Detecting the Emergence of Novel Fish Communities

1. Novelty Detection Framework

The novelty detection framework by Pandolfi et al uses two signals of compositional change to identify novel communities in time series data. These two metrics are: 1) Sequential dissimilarity, defined as composition dissimilarity between time T and T-1 and 2) Minimum dissimilarity, defined as the smallest dissimilarity between time T and any time before T. The mean dissimilarities for both metrics are modeled using GAM's, which allow for flexible expectations of compositional change over time, within one particular time series. Beta regression is then used to establish a distribution around each bin (which represents a point in time). Instantaneous and cumulative dissimilarity values that exceed the arbitrary 95% predictive boundary are categorized as instantaneous and cumulative novelty respectively. Bins/communities that exhibit both cumulative and instantaneous novelty are termed truly novel (*bin and community are used interchangeably here, as a bin represents the community state at a certain time within the time series*).

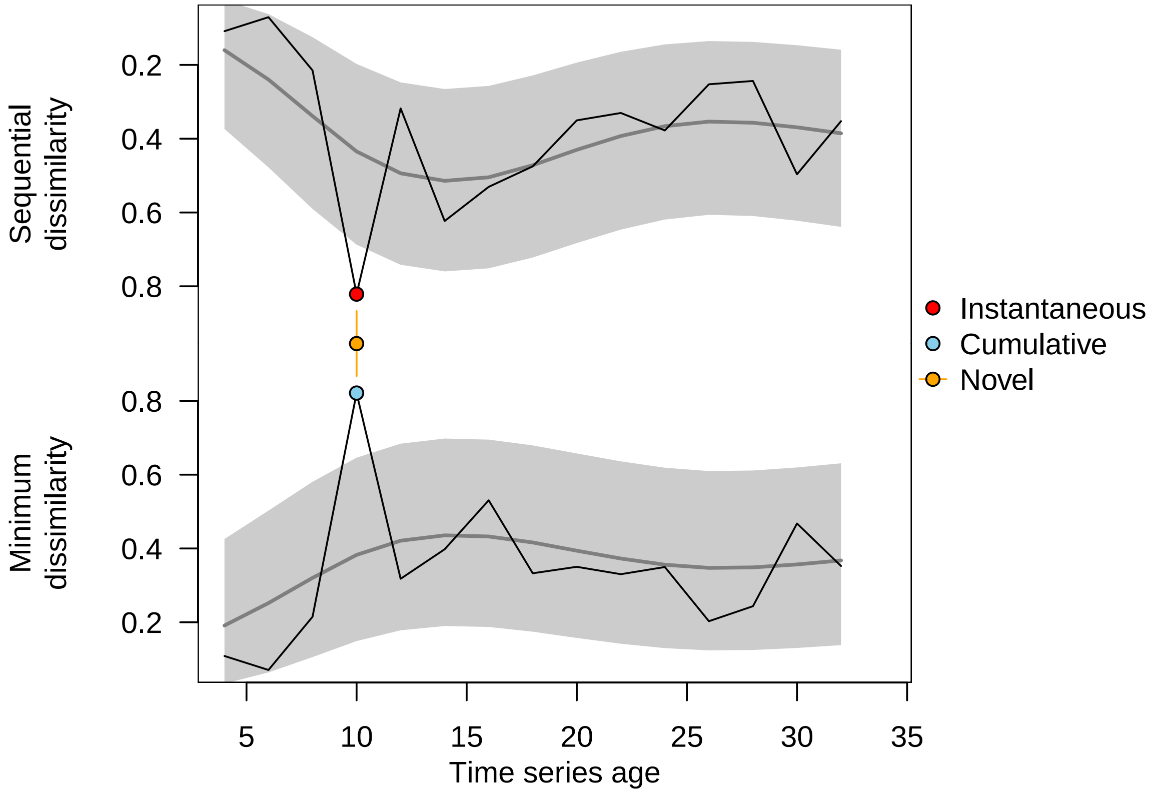

Figure 1: Example of the novelty detection framework applied to a time series. This particular time series shows the emergence of a novel state which persists for the remainder of the time series (as dissimilarities remain low following the novel state).

2. Application of Framework

2.1. Available data

We tested the framework on the RIVFishTime dataset, which consists of 11,386 time series of freshwater fish communities around the world. Sampling events were highly variable between time series; a large number of these time series could not be used. In order to qualify for the detection framework, a time series had to have 1) more than 10 unique time points and 2) more than 5 unique species. This left the following geographical distribution of time series:

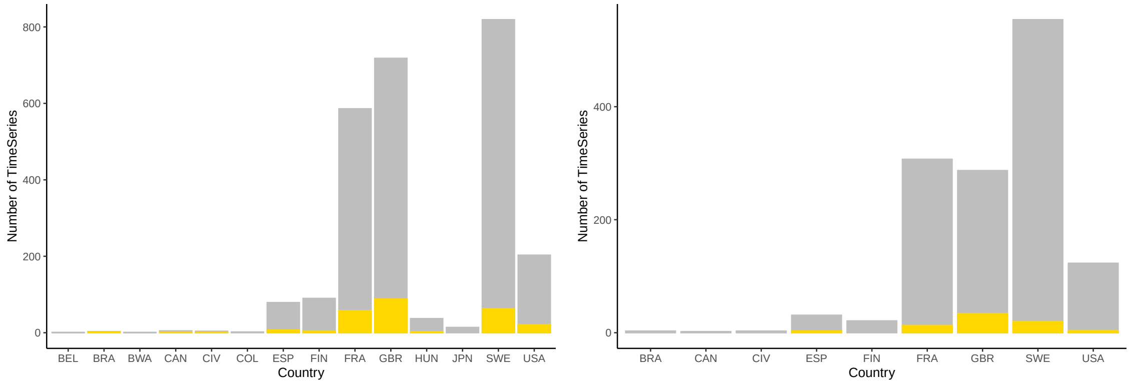

Figure 2: Geographical distribtuion of time series. Left is 1 year bins and right is 2 year bins. Gold denotes number of timeseries where novelty was detected.

The framework was applied to both presence/absence and abundance data, using the Jaccard and Bray-Curtis indexes to calculate dissimilarity scores, respectively. Raw observational data were transformed into community matrices, with rows representing the time point/bin/community and columns representing the species. An example of such a matrix is shown below.

The quantity of data constrained our possible bin widths. There were three options:

1. "Julian days" bin width. A variable bin width which essentially treated every survey in the time series as a unique time point. If there were 4 surveys in year 1 and 3 in year 2, that would amount to 7 time points. This increases the amount of usable data but is far more vulnerable to 'false-positive' novelty due to seasonal dynamics as well as population dynamics that operate over timescales longer than 1-2 years.

2. 1-year bin width. Abundance data from all the time points within a year are averaged. Usage of this bin corrects for seasonal dynamics but remains vulnerable to false positives i.e. brief population anomalies which are not indicative of systemic change.

3. 2-year bin width. Works the same as a 1-year bin but decreases usable data due to increased data requirements. Likely less vulnerable to false positives.

2.1. Results

We examined the rates of novelty using bin widths of 1 and 2 years. The tables below show the percentage of total communities that were identified as one of the three novel states. We identified two design attributes that could skew novelty estimates: 1) Bin Lag, the time gap between bin T and bin T + 1. Time series sometimes featured large, irregular gaps between sampling which could possibly inflate novelty rates, as there is more time for communities to accumulate change. 2) Bin position, which is a measure of how many bins have preceded the current one. The possible affect of this attribute is less intuitive, but may involve a higher chance of detecting novelty early in the time series, when there are less reference communities. Many time series are of varying lengths, and thus the position variable controls for increased/decreased novelty in bins with far more/far less reference bins.

Table 1: Novelty estimates over the entire dataset using absolute abundance data and bin size = 2 (estimates on logit scale)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.3262869	0.0890374	-48.589557	0.0000000
bin.lag	0.1314695	0.0513500	2.560264	0.0104593
position	-0.2966993	0.0966942	-3.068430	0.0021519

Table 2: Novelty estimates over the entire dataset using binary data and bin size = 2 (estimates on logit scale)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.9989490	0.1259018	-39.7051486	0.0000000
bin.lag	-0.0335409	0.1247609	-0.2688414	0.7880517
position	-0.4087415	0.1393212	-2.9338060	0.0033483

Using a 2-year bin, we found the rate of novelty emergence across the RIVFishTime data set to be approximately ~1.3% for abundance data and ~ 0.7% for presence-absence data, after controlling for time series artefacts.

2.2. Potential Issues and Transition modelling

One of the main questions is whether or not truly novel communities can be identified at a decadal timescale, by a framework that has thus far only been tested on paleoecological data sets that run for millions of years. Immediate issues that come to mind are:

1. Erroneous classification of a community as novel due to regular population dynamics that run over timescales of multiple years.
2. Inability to detect novelty due to slow generation times of certain freshwater species; thus true change goes undetected.

A key question was how persistent novelty was in our data. We theorized that the biggest problem in using these data to detect novelty would be the false-positives issue; sudden bursts or decreases in the populations of certain species might elevate dissimilarity above the threshold. In such cases, the novel state would be followed by an instantaneous novel state, signifying a rapid shift back to the type of community observed prior. Abundance data is more sensitive to these types of events than presence/absence data.

These anomalous events are likely to be perceived as novel by the framework. Such anomalous novel events are often characterized by a “Novel -> Instantaneous Novelty” transition, which signifies a transition from a novel state to one that is markedly different from the novel state, but closer to the states that came before. We calculated the effect sizes between the observed and expected transitions found in our data, for varying bin widths and data types. This will indicate which data minimizes the ecological anomalies. However, it must be noted that an N -> I transition does not always mean that the novel state is an artefact. Pandolfi et Al. found a N -> I transition probability that was 10 times higher than expected in their data. It is tempting to say that if Novelty does not persist for more than a year (or some arbitrary time frame), it is not ‘true novelty'. In contrast, one could also argue that novel states are inherently unstable, and more often than not, they fail to persist. This does not negate the fact that there *was* a temporary shift in community structure. Nevertheless, such changes should be distinct from novelty caused by *recurrent* population fluctuations such as those associated with spawning seasons; making these distinctions in a way that is not overly arbitrary is challenging.

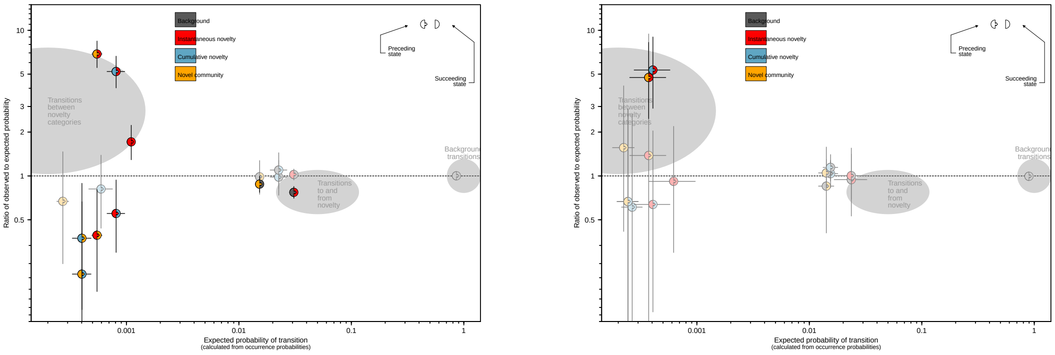

Figure 3: Observed vs. expected transition probabilities for abundance data. Left) Bin size = 1, Right) Bin size = 2

Observed N -> I transitions are approximately 7 and 5 times as high for abundance data using bin widths of 1 and 2, respectively. It is thus likely that the decrease in this particular transition when using 2-year bins is due to the removal of some population spikes (or lows). Abundances in these bins are average abundances of all surveys within that time frame, so the effect of temporary/seasonal extremes is decreased. Furthermore, there is no difference between bin widths when considering the C -> I transition. The cumulative state denotes a state significantly different from previous states, so the fact that this transition is observed 5 times more than expected might be another indicator of the tendency of novel configurations to be short-lived.

3. Analysis of potential drivers of Novelty

Now that we have established the ability of the framework to detect novelty on ecological timescales and identified potential limitations to the method, we can investigate what exactly is driving these novel communities. From the literature, contemporary novel communities are mostly caused by human-mediated exotic invasions, distribution shifts of species tracking climactic niches, and human modification of the environment (i.e. pollution, construction, obstruction etc.). The fact that novel communities are rare in this data set (~1.4 %) means that finding significant associations between novelty and community characteristics might be difficult, especially when considering the possible noise from novel communities that are in fact short term population fluctuations.

3.1 Contribution of invasives

As mentioned, exotic invasions are a likely driver of novel community emergence. We investigated this association as follows:

1. On a country level, assigned each fish species one of two categories, "Native" or "Invader". "Natives" were those species native to the country as well as invasive species that have become established prior to 1970 (the year in which most time series begin). "Invaders" are those species that are non-native to the country and became established there after 1970. We have chosen for these categories due to the novelty framework detecting novelty from a baseline (which is mostly around 1970). Thus, in

quantifying the 'effects of invaders', we need to take this perspective that the base community is 1970, and only new (exotic) species that invade after that arbitrary baseline are true invaders.

2. Computed the absolute change in abundance of the two categories between each bin in a time series, over all time series.
3. Fitted a Binomial GLMM with Logit Link function. This model includes two covariates: Bin Lag and Bin Position, as well as three predictors: Natives Relative Abundance Increase, Invaders Relative Abundance Increase and an interaction. Basin within country was set as a random effect. All fixed effects were scaled. We tested the model on both abundance and binary data to investigate how this would influence results. It is likely that the binary data will be too conservative for this model.

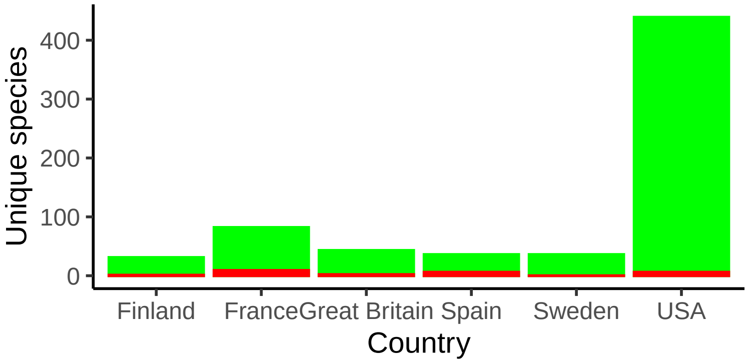

Figure 4: Bar plot showing the number of unique species (green) and the number of unique invaders (red).

3.1.1 Potential Issues

There is substantial variation between countries when it comes to the role of invaders. Sweden and Finland have experienced less exotic invasion than countries like France and the USA since 1970. In fact, the usable data from Finland and Sweden include no observations of any post 1970's invaders. Aggregating all these data into one large frame might obscure some associations.

3.1.2 Results

I have applied the following model to scaled, absolute abundance data. Binary data was insufficient to draw conclusions from and relative abundance data with just two species categories is also problematic (one goes up, the other goes down). Therefore, the use of absolute abundance data (which is more sensitive to temporary population spikes) seems justified.

```
glmer(novelty_category ~ bin_lag + position + NAC_increase * INC_increase + (1 | country/basin))
```

Table 3: GLMM output estimating effect on **True Novelty**

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.4610418	0.1408255	-31.677799	0.0000000
bin_lag	0.1397859	0.0518047	2.698326	0.0069689
position	-0.2821958	0.0994917	-2.836376	0.0045629
NAC_increase	0.2003626	0.0457362	4.380829	0.0000118
INC_increase	0.4660430	0.1475405	3.158745	0.0015845
NAC_increase:INC_increase	-0.0723273	0.0357774	-2.021595	0.0432183

Table 4: GLMM output estimating effect on **Instantaneous Novelty**

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1015185	0.0738324	-42.007576	0.0000000
bin_lag	0.1392790	0.0320538	4.345169	0.0000139
position	-0.1577394	0.0530742	-2.972052	0.0029582
NAC_increase	-0.0595866	0.0353595	-1.685165	0.0919567
INC_increase	0.2699664	0.1191335	2.266083	0.0234463
NAC_increase:INC_increase	-0.0306526	0.0304776	-1.005743	0.3145393

Table 5: GLMM output estimating effect on **Cumulative Novelty**

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.4214874	0.0736246	-46.472057	0.0000000
bin_lag	0.0951977	0.0394552	2.412803	0.0158304
position	-0.2613987	0.0605678	-4.315800	0.0000159
NAC_increase	0.2515228	0.0563882	4.460560	0.0000082
INC_increase	0.4028155	0.1453554	2.771246	0.0055842
NAC_increase:INC_increase	-0.0602114	0.0415776	-1.448170	0.1475696

There is evidence for an association between invaders and novelty, which is in itself not unremarkable, considering the number of invaders as a percentage of total species is quite low. This suggests that these invading species are at least partially associated with novelty emergence. The role of invaders is strongest in the model predicting true novelty. The slope estimate for the increase in invaders is consistently positive and higher than the estimate for native change for all models. The interaction between native and invader change is only significant in the true novelty model, but consistently negative throughout.

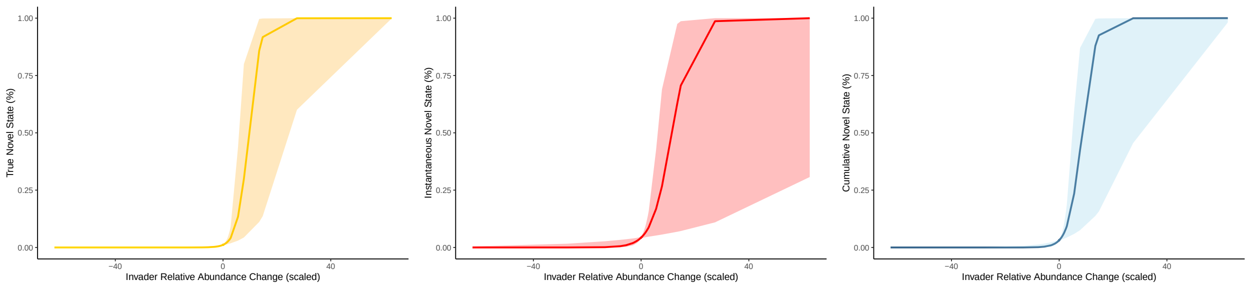

Figure 5: Effect of Invader population fluctuations on the emergence of novelty.

Appendix

A.1. Example timeseries

I have included an example of a (partial) times which shows how different data types can change novelty estimates.

Table 6: Example timeseries containing a sudden increase in *H. molitrix* which is known to form massive schools around mating seasons. Data is average absolute abundance

	Macrhybopsis	Hypophthalmichthys	Neogobius	Cyprinella	Notropis	
	storeriana	molitrix	melanostomus	spiloptera	wickliffi	Category
18	6.0	0.000000	0	0	0.00000	Back
16	0.0	0.000000	0	0	0.00000	Back
14	2.0	21875.000000	0	0	0.00000	Cumul
12	2.0	2.666667	2	4	0.00000	Back
10	2.0	2.600000	0	0	2.00000	Back
8	2.0	9486.666667	2	0	11.33333	Back
6	2.5	2.000000	2	2	2.00000	Back
4	7.0	2.000000	0	2	5.50000	Back

Table 7: Example timeseries containing a sudden increase in *H. molitrix* which is known to form massive schools around mating seasons. Data is relative abundance

	Macrhybopsis	Hypophthalmichthys	Neogobius	Cyprinella	Notropis	
	storeriana	molitrix	melanostomus	spiloptera	wickliffi	Category
18	0.0015863	0.0000000	0.0000000	0.0000000	0.0000000	Back
16	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	Back
14	0.0000769	0.8405778	0.0000000	0.0000000	0.0000000	Novel
12	0.0008146	0.0010861	0.0008146	0.0016292	0.0000000	Back
10	0.0057606	0.0074888	0.0000000	0.0000000	0.0057606	Back
8	0.0001178	0.5588973	0.0001178	0.0000000	0.0006677	Back
6	0.0105160	0.0084128	0.0084128	0.0084128	0.0084128	Back
4	0.0011126	0.0003179	0.0000000	0.0003179	0.0008742	Back

Table 8: Example timeseries containing a sudden increase in *H. molitrix* which is known to form massive schools around mating seasons. Data is presence absence

	Macrhybopsis	Hypophthalmichthys	Neogobius	Cyprinella	Notropis	
	storeriana	molitrix	melanostomus	spiloptera	wickliffi	Category
18	1	0	0	0	0	Back
16	0	0	0	0	0	Cumul
14	1	1	0	0	0	Back
12	1	1	1	1	0	Back
10	1	1	0	0	1	Back
8	1	1	1	0	1	Back
6	1	1	1	1	1	Back

	Macrhybopsis storeriana	Hypophthalmichthys molitrix	Neogobius melanostomus	Cyprinella spiloptera	Notropis wickliffi	Category
4	1	1	0	1	1	Back

From these examples time series, two things become apparent: Single species fluctuations associated with seasonal behavior (or the fact that the person doing the survey happened to stumble upon a large school by chance) have the potential to skew results. All of these matrices are from the same time series. Use of relative abundance data 'standardizes' the magnitude of a change to a value between 0 and 1, whilst preserving the ability to pick up on systemic changes in abundance, which is not possible using presence/absence data. In fact, presence/absence data completely ignores the big fluctuation in bin 14, and instead picks up on a cumulative change in bin 16 (*note, not all rows are shown*). All data types have their pro's and con's. Relative abundance and presence/absence data are more conservative and likely reduce the number of false-positives, but in doing so also largely ignore the strength of an absolute decrease/increase in population, which hinder modeling of the true effects of potential drivers.