alpha level, verifying the fact that a difference is evident.

We tested the Taxon against 75% attainment again via a Kruskal-Wallis test. *The p-value was 0.00645*, again an order of magnitude lower than the α level of 0.05.

4.4 Testing Lag

We also performed statistical tests concerning whether the different taxa behave differently concerning the case of time, source or both being found in lag with respect to the schema evolution for the entire lifetime of a project. We performed a Chi-square as well as a two-sided Fisher test for all the three categories depicted in the following Figure, assuming an α level of 0.05.

Time lags schema always	FROZEN (34 prjs)	ALMOST FROZEN (65 prjs)	FocusedShotn FROZEN (25 prjs)	MODERATE (29 prjs)	FocusedShotn LOW (20 prjs)	ACTIVE (22 prjs)	*Grand Total (195 prjs)*
FALSE	50.00%	47.69%	64.00%	62.07%	80.00%	68.18%	*57.95%*
TRUE	50.00%	50.77%	36.00%	37.93%	15.00%	31.82%	***41.03%***
(blank)	0.00%	1.54%	0.00%	0.00%	5.00%	0.00%	*1.03%*
Grand Total	***100.00%***	***100.00%***	***100.00%***	***100.00%***	***100.00%***	***100.00%***	***100.00%***

Source lags schema always	FROZEN (34 prjs)	ALMOST FROZEN (65 prjs)	FocusedShotn FROZEN (25 prjs)	MODERATE (29 prjs)	FocusedShotn LOW (20 prjs)	ACTIVE (22 prjs)	*Grand Total (195 prjs)*
FALSE	52.94%	64.62%	72.00%	72.41%	90.00%	86.36%	*69.74%*
TRUE	47.06%	33.85%	28.00%	27.59%	5.00%	13.64%	***29.23%***
(blank)	0.00%	1.54%	0.00%	0.00%	5.00%	0.00%	*1.03%*
Grand Total	***100.00%***	***100.00%***	***100.00%***	***100.00%***	***100.00%***	***100.00%***	***100.00%***

Both time and source lag schema always	FROZEN (34 prjs)	ALMOST FROZEN (65 prjs)	FocusedShotn FROZEN (25 prjs)	MODERATE (29 prjs)	FocusedShotn LOW (20 prjs)	ACTIVE (22 prjs)	*Grand Total (195 prjs)*
FALSE	52.94%	66.15%	76.00%	72.41%	90.00%	86.36%	*70.77%*
TRUE	47.06%	32.31%	24.00%	27.59%	5.00%	13.64%	***28.21%***
(blank)	0.00%	1.54%	0.00%	0.00%	5.00%	0.00%	*1.03%*
Grand Total	***100.00%***	***100.00%***	***100.00%***	***100.00%***	***100.00%***	***100.00%***	***100.00%***

Figure 4 Breakdown of Lags per Taxon

The data used have the form:

columns: the 6 taxa

rows: one row for FALSE (xxx does not ALWAYS lag schema evo), and, one row for TRUE (xxx ALWAYS lags schema evo), with xxx in {*TimeLagAlways, SourceLagAlways, BothLagAlways*}

cells: the number of projects that pertain to the respective [row,col]

Practically, we are testing whether Taxon and *LagsAlways* have a relationship. We omit the 2 projects with extremely short period that could not be accounted for in the "always" measurements.

The hypotheses for both the Chi-square test and the Fisher's exact test are the same, i.e.:

H0: the column and row variables are independent, there is no relationship between them.

H1: the variables are dependent, there is a relationship between the two categorical variables.

Concerning the *Time Lag*, both the Chi-square and the Fisher test produce a p-value of 0.07. Concerning both the *SourceLagAlways* and the *BothLagAlways*, the Chi-square tests give a p-value of 0.02 and the Fisher test a p-value of 0.01, i.e., both lower than the significance level of 0.05. Since the p-value is less than the significance level, we can reject the null hypothesis, i.e., it is possible that there is a significant relationship between the two categorical variables (*Taxon* and *LagsAlways*). We should note that all the p-values are still in the area of 1% which is not extremely small. However, the combination of the data (producing a trend) and the p-values allows us to state that *concerning the issue of whether the schema evolves in advance of time, source code or both, the differences between the taxa are clearly visible and in the two last categories statistically significant too*.

5 Testing all other attributes

5.1 Synchronicity and Attainment over Total Schema Activity

Total Schema Activity measures the amount of change that occurred to the schema in terms of attributes. Specifically, we count attributes born with new tables, injected to existing ones, deleted with deleted tables, ejected from tables with the table surviving, as well as attributes going through a data type change or a change in the primary key of the table.

3 outliers were removed (only). The statistical tests of both 10% synchronicity and 75% attainment show p-values close to the ones before the removal of the outliers. *Both p-values are lower than the alpha level of 0.05*, with (a) 10% synchronicity one or two orders of magnitude lower (0.0009 for all the dataset and 0.003 when outliers are removed), and, (b) 75% attainment around 0.005 in both cases, i.e., an order of magnitude lower than the alpha level.

Concerning **synchronicity**: both small values of synchronicity and very high values of synchronicity are clustered around small volumes of activity. The range of 40% - 70% synchronicity shows a higher IQR, median and mean of activity than the other ranges.

At the same time, high volumes of schema activity are present in all the range 0% - 70% of synchronicity. This means that high levels of synchronicity are not accompanied by high schema change, which is somehow expected, given the existence of projects that do not change much either the source or the schema (i.e., high synchronicity and low activity, as shown in the right-hand side of the violin plot).

On the other hand, actively evolving schemata seem to be in a 40% - 70% synchronicity rate (esp., if they are long-lasting, which means that after the "5-year threshold" they start getting out of sync).

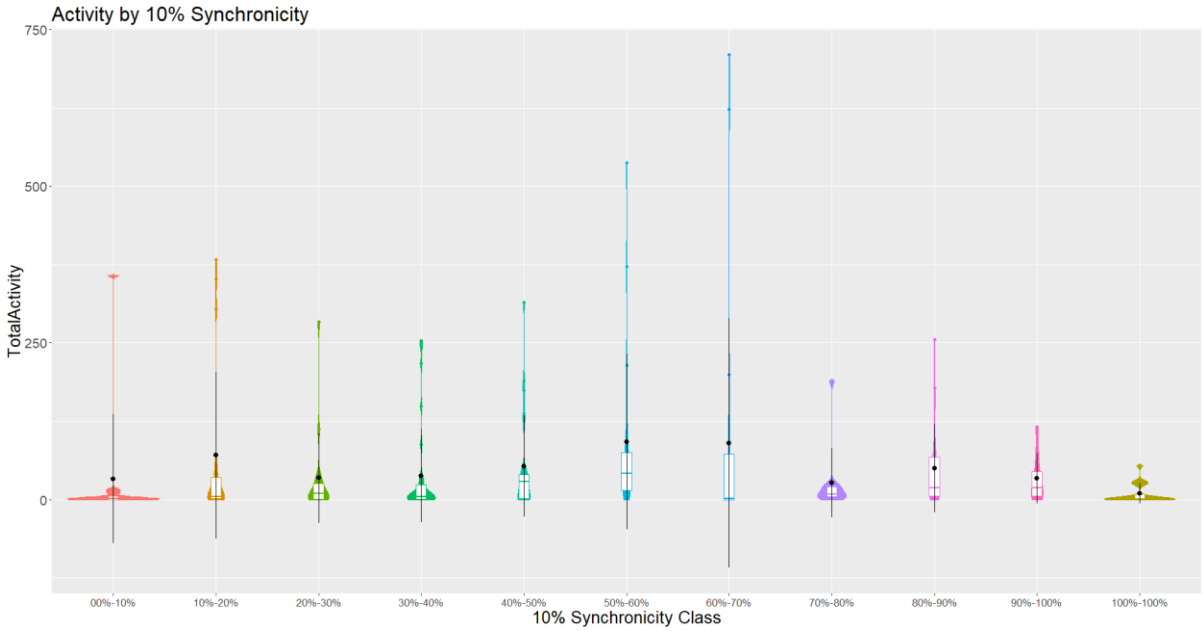


Figure 5 Activity over 10% synchronicity

Concerning **attainment**: there are projects with a high-volume activity in several ranges. As already mentioned, the vast majority of projects attain 75% of change very early, esp. at the originating version. Of particular interest is the range 60%-80% (the reasonable place to see a 75% attainment)

which has the longest IQR's and means as well as high volume changes. Practically: since both Activity and 75% attainment refer to schema (attn: schema, not project) changes, the high-level activities are probably produced with a close-to-linear pace. At the same time, see also some similar activity quickly gathered in the 10%-20% interval.

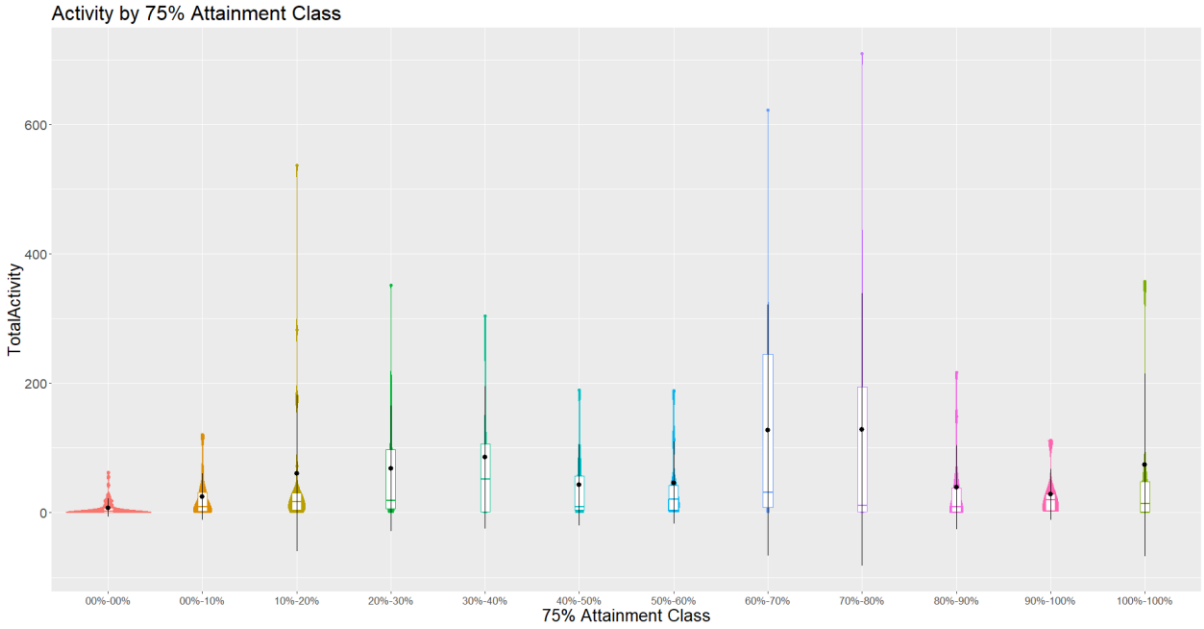


Figure 6 Activity over 75% attainment

5.2 Synchronicity and Attainment over TableDelta

TableDelta refers to the total number of tables that were created or removed during the life of the schema. It is a coarse approximation of activity, measured in tables rather than attributes.

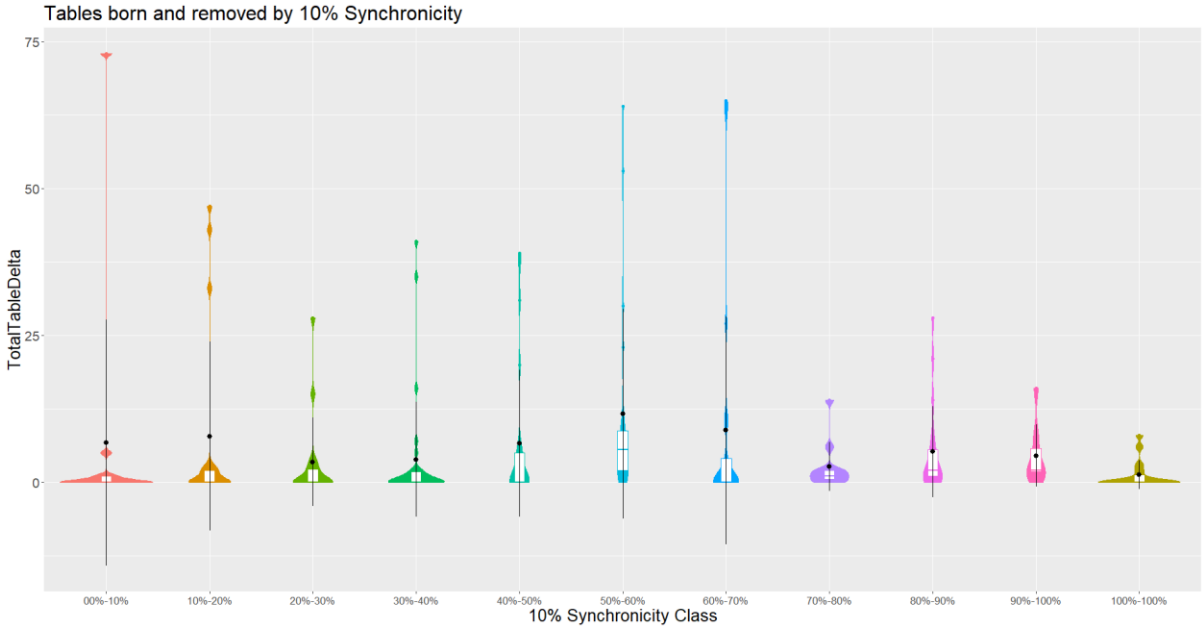


Figure 7 Table Delta over 10% synchronicity

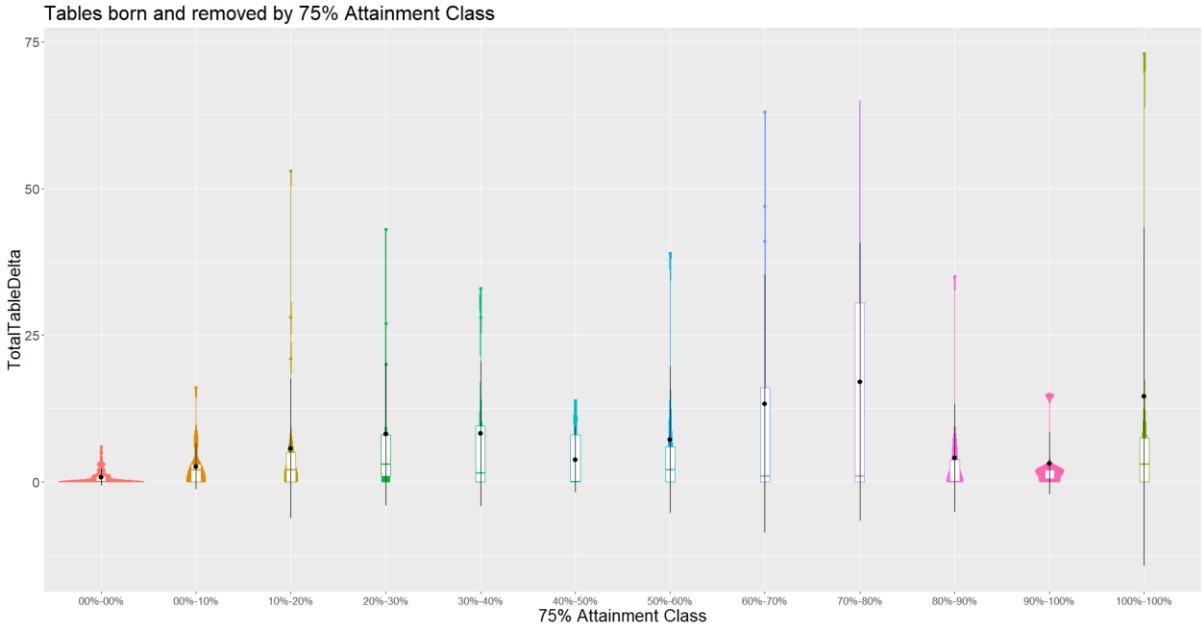


Figure 8 Table Delta over 75% attainment

2 outliers were removed (only). The statistical tests of both 10% synchronicity and 75% attainment show p-values very close to the ones before the removal of the 2 outliers (which is expected of course, just 2 outliers). *Both p-values are lower than the alpha level of 0.05*, with (a) 10% synchronicity one

order of magnitude lower (0.006 for all the dataset and 0.008 when outliers are removed), and (b) 75% attainment being fairly close to the alpha level (p-value is 0.03).

Concerning **synchronicity**: there is a small difference of median and IQR for the interval 50%-60% -- otherwise all other ranges are very similar (quite expected due to the predominance of projects having a value of zero table delta's in their schema size).

Concerning **attainment**: the groups have similar medians, but the IQR's seem as a combination of two bell curves having an exceptional peak at the 30%-40%, and, 60%-80%. The left-hand-side population is mostly having very small deltas (see the IRQs); the center-based population has long IQR's meaning that it hosts all sorts of projects with respect to their table deltas. There are only 19 projects exceeding 20 tables difference (21 with the outliers) which are spread fairly uniformly throughout the entire range of attainments.

5.3 Synchronicity and Attainment over Schema Update Period

Schema Update Period is the period between the birth of the schema and the last update ever performed to it, and is measured in months.

9 outliers were removed. The statistical tests of 10% synchronicity and 75% attainment differ. *10% synchronicity* shows p-values in the area of 0.001 (full data set) and 0.003 (outliers removed), *lower than the alpha level of 0.05*. On the other hand, *75% attainment exceeds the alpha level* (0.07 for all the dataset and 0.13 when outliers are removed).

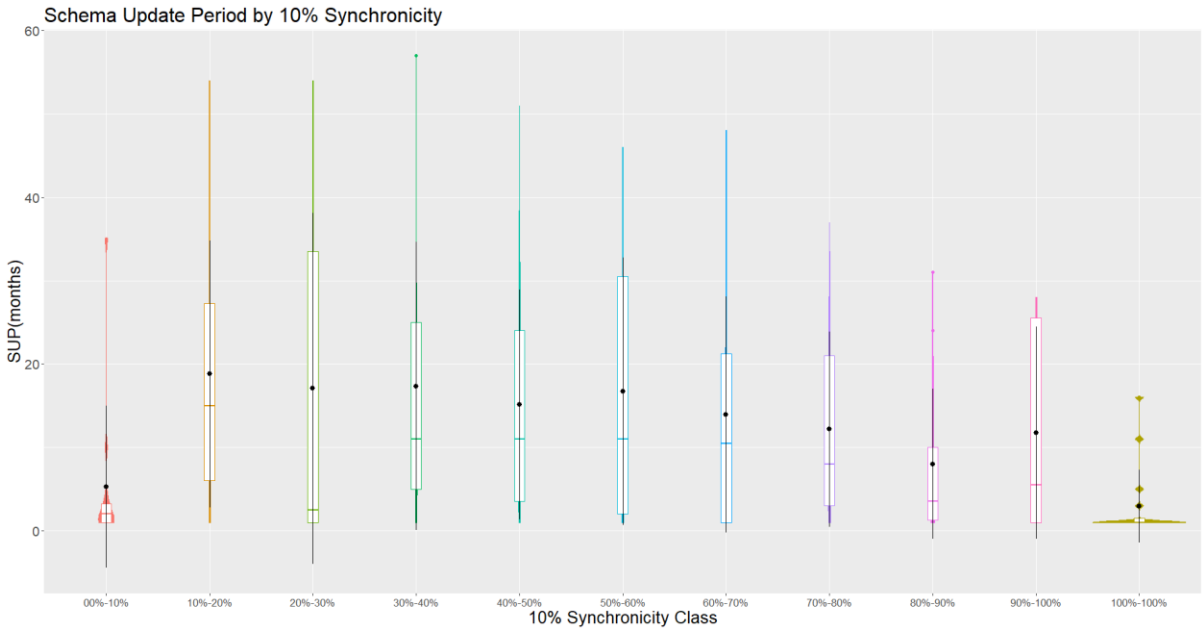


Figure 9 SUP over 10% Synchronicity

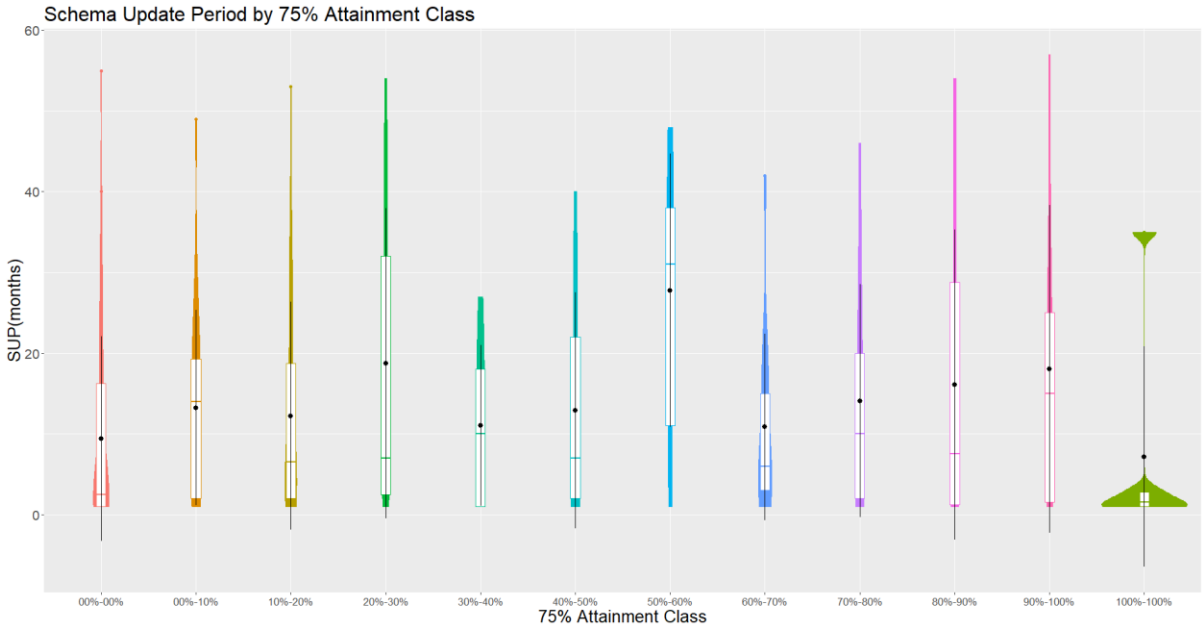


Figure 10 SUP over 75% attainment

Concerning **Synchronicity**: with the exception of the 0%-10% and the 100% intervals, the rest of the boxplots seem very close. Mean values and medians are very close, and high-level schema update periods are found in all the range between 10% - 70%. No major correlation of SUP with synchronicity

seems to be present, other than a tendency of high values of synchronicity (80%-90% and 100%) being correlated with short update periods; this also holds for projects with short SUP and high project activity, having a synchronicity of 0%-10%.

Concerning **Attainment**: with the exception of the 50%-60% and the 100% intervals (which are very small groups with 9 and 6 members, respectively), the rest of the boxplots seem very close.

Lesson learned: differences in synchronicity and attainment are NOT due to SUP!

5.4 Synchronicity and Attainment over Schema Size at Start

Schema Size at Start (also: *Tables at Start*) is the number of tables with which the schema is born. We are interested, of course, on whether this information can have any relationship to how the schema evolves later.

6 outliers were removed. The statistical tests of both 10% synchronicity and 75% attainment differ. *10% synchronicity* shows p-values in the area of 0.007 (full data set) *lower than the alpha level of 0.05*, and, 0.3 (outliers removed), *higher than the alpha level of 0.05*. Here the effect of outlier removal is significant (the only case where this happens) and has to do with the very high values attained by some of the removed schema sizes. On the other hand, *75% attainment exceeds the alpha level* in both cases (0.3 for all the dataset and 0.26 when outliers are removed). Overall, *we cannot say that the differences we observe are statistically significant.*

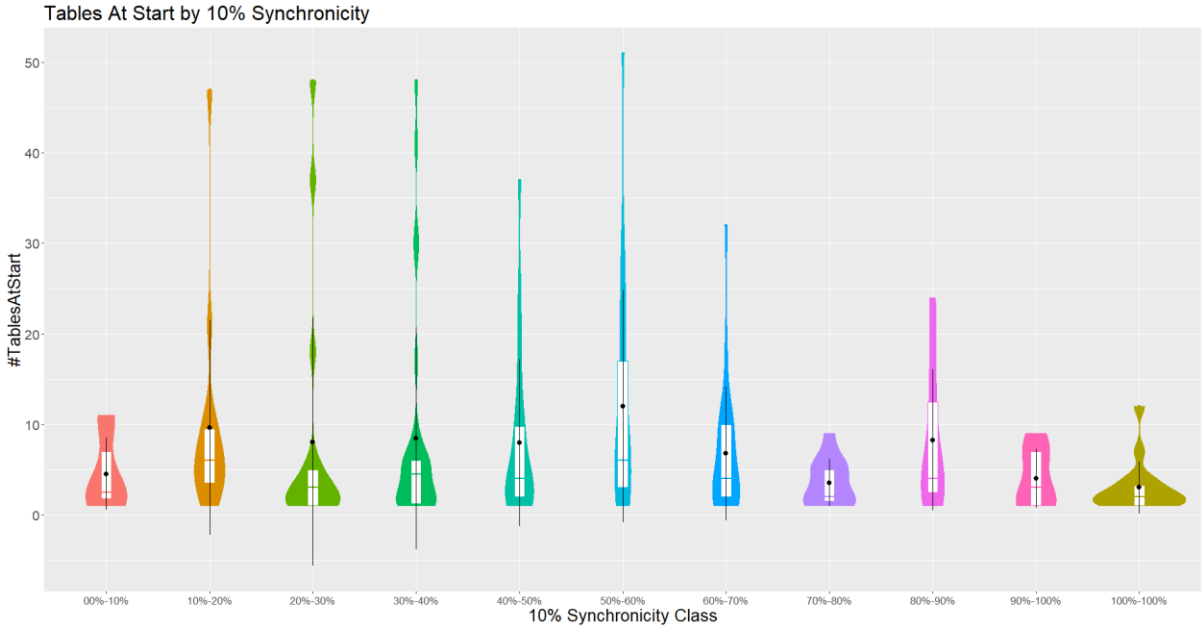


Figure 11Tables at Start over 10% Synchronicity

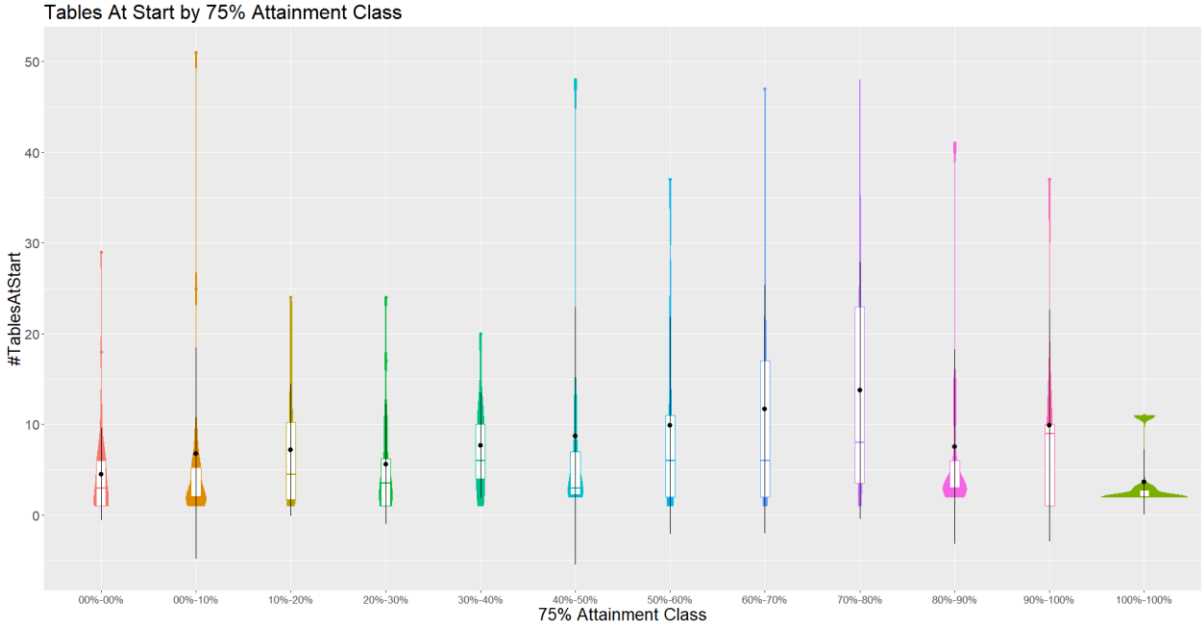


Figure 12 Tables at Start over 75% Attainment

Overall, there is not much to say about how schemata behave with respect to their **synchronicity** given their originating schema size. To a large extent this is due to the predominance of small schemata throughout the data set. Large schema sizes are found anywhere between 10% and 60% synchronicity, practically saying that the high-levels of synchronicity typically concern small schemata.

In a rather inverse picture concerning **attainment**, large originating schemata are found in the area 40% - 90% of PUP for 75% attainment, which is, again, a too broad area to allow us to infer much. At the same time, we can observe that medians, mean and IQR values are higher in the area of 60% - 80% of PUP (again: where 75% attainment would normally be expected to be).

Combined with the same respective Figures for the *Activity* and *TableDelta* attributes (showing a similar behavior for early attainment values) we can say that early attainment comes with smaller originating schemata and smaller activity too, practically producing a "build-small-n-freeze-soon" effect. On the other hand, middle-level attainment can include larger initiating schemata and larger amounts of activity.

An overall impression is that schemata that are more alive in all respects come at the area of 50%-80% 75% attainment, which is quite reasonable. The inverse does not hold however: these ranges also have small and low-activity schemata too.

5.5 Synchronicity and Attainment over Project File Update Activity

Project File Update activity is the total number of files being updated in the lifetime of a project.

7 outliers were removed. The statistical tests of both 10% synchronicity and 75% attainment show *p-values lower than the alpha level of 0.05*, with (a) 10% synchronicity one order of magnitude lower (0.005 for all the dataset and 0.01 when outliers are removed), and (b) 75% attainment being extremely far from the alpha level (p-value is in the area of 10E-7).

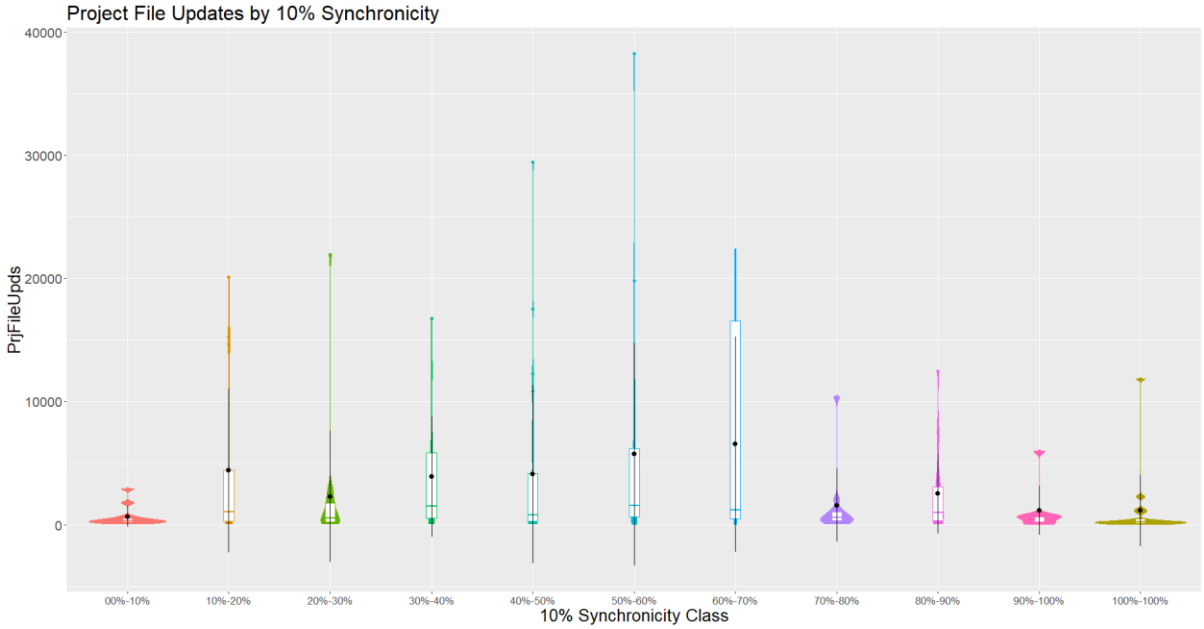


Figure 13 Project File Updates over 10% Synchronicity

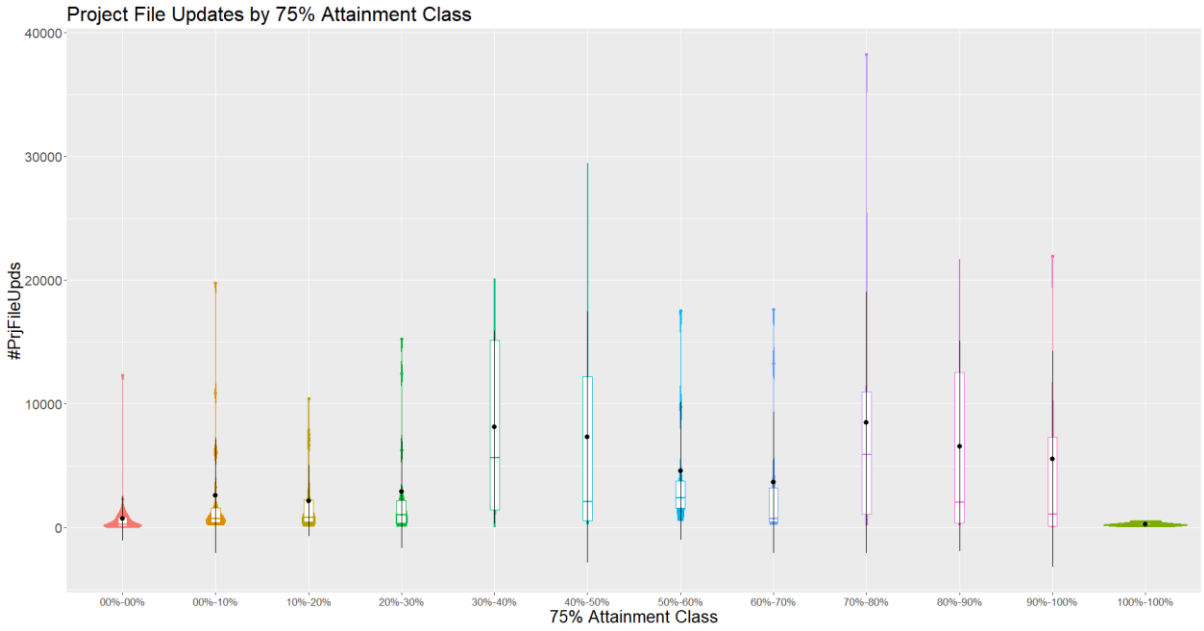


Figure 14 Project File Updates over 75% Attainment

Concerning **synchronicity:** all medians are close, with mid-range attainment periods (between 30% and 70%) having medians that are slightly higher, albeit without a trend in either the medians or the IQRs. It is evident that the projects with very high number of updates are in the area of 40% - 70% of

10% synchronicity. However, the remaining low and the high values of the x-axis of the violin plot, show significantly smaller values of file updates, in terms of all their characteristics.

Concerning **attainment**: all medians are close, with the exception of 30%-40% and 70%-80%. The IQRs also have different ranges. 30%-50% as well as 70%-100% (100% excluded) have large IQR's and mean values of project update. These are also the areas where the projects with the very high file update activity are found. The most important finding of the violin plot is that the four largest groups of attainment, i.e., the 4 groups in the area 0% - 30%, concern projects with smaller volume of file update and earlier attainment – i.e., projects which have their schema evolution completed early appear to not change a lot either with respect to their source code.

5.6 Synchronicity and Attainment over Project Update Period

The Project Update Period is the time period between the first and the last commit of a project's history that we observed, and we measure it in months.

11 outliers were removed. The statistical tests of both 10% synchronicity and 75% attainment show *p-values at least an order of magnitude lower than the alpha level of 0.05*, with (a) 10% synchronicity having 0.0006 for all the dataset and 0.001 when outliers are removed, and, (b) 75% attainment having 0.001 for all the dataset and 0.006 when outliers are removed. Overall, the p-values are reasonably small, although, as we shall see, this is due mostly to "local" phenomena.

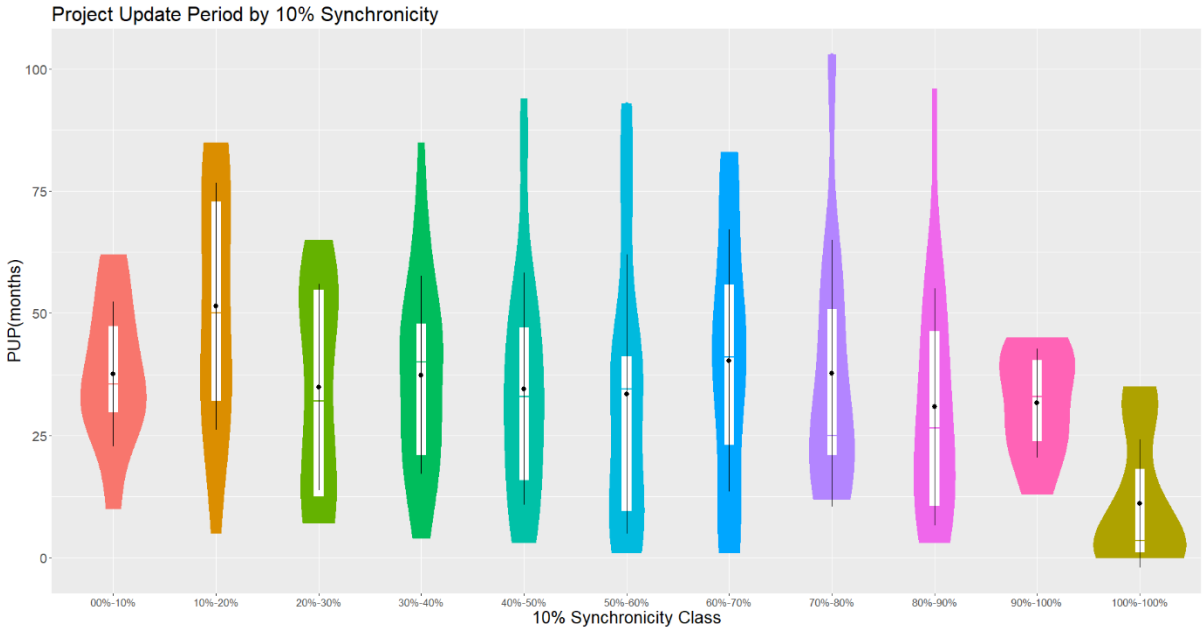


Figure 15 PUP over 10% Synchronicity

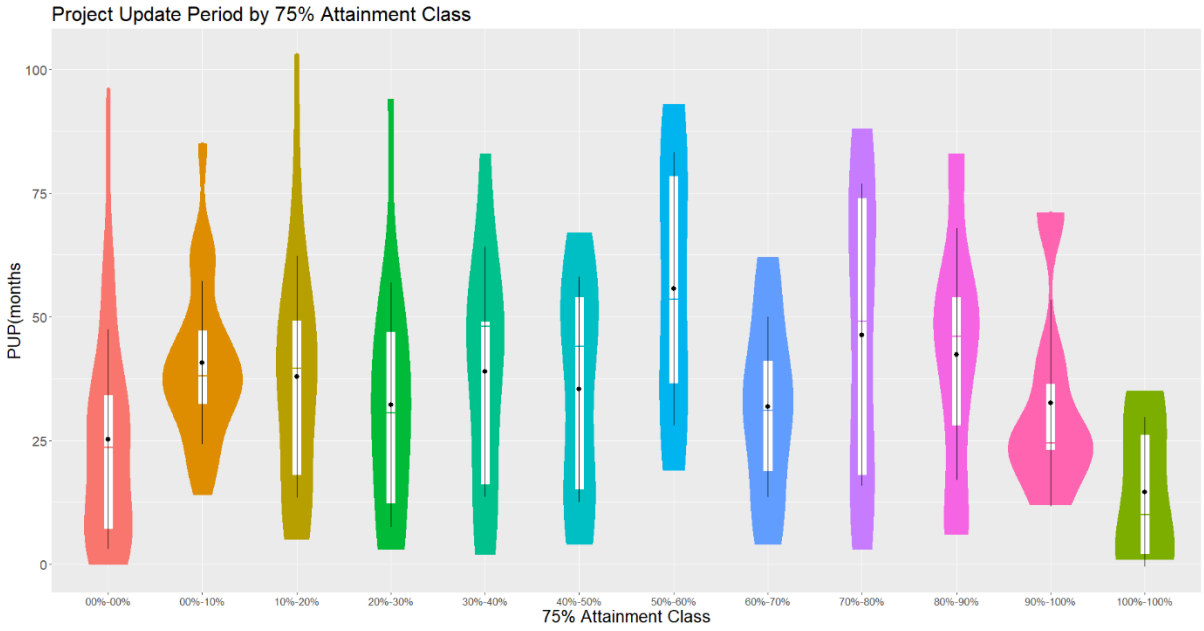


Figure 16 PUP over 75% Attainment

Concerning **synchronicity**: all medians and IQR's are close, with the exception of the range of 10%-20% which has a higher PUP range (expectedly, high duration and synchronicity are inversely correlated) and 100% with a lower PUP (total synchronicity is achieved in rather short project durations). 8 of the 11 outliers were in the range 10%-40% too. Overall, to the extent that projects are rather uniformly spread over the synchronicity intervals, there is no real diversification of PUP with respect to synchronicity, despite the low p-value.

Concerning **attainment**: the medians and the mean values of the different violins are not far from one another (in the area of 30 – 50 months), with the exceptions of the two edges of the x-axis, esp., the right-hand side. Remember that the boxplot does not show that half projects attain 75% of schema evo in the first 20% of their life: the largest group, 0-10% of PUP for 75% attainment, comprising 48 out of 195 projects, is fairly lower than the rest of the groups in terms of PUP. The very small group of 100% is also quite low. Every other group is pretty much close, with the exception of the small group of 50%-60% and 70%-80% having higher median and mean values than the rest as well as IQRs reaching high values (remember that this is 75% attainment, which means that there is a slight precedence of schema evolution with respect to time, and this appears to be the case for the lengthiest projects). In other words, the lengthiest projects are mostly having a reasonably expected 75% attainment of schema activity in the area of 50%-80%.

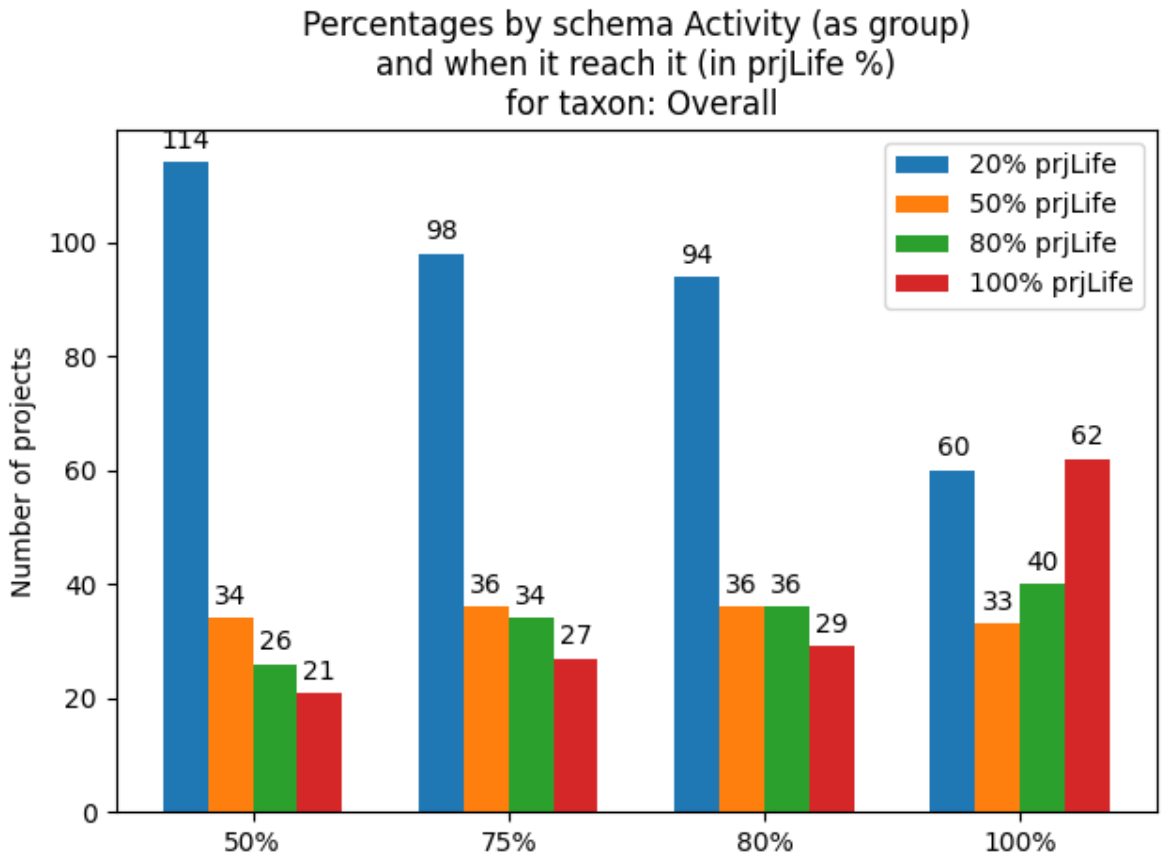
Overall, PUP and synchronicity are not strongly related, whereas high PUP durations in the area of 50%-80% for 75% attainment.

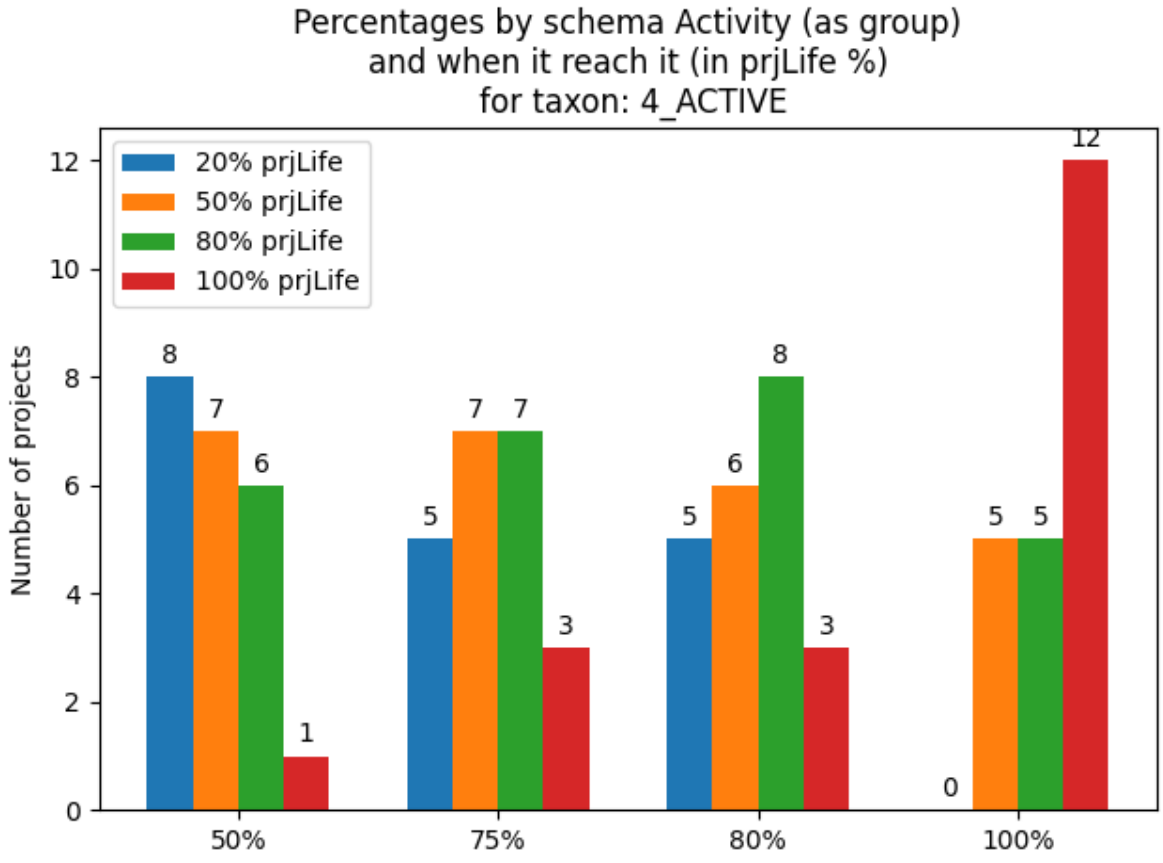
ACCOMPANYING CHARTS

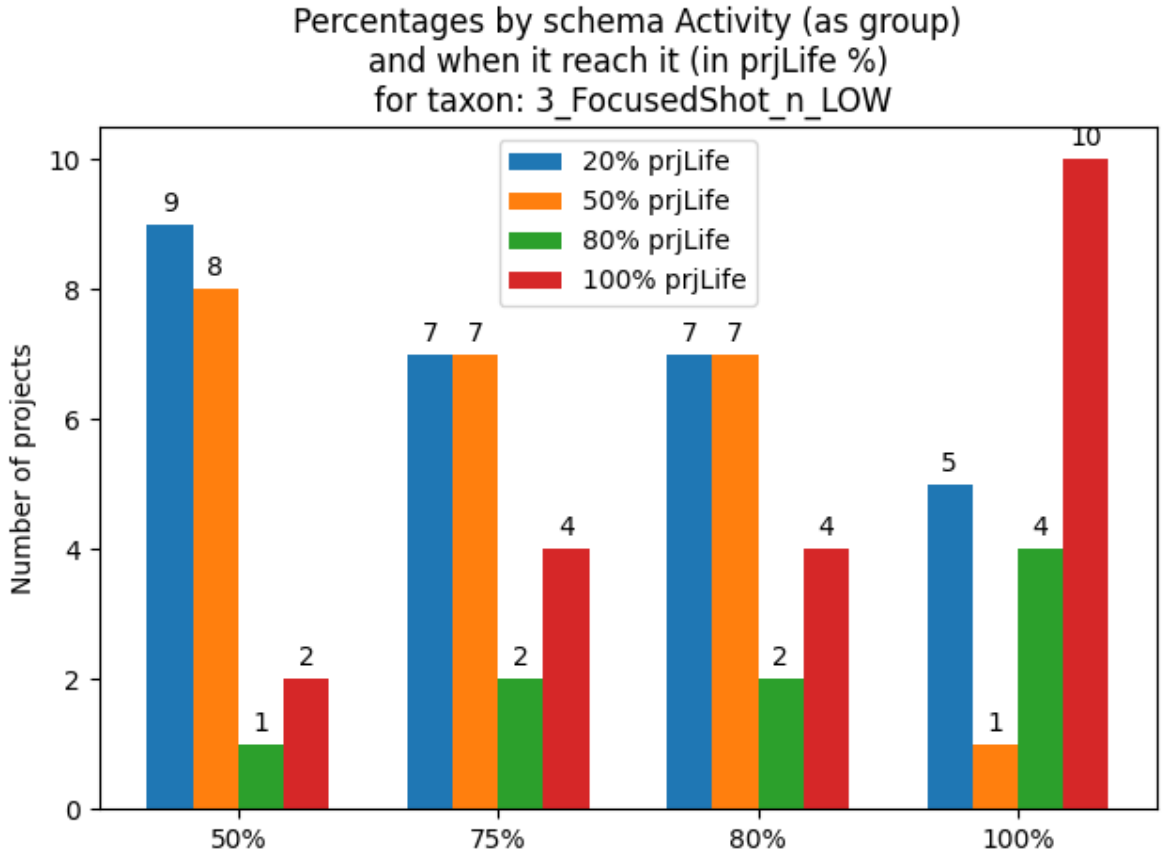
6 Attainment Charts

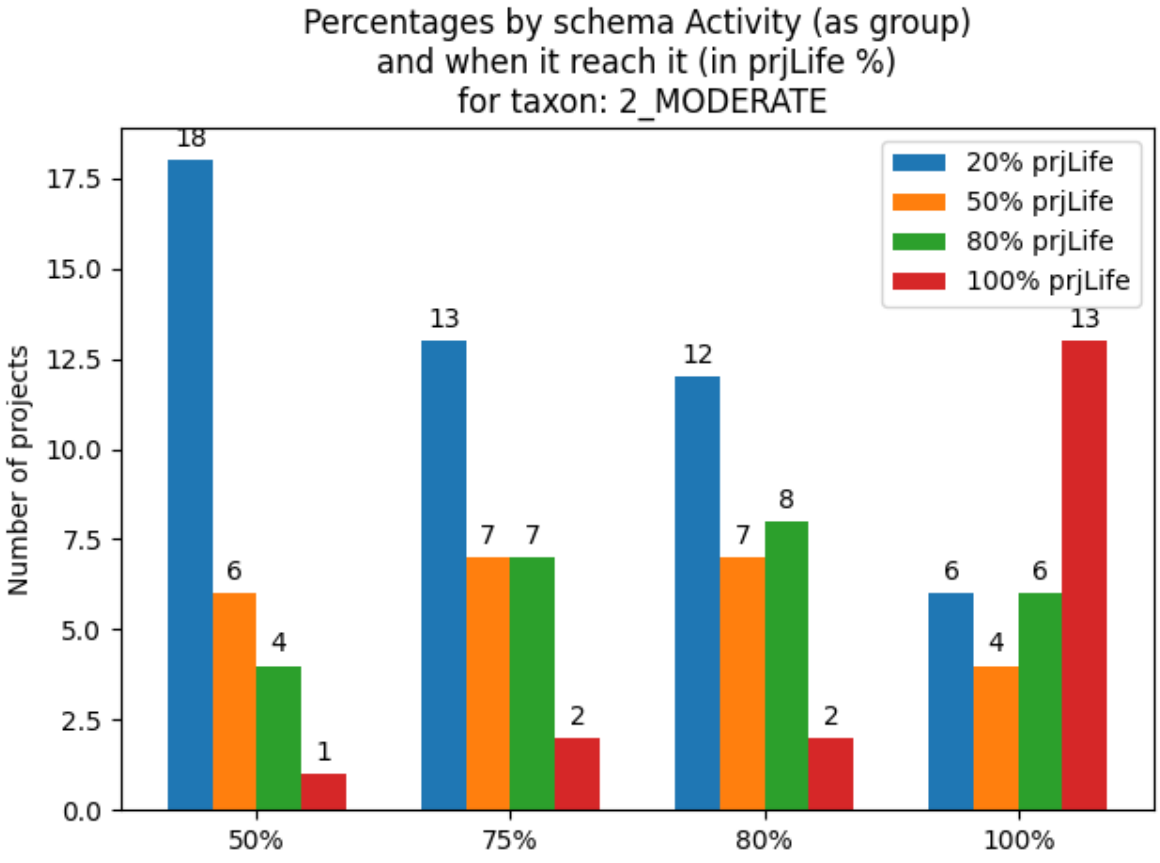
Number of projects that reached x% of schema activity, and when

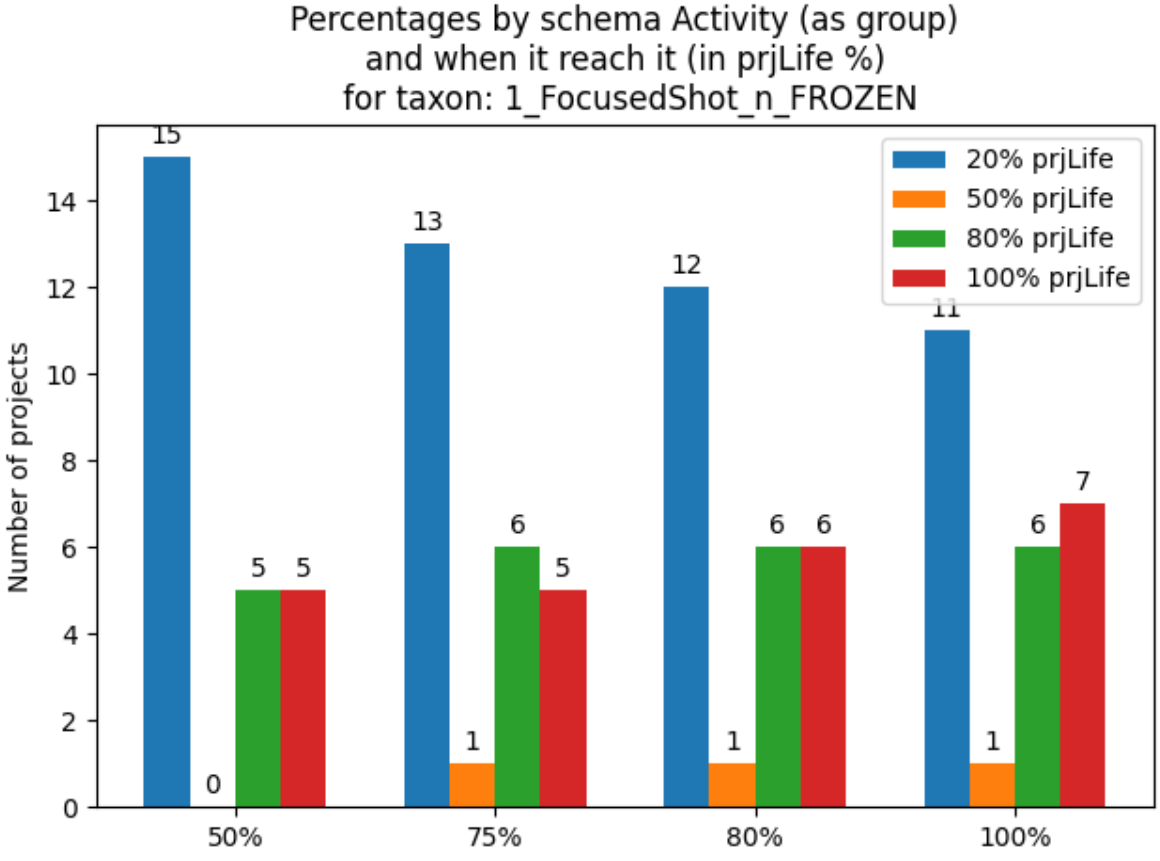
Taxon	Project Life	Schema Activity 50%	Schema Activity 75%	Schema Activity 80%	Schema Activity 100%
0_FROZEN	• 20% • 50% • 80% • 100%	• 21 • 5 • 5 • 2	• 21 • 6 • 5 • 2	• 21 • 6 • 5 • 2	• 21 • 6 • 5 • 2
1_ALMOST_FROZEN	• 20% • 50% • 80% • 100%	• 43 • 7 • 5 • 10	• 39 • 8 • 7 • 11	• 37 • 9 • 7 • 12	• 17 • 16 • 14 • 18
1_FocusedShot_n_FROZEN	• 20% • 50% • 80% • 100%	• 15 • 0 • 5 • 5	• 13 • 1 • 6 • 5	• 12 • 1 • 6 • 6	• 11 • 1 • 6 • 7
2_MODERATE	• 20% • 50% • 80% • 100%	• 18 • 6 • 4 • 1	• 13 • 7 • 7 • 2	• 12 • 7 • 8 • 2	• 6 • 4 • 6 • 13
3_FocusedShot_n_LOW	Task 5 • 20% • 50% • 80% • 100%	• 9 • 8 • 1 • 2	• 7 • 7 • 2 • 4	• 7 • 7 • 2 • 4	• 5 • 1 • 4 • 10
4_ACTIVE	• 20% • 50% • 80% • 100%	• 8 • 7 • 6 • 1	• 5 • 7 • 7 • 3	• 5 • 6 • 8 • 3	• 0 • 5 • 5 • 12
Overall	• 20% • 50% • 80% • 100%	• 114 • 34 • 126 • 21	• 98 • 36 • 34 • 27	• 94 • 36 • 36 • 29	• 60 • 33 • 40 • 62

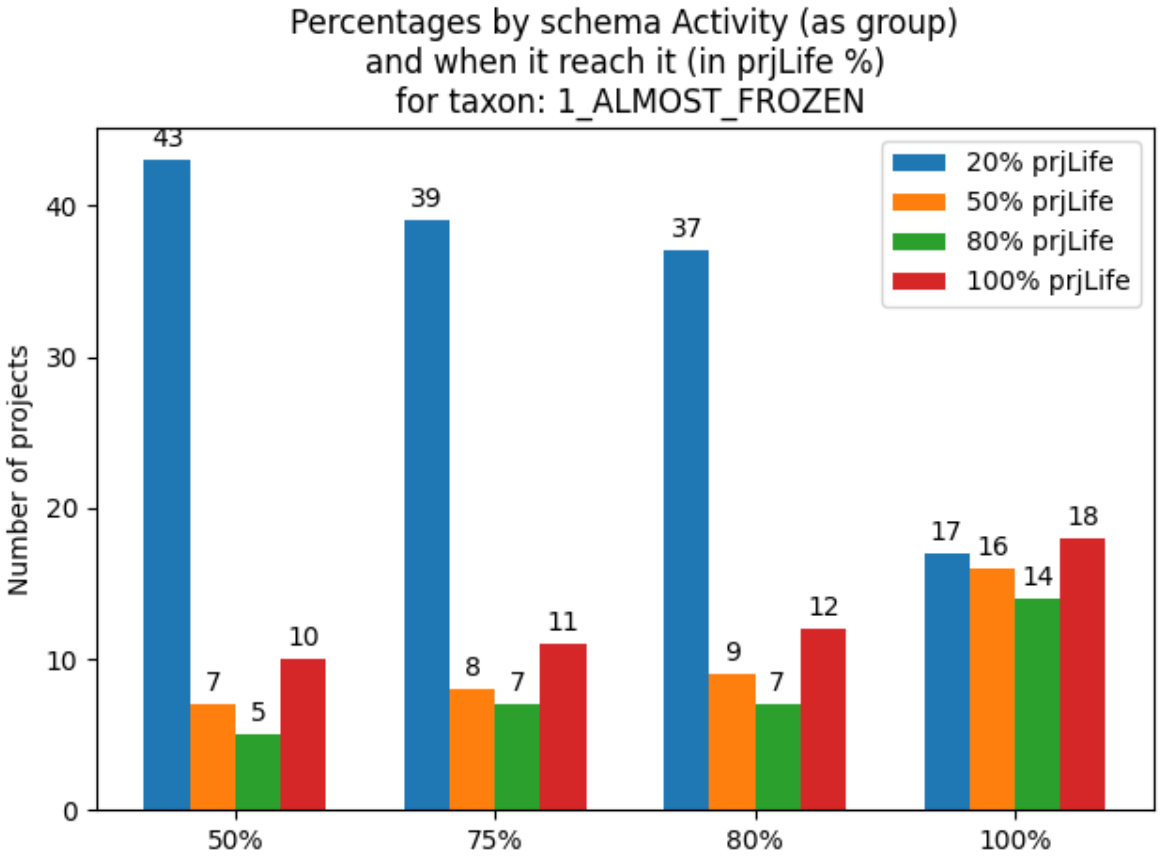

Percentages by schema Activity (as group)
and when it reach it (in prjLife %)
for taxon: Overall
Number of projects
20% prjLife
50% prjLife
80% prjLife
100% prjLife
114
34
26
21
98
36
34
27
94
36
36
29
60
33
40
62
0
20
40
60
80
100
50%
75%
80%
100%

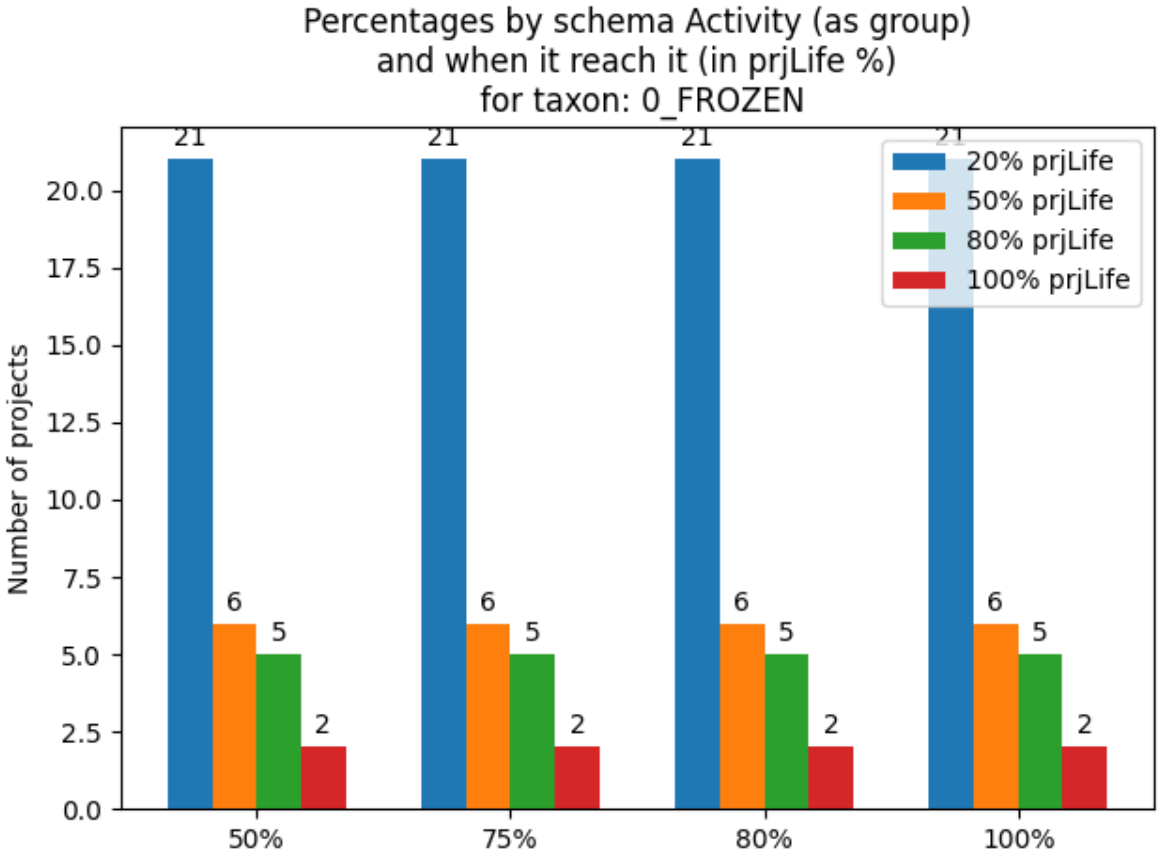

Percentages by schema Activity (as group)
and when it reach it (in prjLife %)
for taxon: 4_ACTIVE
20% prjLife
50% prjLife
80% prjLife
100% prjLife
Number of projects
0
2
4
6
8
10
12
8
7
6
1
5
7
7
3
5
6
8
3
0
5
5
12
50%
75%
80%
100%


Percentages by schema Activity (as group)
and when it reach it (in prjLife %)
for taxon: 3_FocusedShot_n_LOW
20% prjLife
50% prjLife
80% prjLife
100% prjLife
Number of projects
0
2
4
6
8
10
9
8
1
2
7
7
2
4
7
7
2
4
5
1
4
10
50%
75%
80%
100%

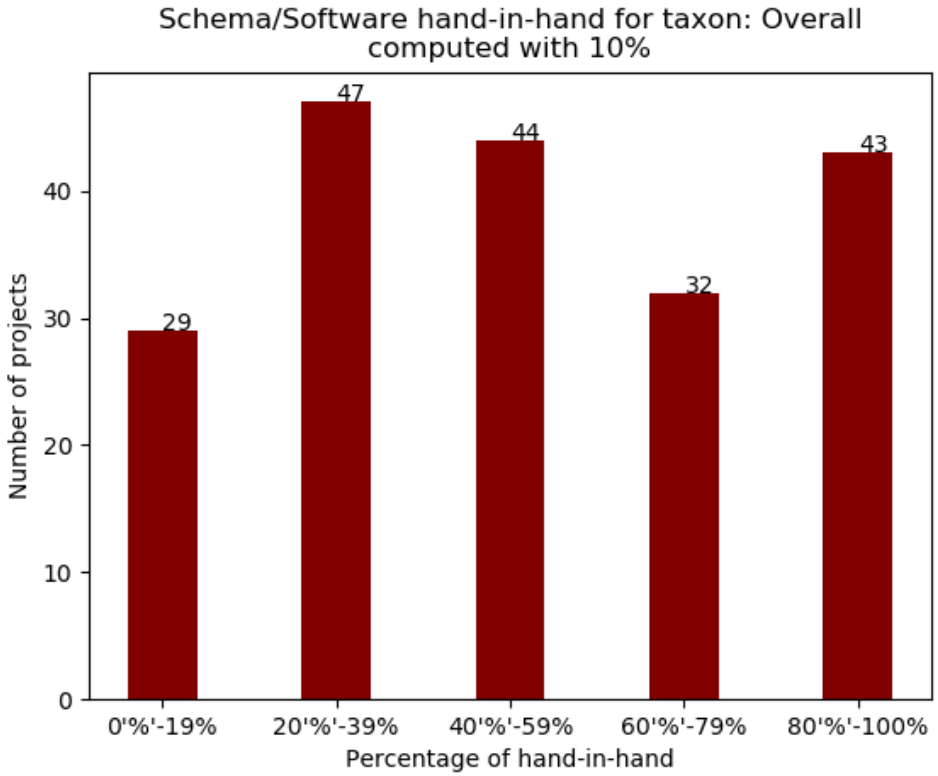

Percentages by schema Activity (as group)
and when it reach it (in prjLife %)
for taxon: 2_MODERATE
Number of projects
20% prjLife
50% prjLife
80% prjLife
100% prjLife
50%
75%
80%
100%
18
6
4
1
13
7
7
2
12
7
8
2
6
4
6
13


Percentages by schema Activity (as group)
and when it reach it (in prjLife %)
for taxon: 1_FocusedShot_n_FROZEN
Number of projects
20% prjLife
50% prjLife
80% prjLife
100% prjLife
50%
75%
80%
100%
15
0
5
5
13
1
6
5
12
1
6
6
11
1
6
7


Percentages by schema Activity (as group)
and when it reach it (in prjLife %)
for taxon: 1_ALMOST_FROZEN
Number of projects
20% prjLife
50% prjLife
80% prjLife
100% prjLife
43
7
5
10
39
8
7
11
37
9
7
12
17
16
14
18
50%
75%
80%
100%


Percentages by schema Activity (as group)
and when it reach it (in prjLife %)
for taxon: 0_FROZEN
Number of projects
20% prjLife
50% prjLife
80% prjLife
100% prjLife
21
6
5
2
50%
75%
80%
100%

7 Synchronicity Charts



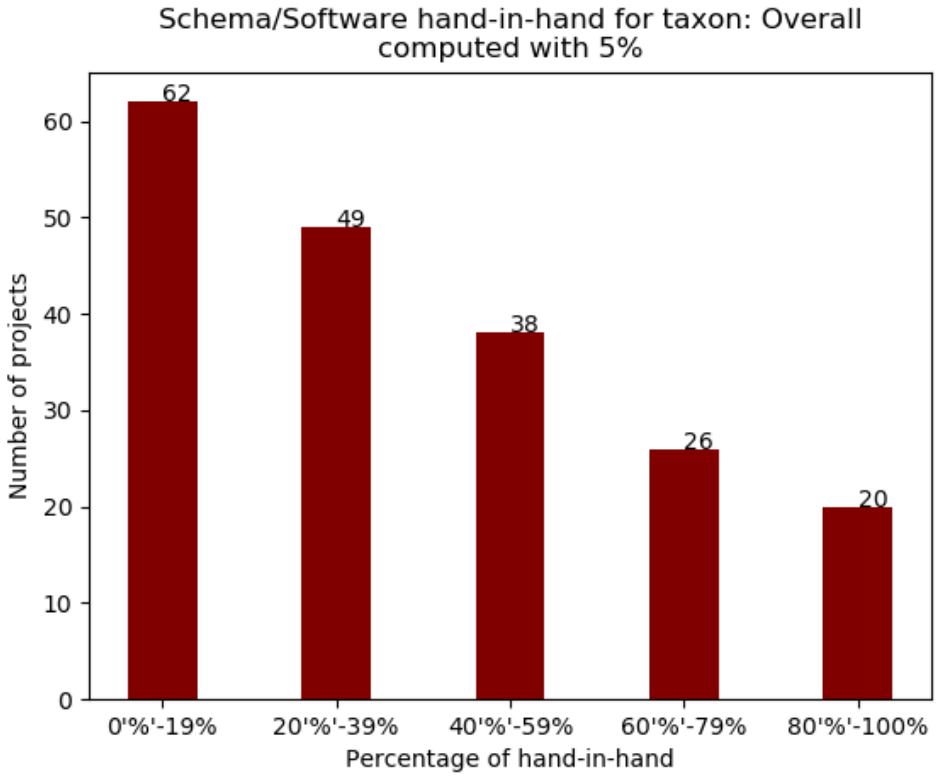
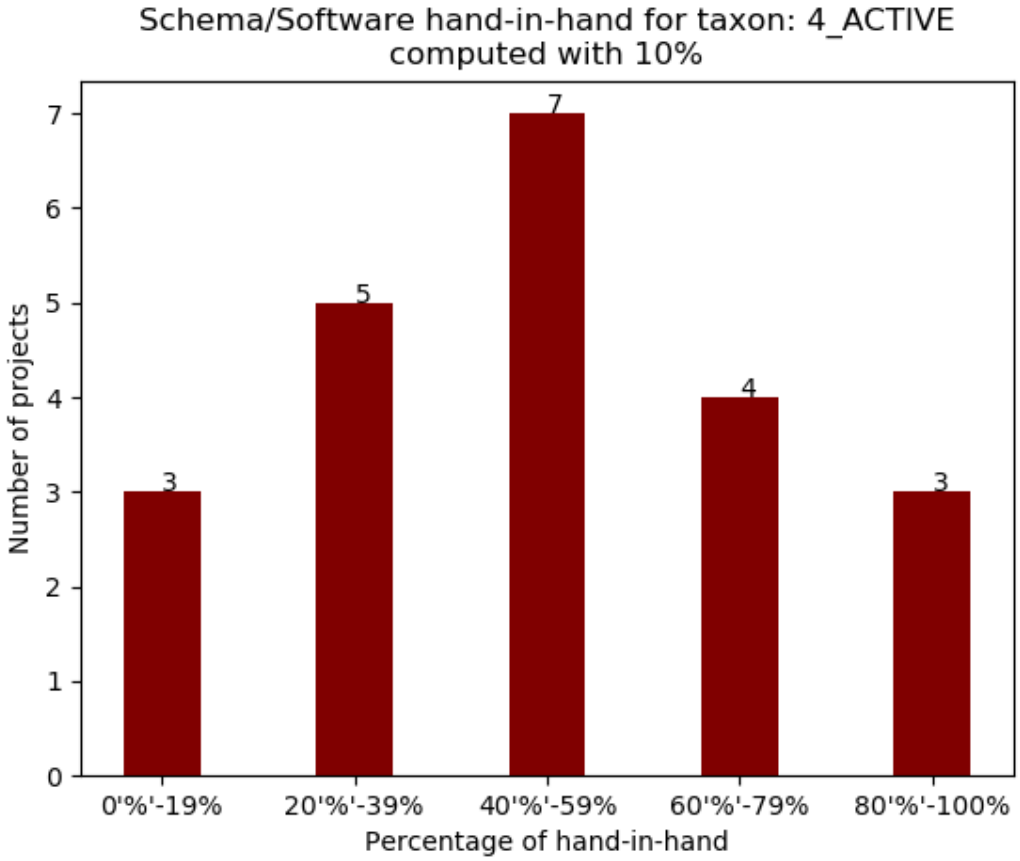


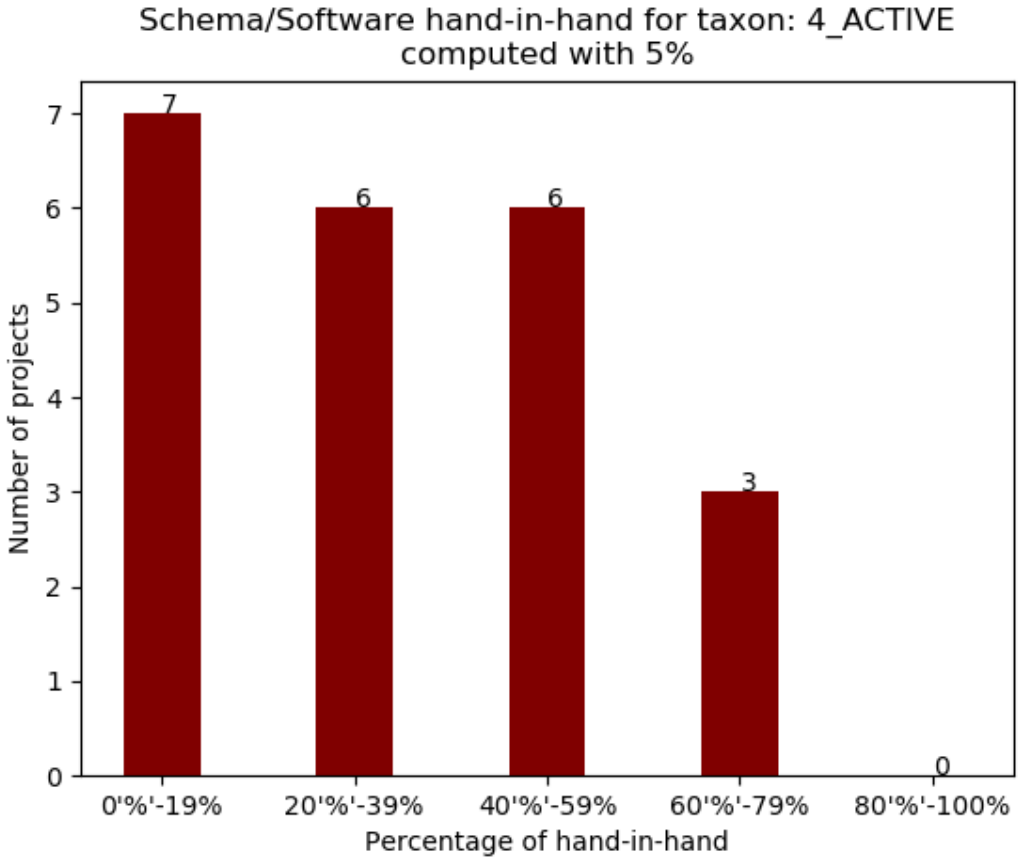
Table with 10% hand-in-hand

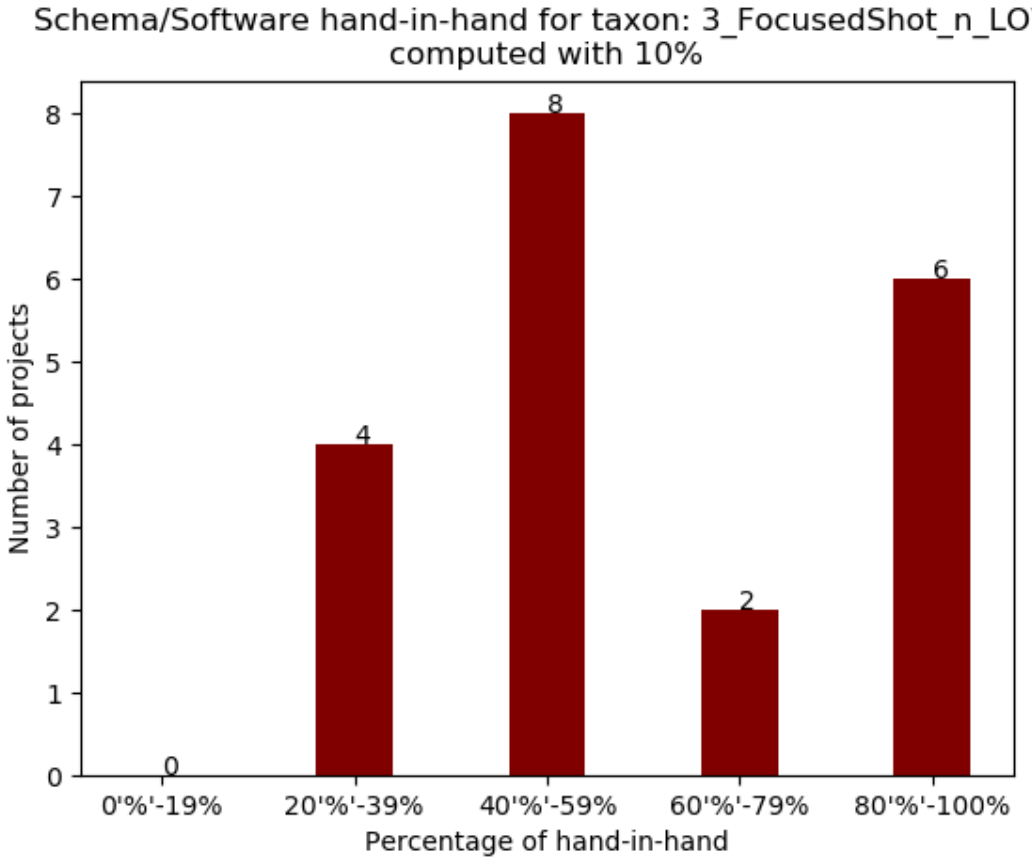
Taxon	0%-19%	20%-39%	40%-59%	60%-79%	80%-100%
0_FROZEN	6	9	4	7	8
1_ALMOST_FROZEN	13	18	9	12	13
1_FocusedShot_n_FROZEN5	2	4	5	5	9
2_MODERATE	5	7	11	2	4
3_FocusedShot_n_LOW	0	4	8	2	6
4_ACTIVE	3	5	7	4	3
Overall	29	47	44	32	43

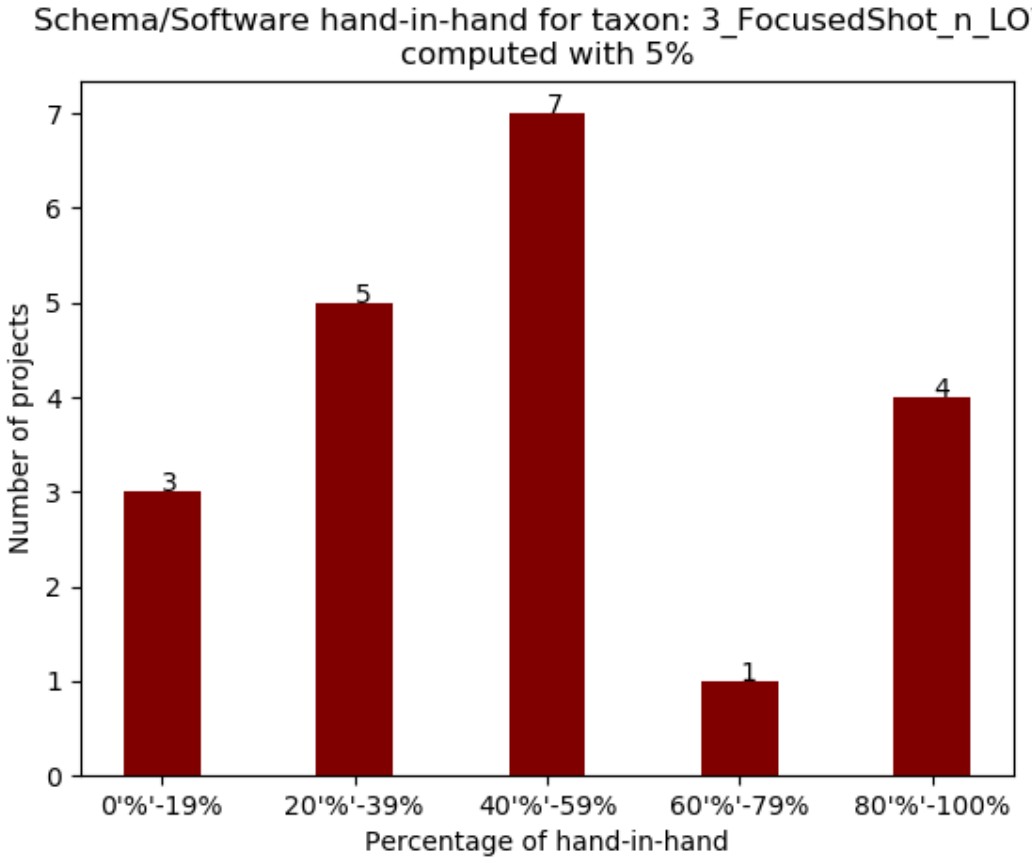
Table with 5% hand-in-hand

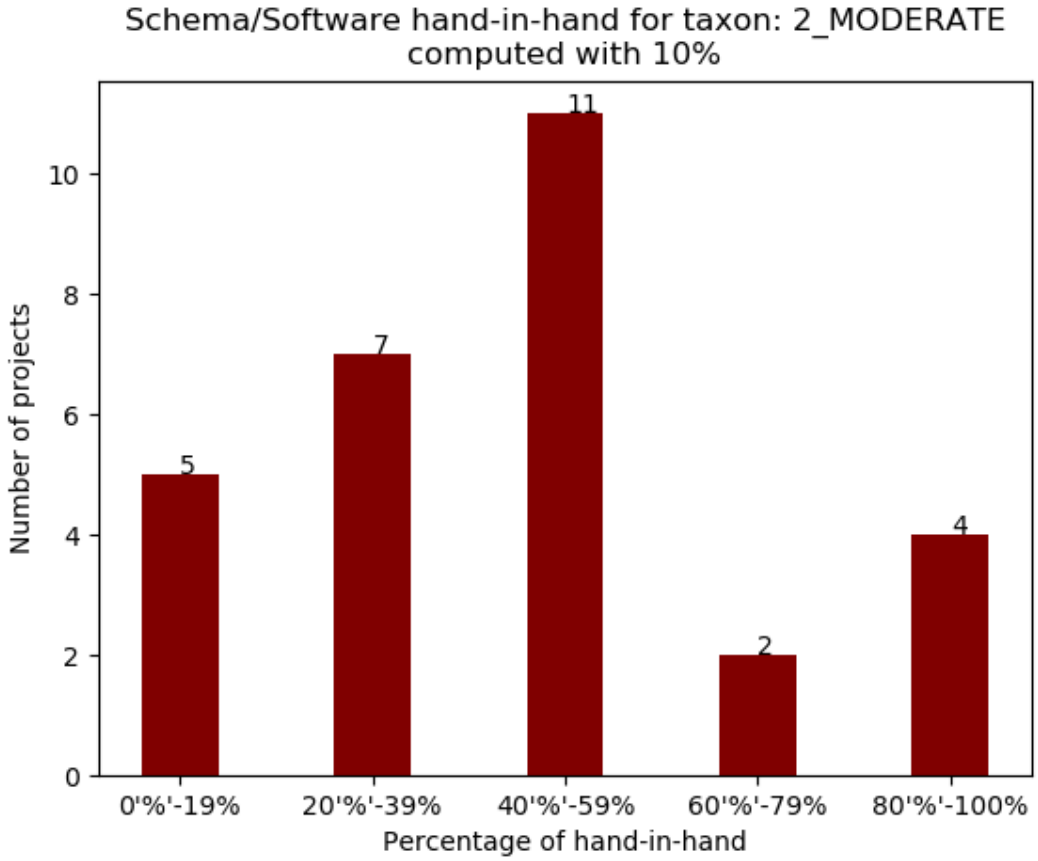
Taxon	0%-19%	20%-39%	40%-59%	60%-79%	80%-100%
0_FROZEN	13	5	6	5	5
1_ALMOST_FROZEN	23	15	10	11	6
1_FocusedShot_n_FROZEN5	5	8	4	4	4
2_MODERATE	11	10	5	2	1
3_FocusedShot_n_LOW	3	5	7	1	4
4_ACTIVE	7	6	6	3	0
Overall	62	49	38	26	20

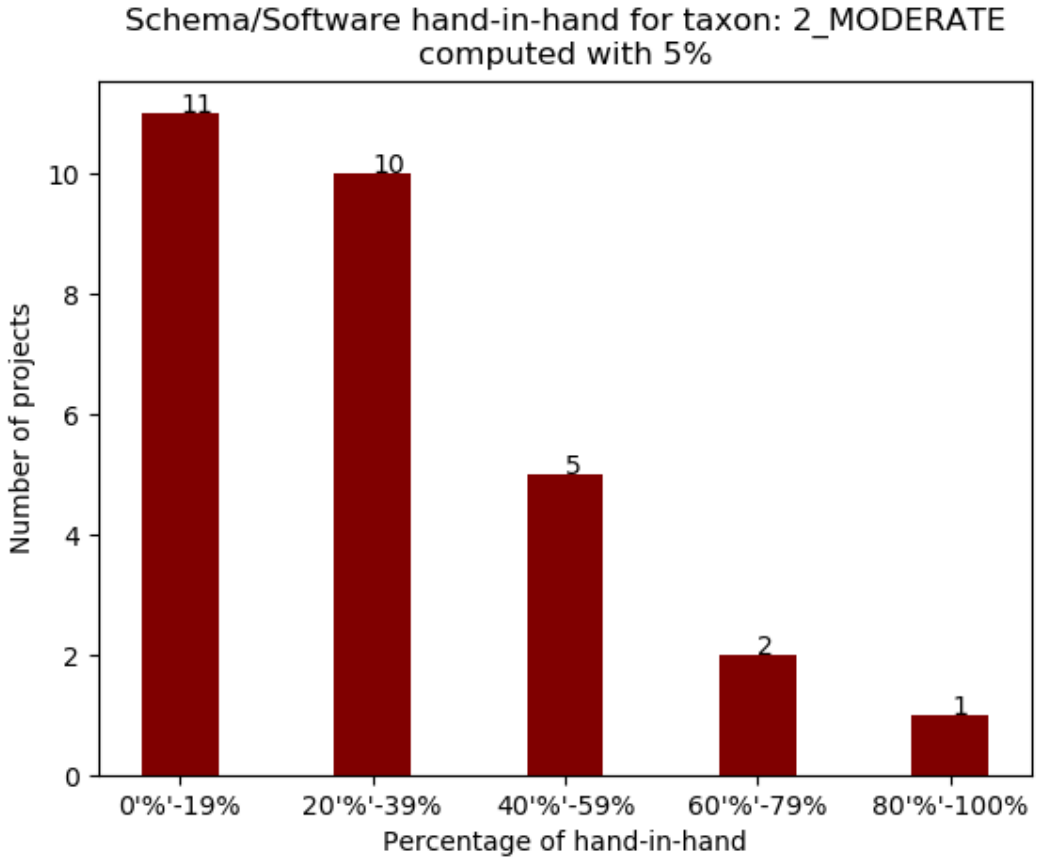

Schema/Software hand-in-hand for taxon: 4_ACTIVE
computed with 10%
Number of projects
Percentage of hand-in-hand
0'%'-19%
20'%'-39%
40'%'-59%
60'%'-79%
80'%'-100%
3
5
7
4
3

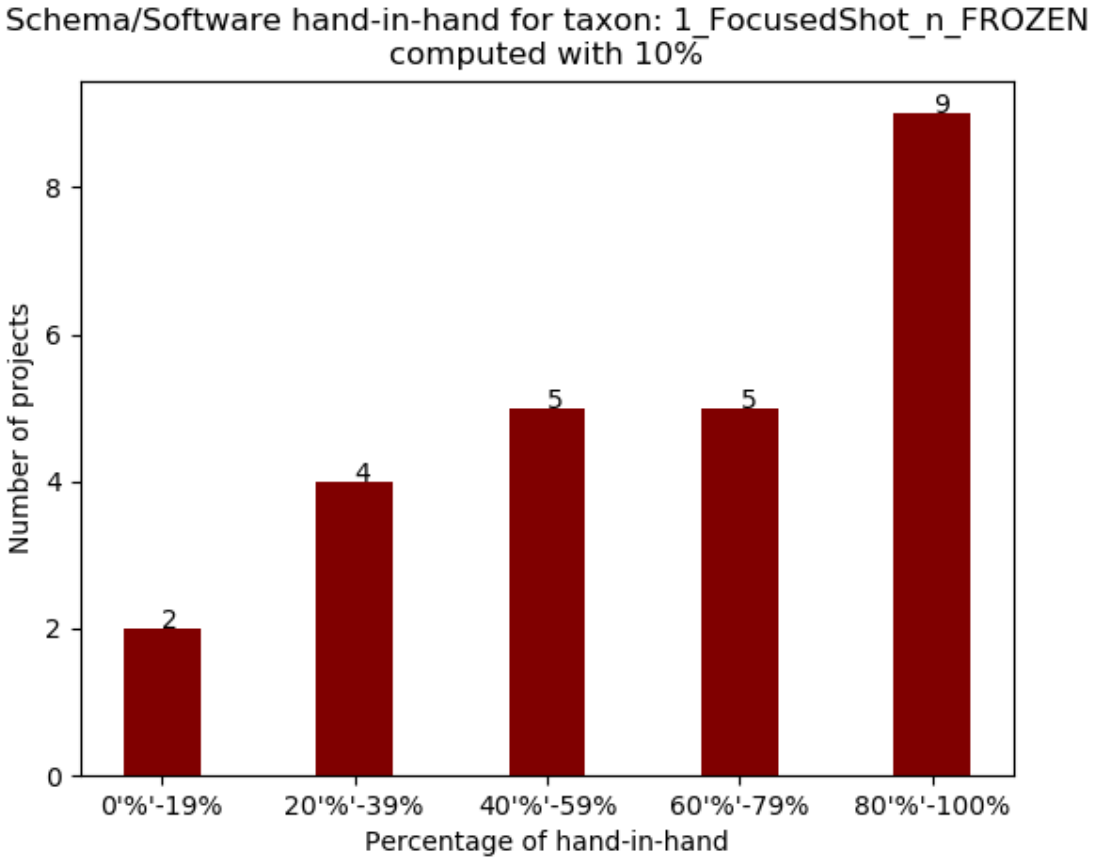

Schema/Software hand-in-hand for taxon: 4_ACTIVE
computed with 5%
Number of projects
Percentage of hand-in-hand
0'%'-19%
20'%'-39%
40'%'-59%
60'%'-79%
80'%'-100%
7
6
6
3
0

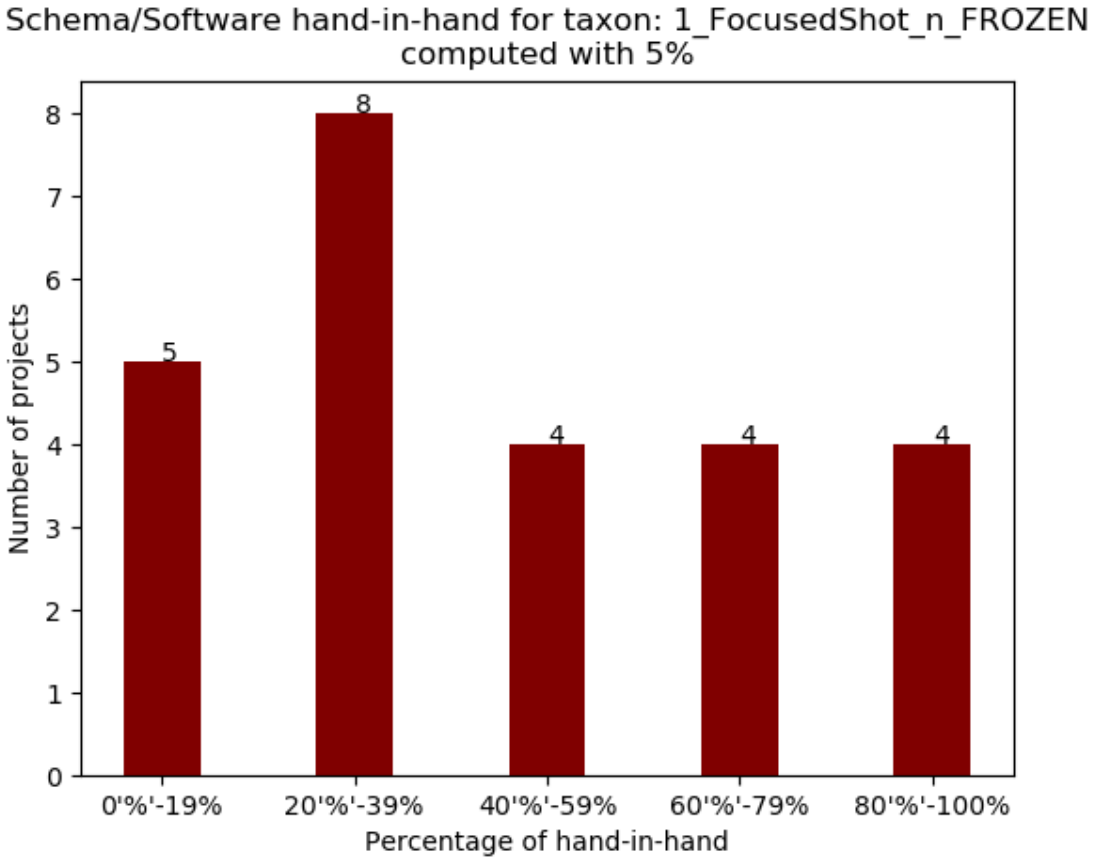

Schema/Software hand-in-hand for taxon: 3_FocusedShot_n_LOW
computed with 10%
Number of projects
Percentage of hand-in-hand
0'%'-19%
20'%'-39%
40'%'-59%
60'%'-79%
80'%'-100%
0
4
8
2
6

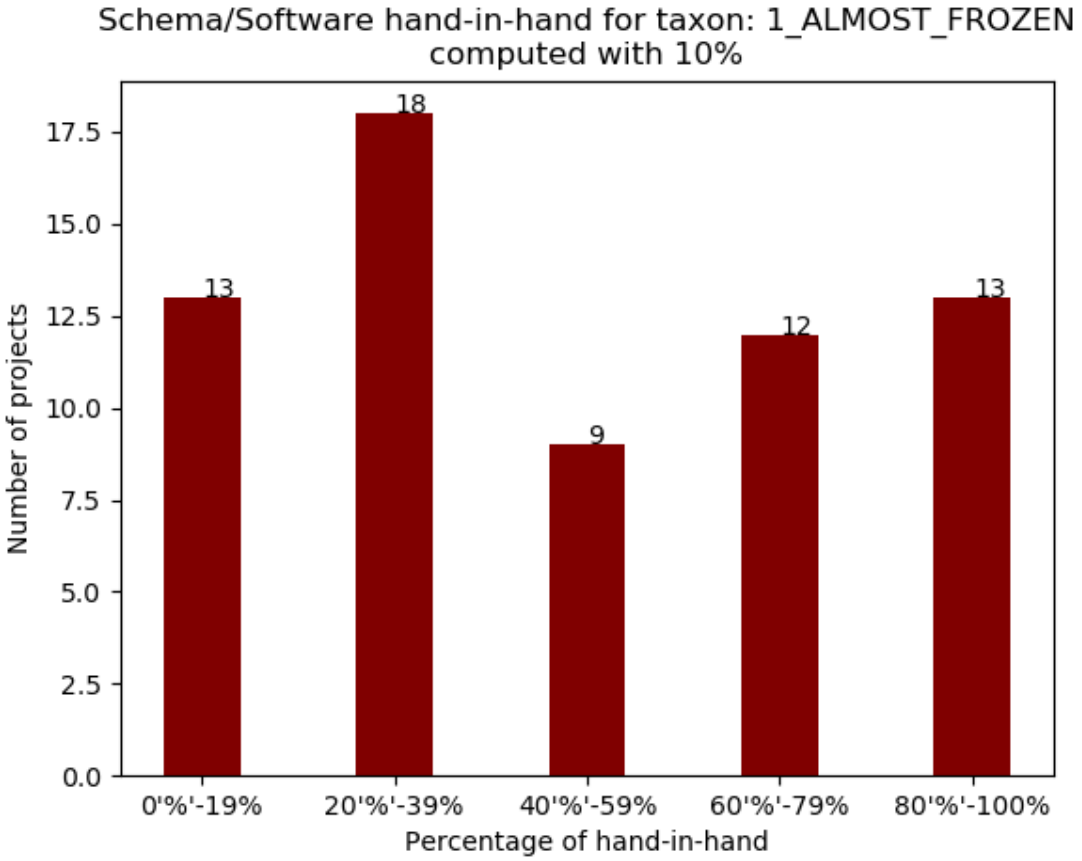

Schema/Software hand-in-hand for taxon: 3_FocusedShot_n_LOW
computed with 5%
Number of projects
Percentage of hand-in-hand
0'%'-19%
20'%'-39%
40'%'-59%
60'%'-79%
80'%'-100%
3
5
7
1
4

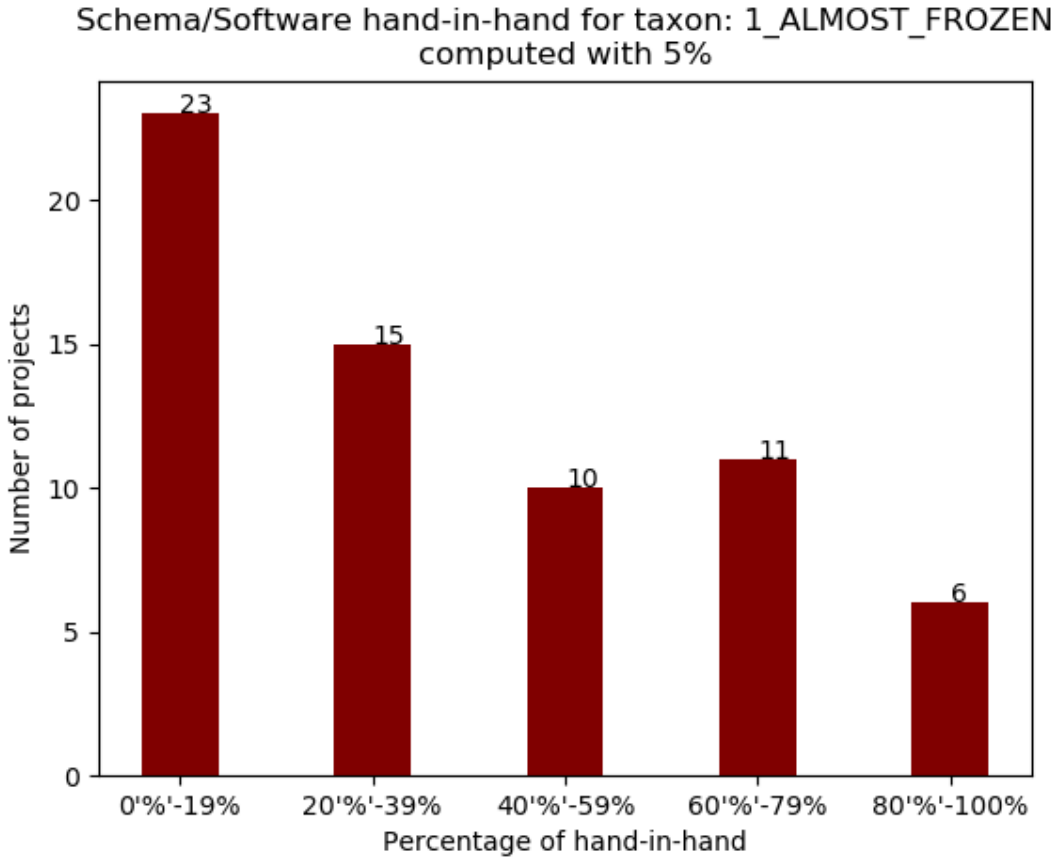

Schema/Software hand-in-hand for taxon: 2_MODERATE
computed with 10%
Number of projects
Percentage of hand-in-hand
0'%'-19%
20'%'-39%
40'%'-59%
60'%'-79%
80'%'-100%
5
7
11
2
4

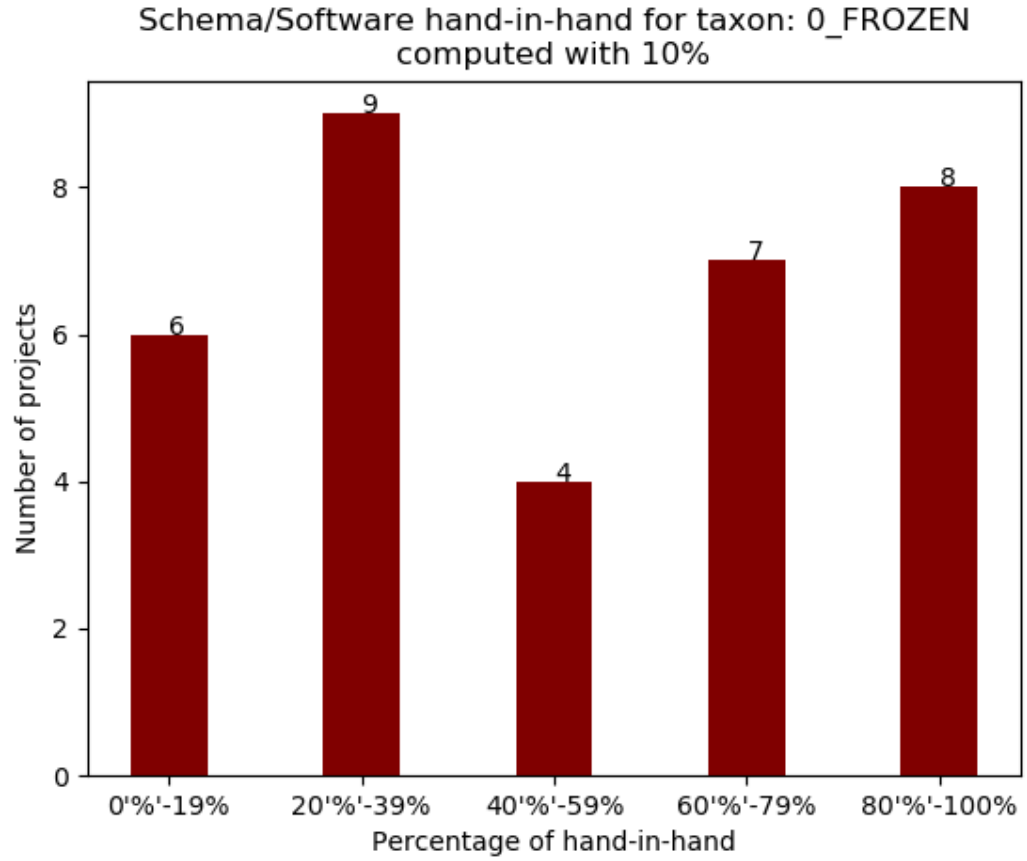

Schema/Software hand-in-hand for taxon: 2_MODERATE
computed with 5%
Number of projects
Percentage of hand-in-hand
0'%'-19%
20'%'-39%
40'%'-59%
60'%'-79%
80'%'-100%
11
10
5
2
1

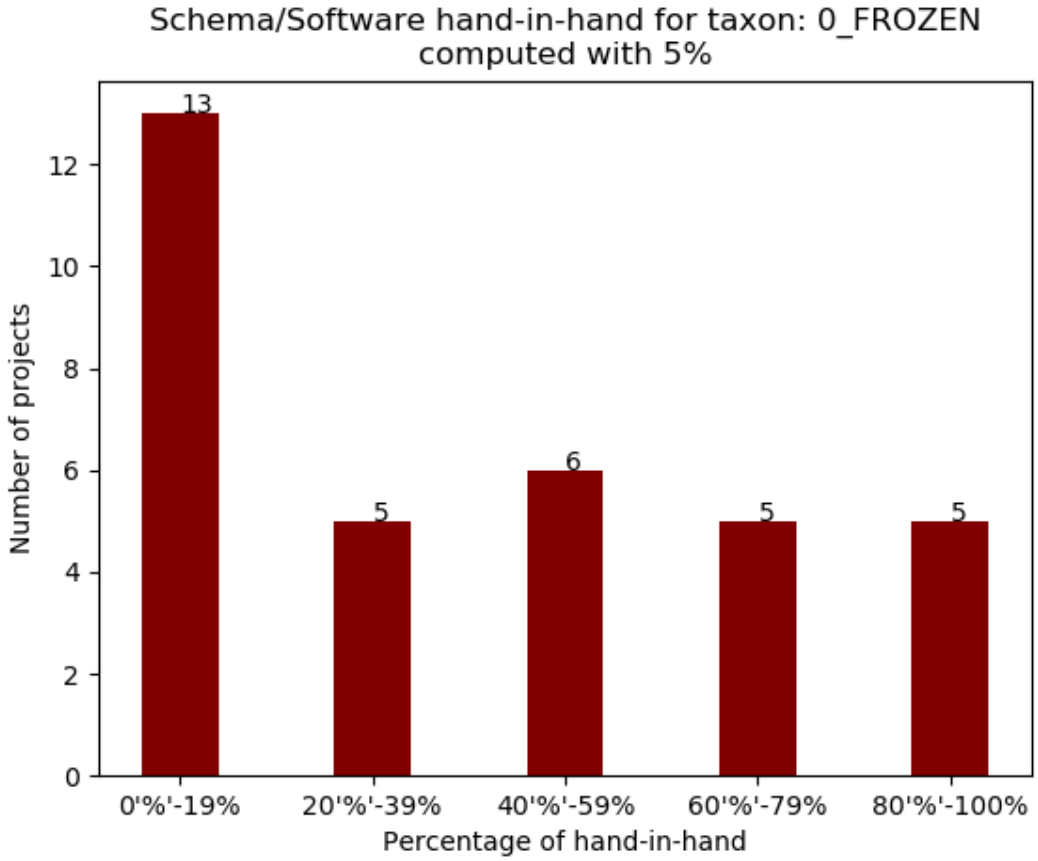

Schema/Software hand-in-hand for taxon: 1_FocusedShot_n_FROZEN
computed with 10%
Number of projects
Percentage of hand-in-hand
0'%'-19%
20'%'-39%
40'%'-59%
60'%'-79%
80'%'-100%
2
4
5
5
9


Schema/Software hand-in-hand for taxon: 1_FocusedShot_n_FROZEN
computed with 5%
Number of projects
Percentage of hand-in-hand
0'%'-19%
20'%'-39%
40'%'-59%
60'%'-79%
80'%'-100%
5
8
4
4
4


Schema/Software hand-in-hand for taxon: 1_ALMOST_FROZEN
computed with 10%
Number of projects
Percentage of hand-in-hand
0'%'-19%
20'%'-39%
40'%'-59%
60'%'-79%
80'%'-100%
18
13
9
12
13
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

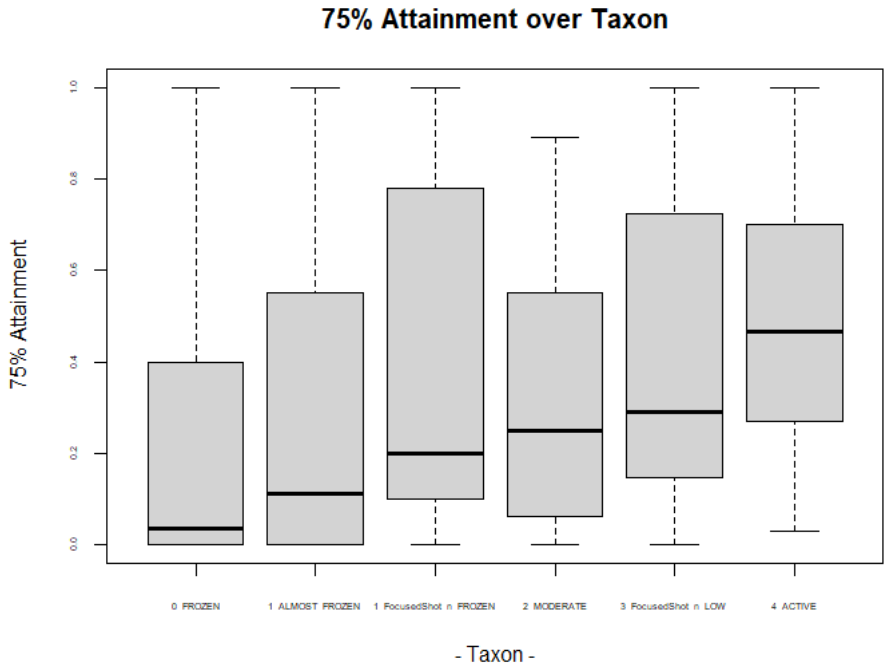

Schema/Software hand-in-hand for taxon: 1_ALMOST_FROZEN
computed with 5%
Number of projects
Percentage of hand-in-hand
0'%'-19%
20'%'-39%
40'%'-59%
60'%'-79%
80'%'-100%
23
15
10
11
6
0
5
10
15
20

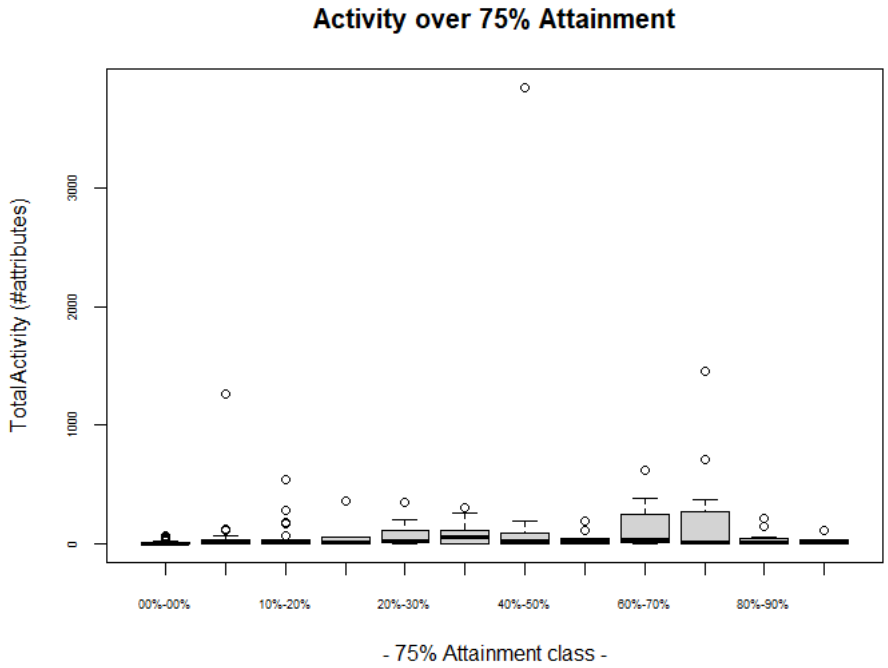

Schema/Software hand-in-hand for taxon: 0_FROZEN
computed with 10%
Number of projects
Percentage of hand-in-hand
0'%'-19%
20'%'-39%
40'%'-59%
60'%'-79%
80'%'-100%
6
9
4
7
8
0
2
4
6
8

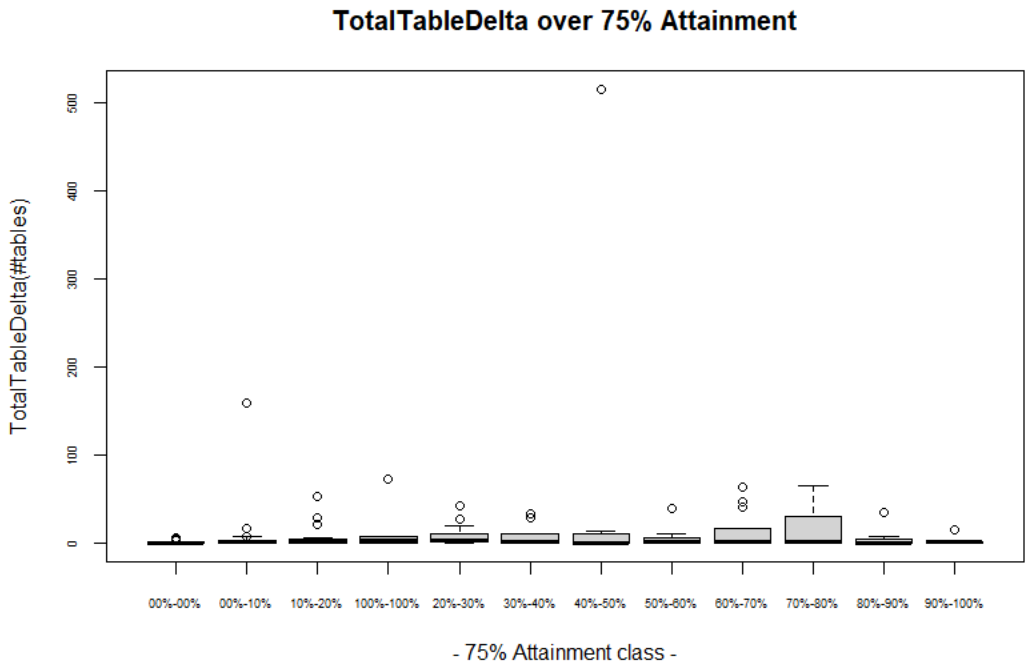

Schema/Software hand-in-hand for taxon: 0_FROZEN
computed with 5%
Number of projects
Percentage of hand-in-hand
0'%'-19%
20'%'-39%
40'%'-59%
60'%'-79%
80'%'-100%
13
5
6
5
5
0
2
4
6
8
10
12

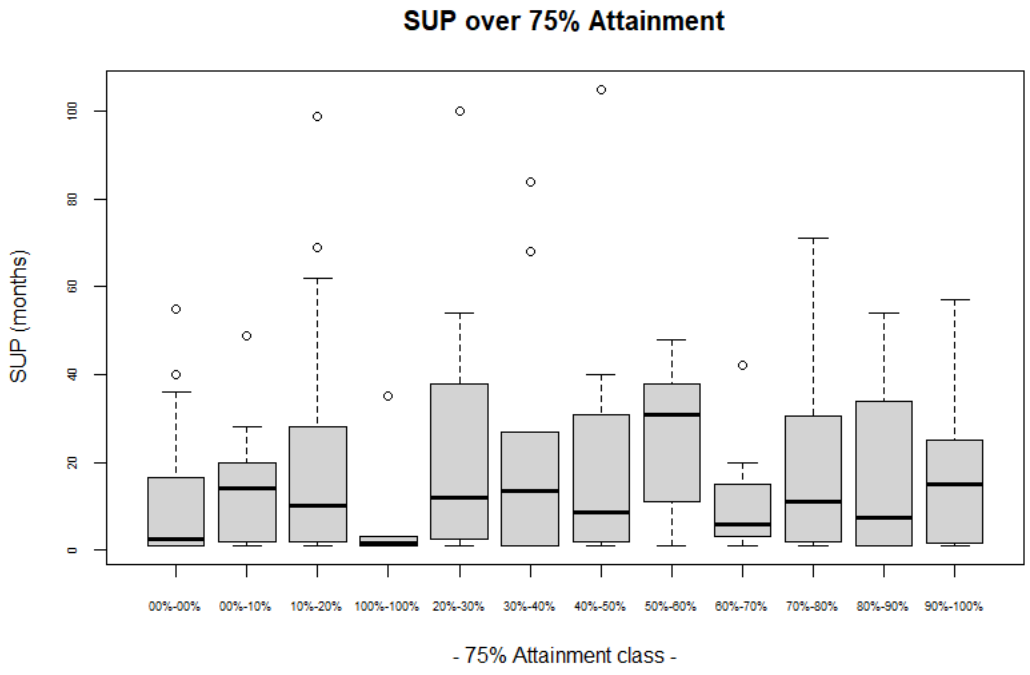
8 Attainment Boxplots over Potentially Interesting Attributes

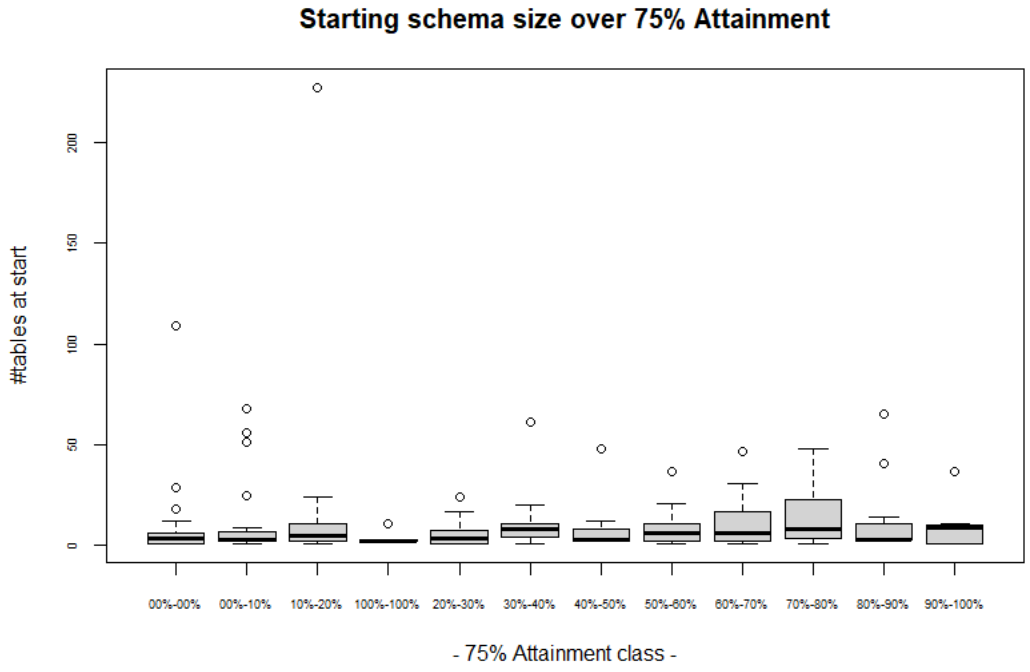
(no outliers removed)

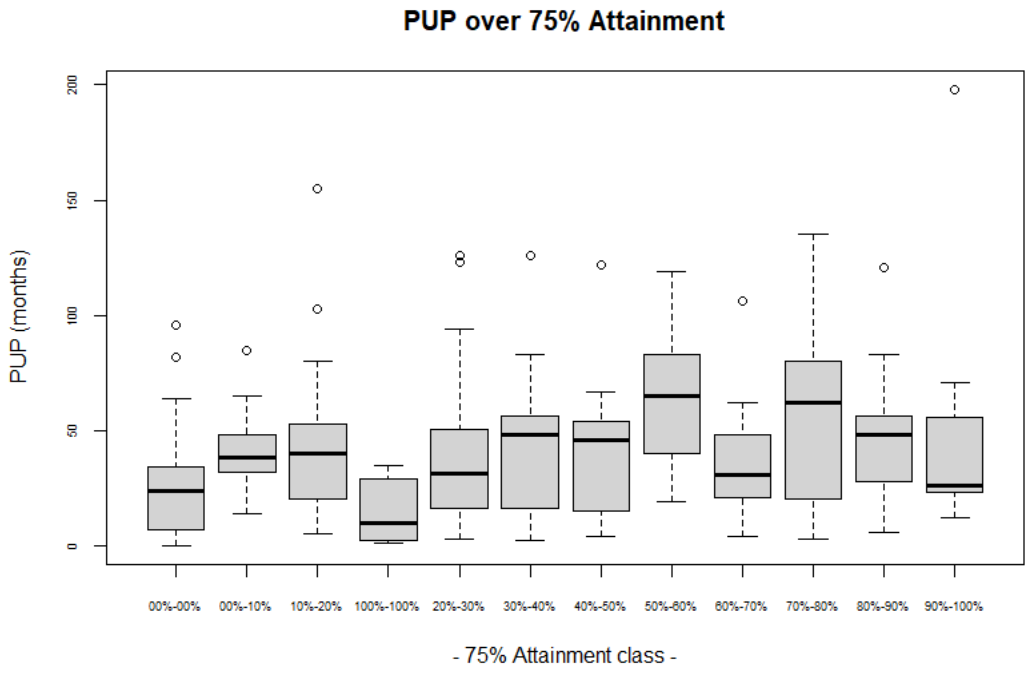


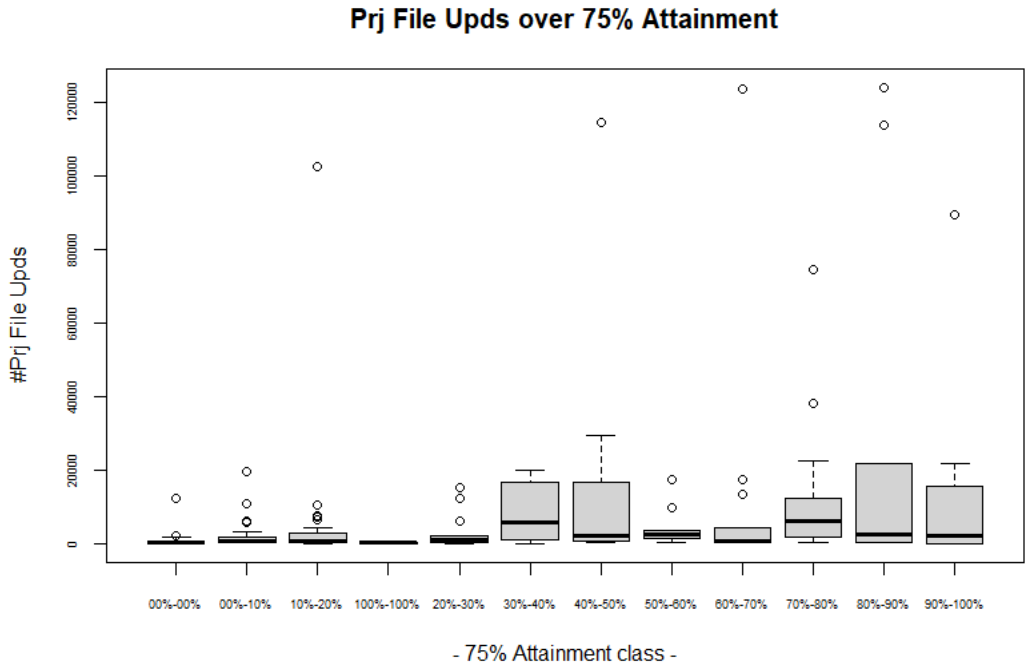



TotalTableDelta over 75% Attainment
TotalTableDelta(#tables)
0
100
200
300
400
500
00%-00%
00%-10%
10%-20%
100%-100%
20%-30%
30%-40%
40%-50%
50%-60%
60%-70%
70%-80%
80%-90%
90%-100%
- 75% Attainment class -


SUP over 75% Attainment
SUP (months)
0
20
40
60
80
100
00%-00%
00%-10%
10%-20%
100%-100%
20%-30%
30%-40%
40%-50%
50%-60%
60%-70%
70%-80%
80%-90%
90%-100%
- 75% Attainment class -

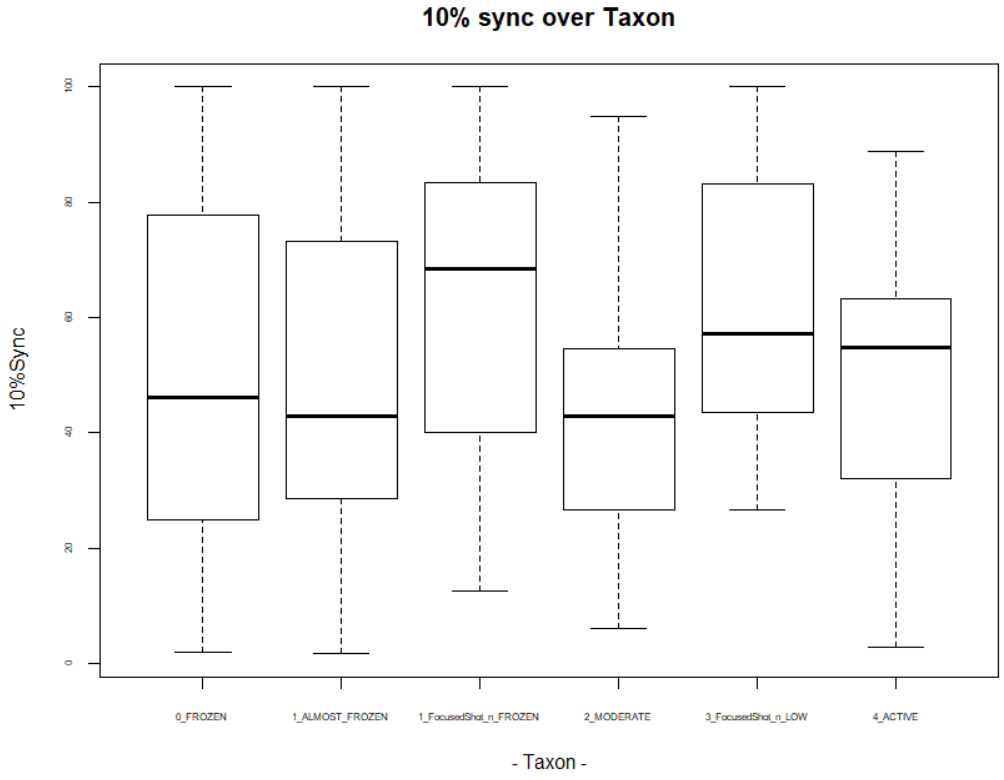

Starting schema size over 75% Attainment
#tables at start
0
50
100
150
200
00%-00%
00%-10%
10%-20%
100%-100%
20%-30%
30%-40%
40%-50%
50%-60%
60%-70%
70%-80%
80%-90%
90%-100%
- 75% Attainment class -

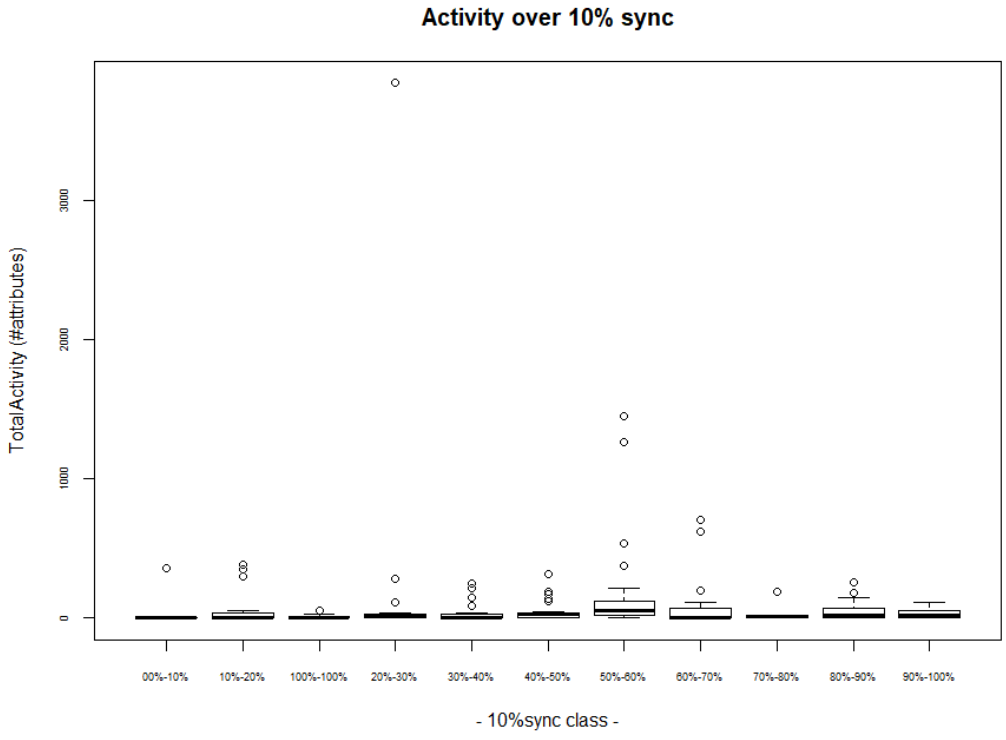

PUP over 75% Attainment
PUP (months)
0
50
100
150
200
00%-00%
00%-10%
10%-20%
100%-100%
20%-30%
30%-40%
40%-50%
50%-60%
60%-70%
70%-80%
80%-90%
90%-100%
- 75% Attainment class -

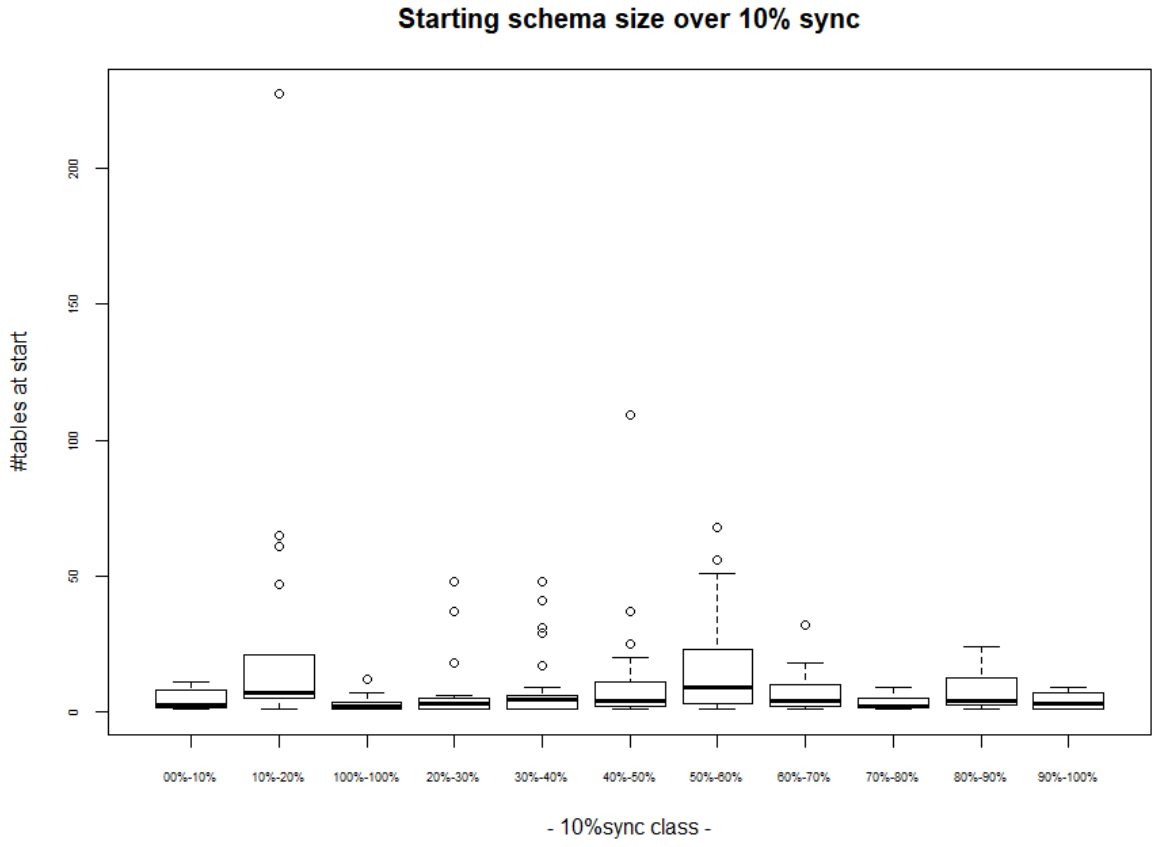

Prj File Upds over 75% Attainment
#Prj File Upds
0
20000
40000
60000
80000
100000
120000
00%-00%
00%-10%
10%-20%
100%-100%
20%-30%
30%-40%
40%-50%
50%-60%
60%-70%
70%-80%
80%-90%
90%-100%
- 75% Attainment class -

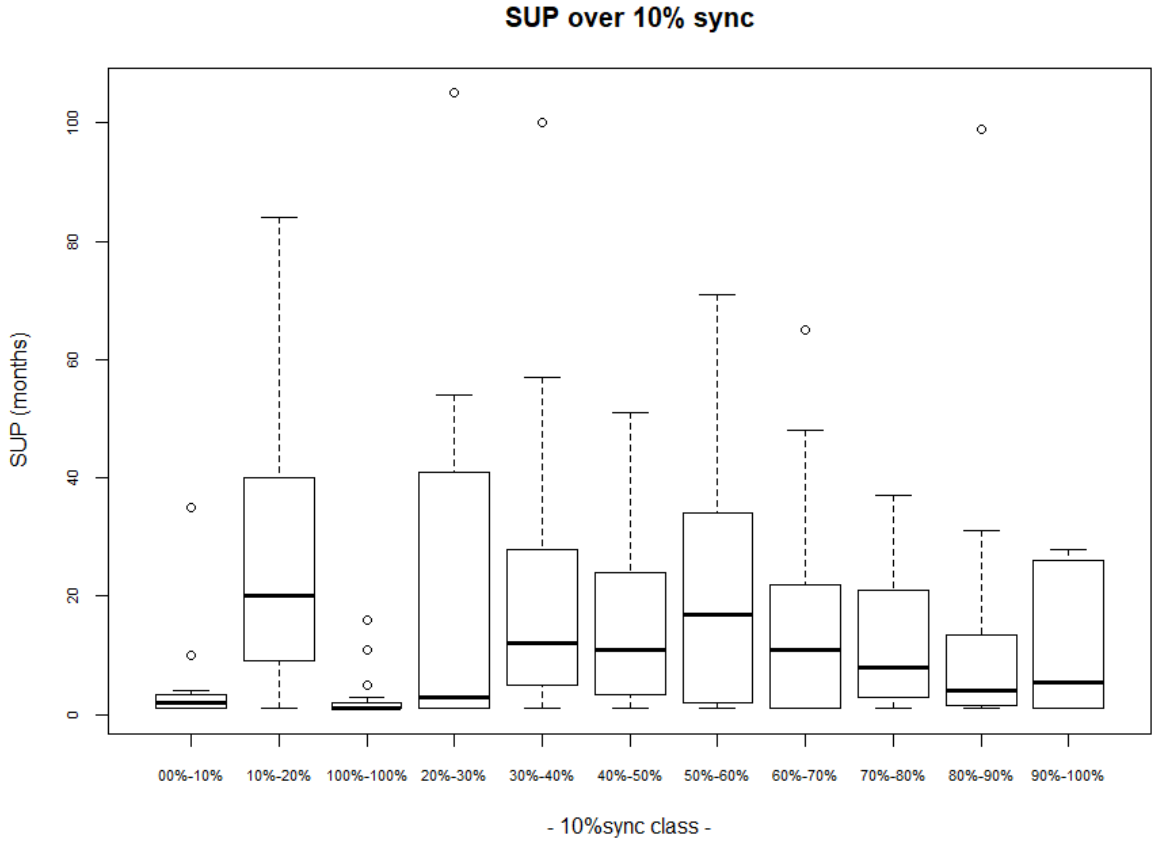
9 Synchronicity Boxplots over Potentially Interesting Attributes

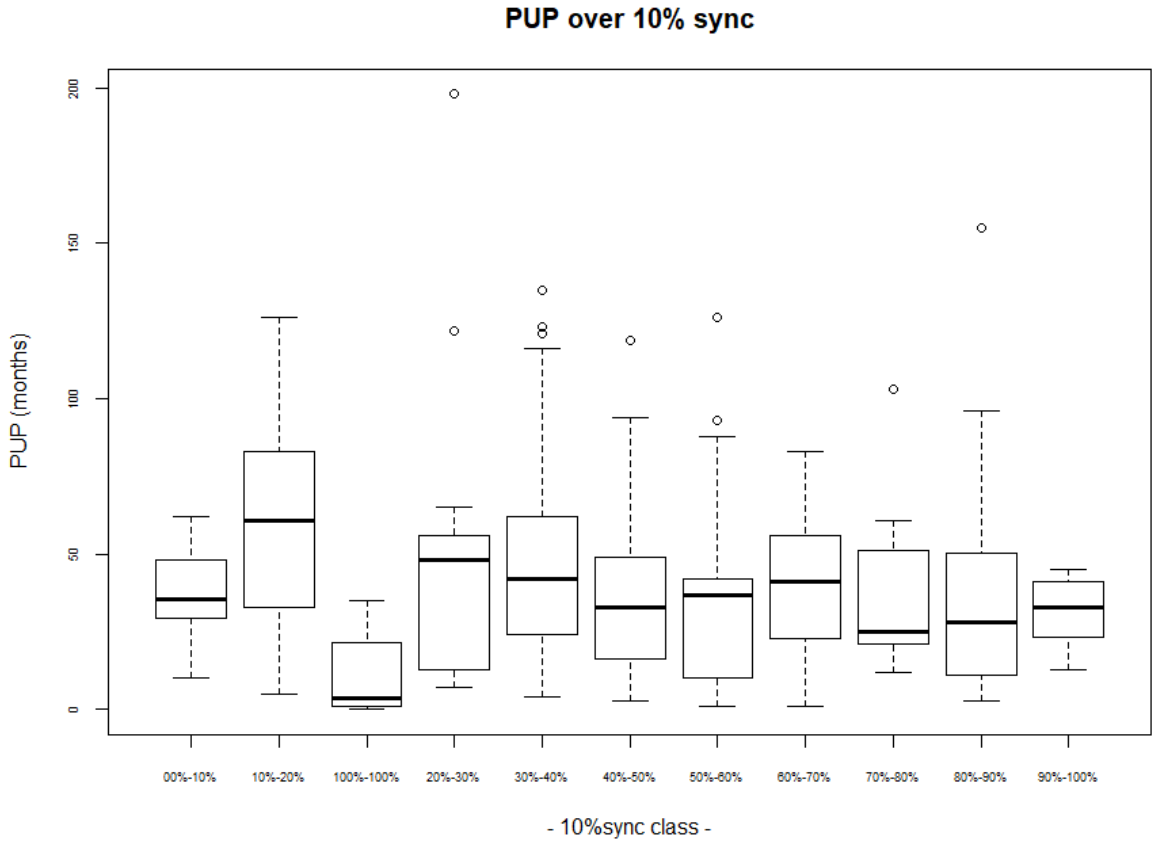
(no outliers removed)

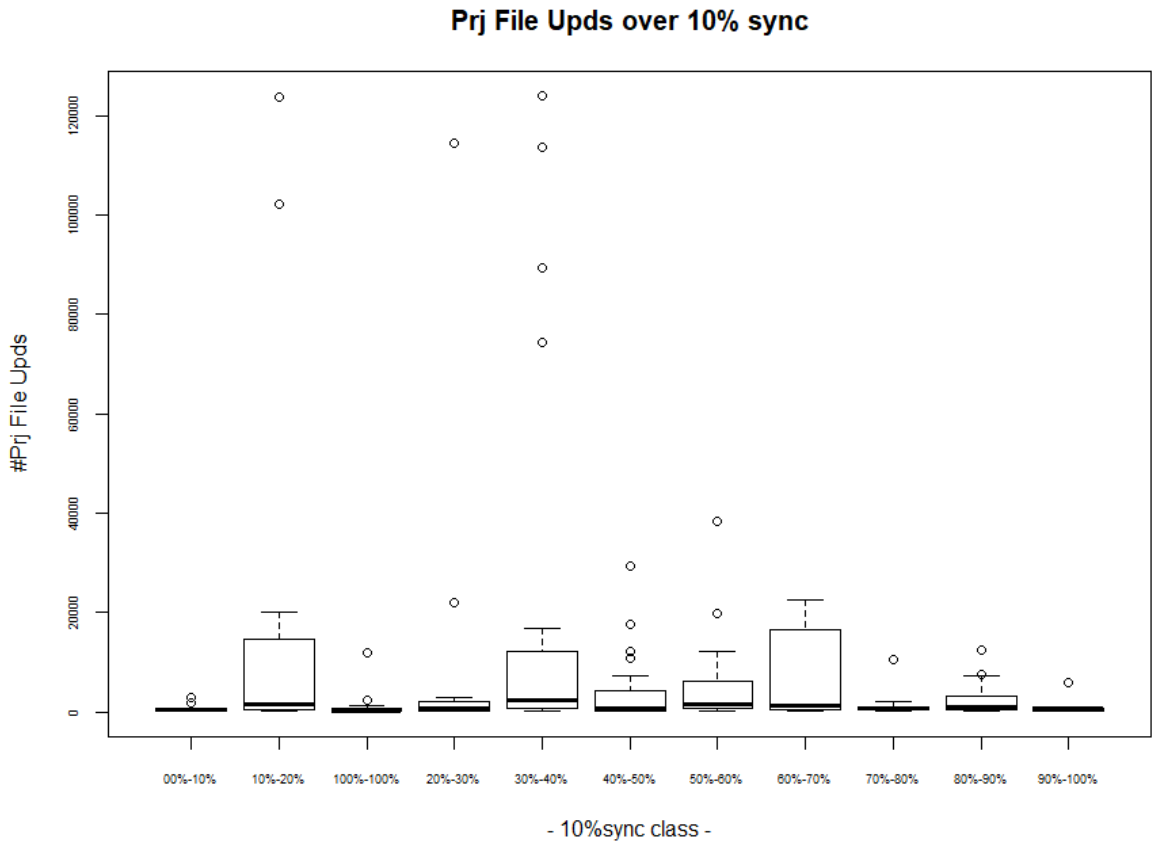


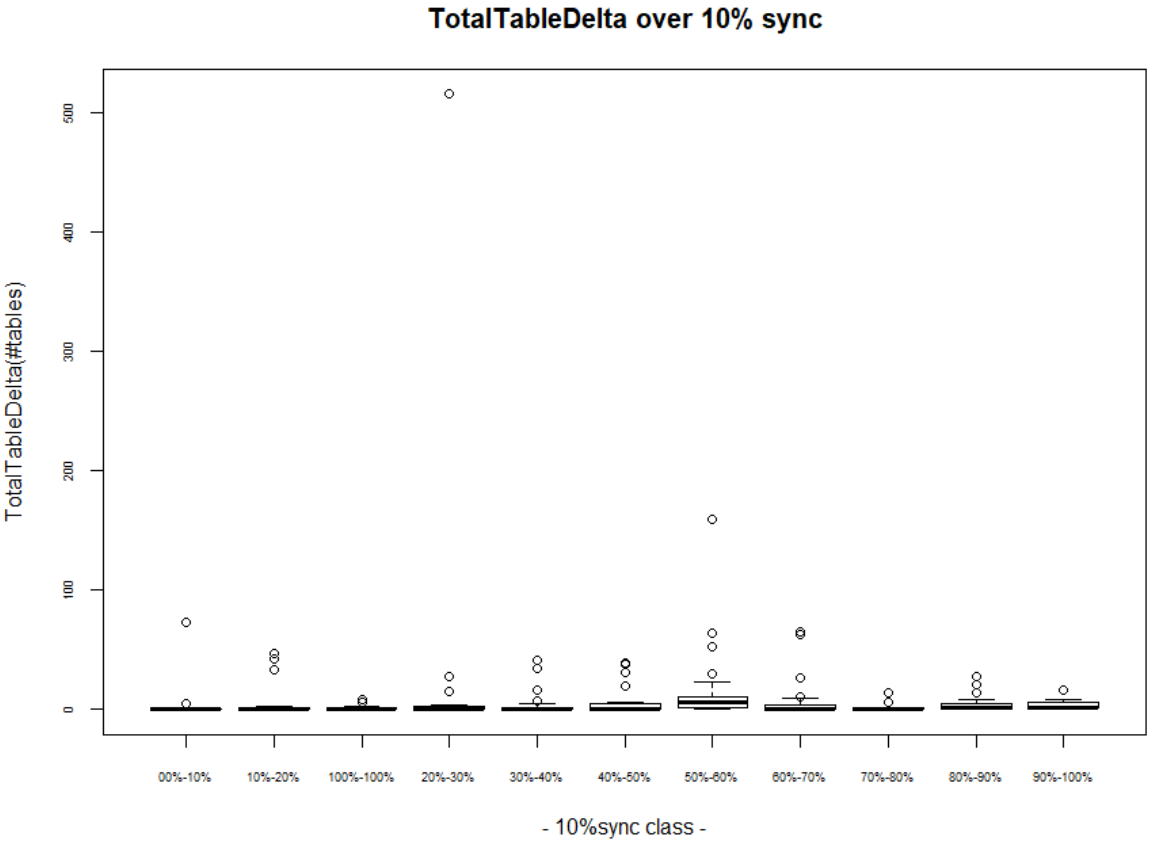



Starting schema size over 10% sync
#tables at start
0
50
100
150
200
00%-10%
10%-20%
100%-100%
20%-30%
30%-40%
40%-50%
50%-60%
60%-70%
70%-80%
80%-90%
90%-100%
- 10%sync class -


SUP over 10% sync
SUP (months)
0
20
40
60
80
100
00%-10%
10%-20%
100%-100%
20%-30%
30%-40%
40%-50%
50%-60%
60%-70%
70%-80%
80%-90%
90%-100%
- 10%sync class -


PUP over 10% sync
PUP (months)
0
50
100
150
200
00%-10%
10%-20%
100%-100%
20%-30%
30%-40%
40%-50%
50%-60%
60%-70%
70%-80%
80%-90%
90%-100%
- 10%sync class -


Prj File Upds over 10% sync
#Prj File Upds
0
20000
40000
60000
80000
100000
120000
00%-10%
10%-20%
100%-100%
20%-30%
30%-40%
40%-50%
50%-60%
60%-70%
70%-80%
80%-90%
90%-100%
- 10%sync class -


TotalTableDelta over 10% sync
TotalTableDelta(#tables)
0
100
200
300
400
500
00%-10%
10%-20%
100%-100%
20%-30%
30%-40%
40%-50%
50%-60%
60%-70%
70%-80%
80%-90%
90%-100%
- 10%sync class -