# Data Checks & Re-Analysis

Cole B. Brookson

2022-07-26

First check is regarding Farms w/in or outside of BATI control. This is discussed in `Item 1` here

## Item 1 - Farms Used in Analysis

We want to know which farms are under BATI control. The below farms printed here are all the farms that are in the `.csv` file that Sean provided to Cole from the Nation.

```
##  [1] "Arrow Pass"      "Burdwood"        "Cecil Island"    "Cliff Bay"
##  [5] "Cypress Harbour" "Doctor Islets"   "Glacier Falls"   "Humphrey Rock"
##  [9] "Larsen Island"   "Midsummer"       "Port Elizabeth"  "Potts Bay"
## [13] "Sargeaunt Pass"  "Sir Edmund Bay"  "Swanson"         "Upper Retreat"
## [17] "Wicklow Point"
```

There are 17 farms in the BATI dataset, but there are additional farms in the DFO data (available from the government portal) that have data after 2011, and have data from pre 2011 in the Marty dataset. The farms that are in the DFO dataset from 2011 onwards are:

```
##  [1] "Arrow Pass"      "Burdwood"        "Cecil Island"    "Cliff Bay"
##  [5] "Cypress Harbour" "Doctor Islets"   "Glacier Falls"   "Humphrey Rock"
##  [9] "Larsen Island"   "Maude Island"    "Midsummer"       "Noo-la"
## [13] "Port Elizabeth"  "Potts Bay"       "Sargeaunt Pass"  "Simmonds Point"
## [17] "Sir Edmund Bay"  "Swanson"         "Tsa-ya"          "Upper Retreat"
## [21] "Wa-kwa"          "Wehlis Bay"      "Wicklow Point"
```

## Farms Missing From BATI Control

So the farms that there are DFO data for, but that are not under BATI control are:

```
## [1] "Maude Island"   "Noo-la"         "Simmonds Point" "Tsa-ya"
## [5] "Wa-kwa"         "Wehlis Bay"
```

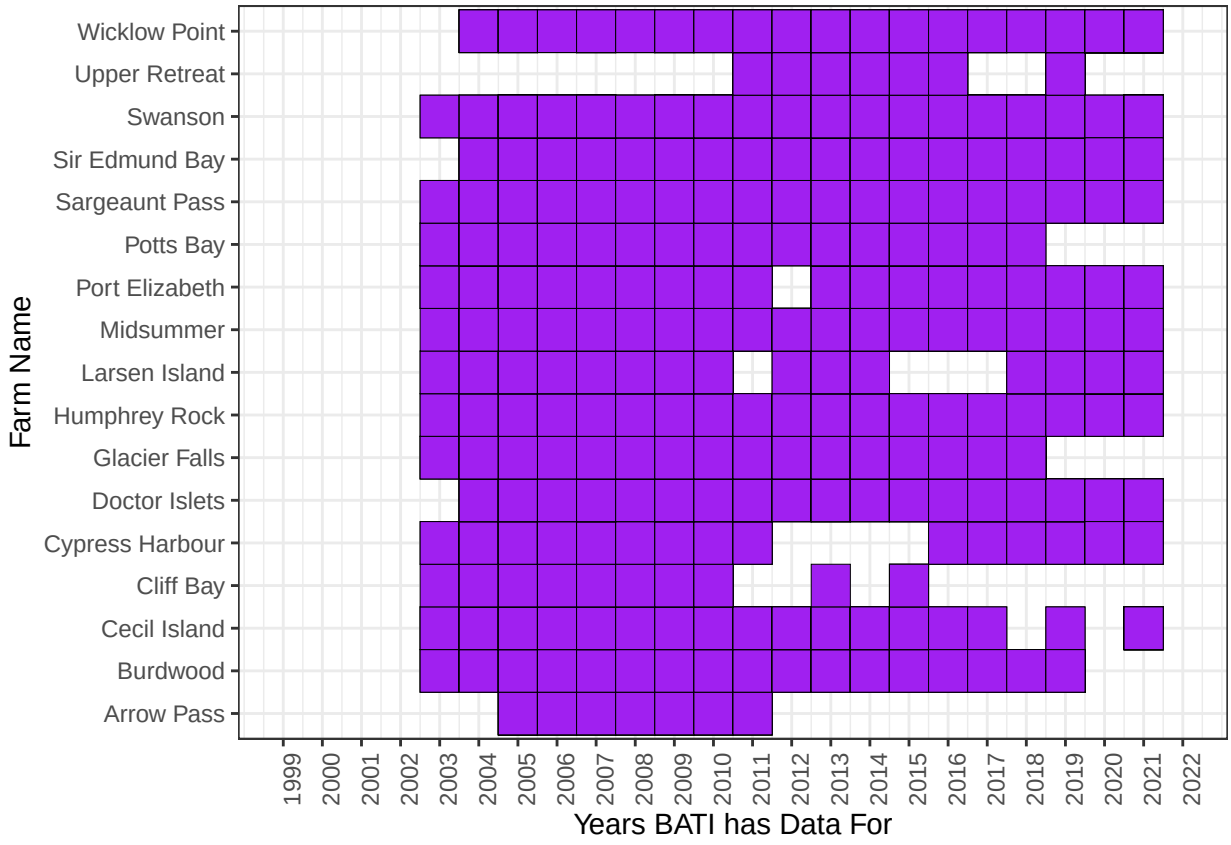
With respect the Marty (2010) dataset, most farms used in that dataset are farms that DFO has data for after 2011, but some farms don't have data past the first few years of the Marty dataset. Since nowhere in the Marty (2010) SI is there a set of common names for the farms, I named them `NA_XX` for convenience. The names of the farms in the Marty dataset are:

```
##  [1] Arrow Pass      Burdwood        Cecil Island    Cliff Bay
##  [5] Cypress Harbour Doctors Islets  Glacier Falls   Humphrey Rock
##  [9] Larson Island   Maude Island    Midsummer       NA_12
## [13] NA_14           NA_15           NA_7            Noo-la
## [17] Port Elizabeth  Potts Bay       Sargeaunt Pass  Simmonds Point
## [21] Sir Edmund Bay  Swanson         Upper Retreat   Whlis Bay
## [25] Wicklow Point
## 25 Levels: Arrow Pass Burdwood Cecil Island Cliff Bay ... Wicklow Point
```

So there are three (overlapping data sources). Their coverage is the following:

## BATI Coverage

Timeline of which farms and for which years data exists for:

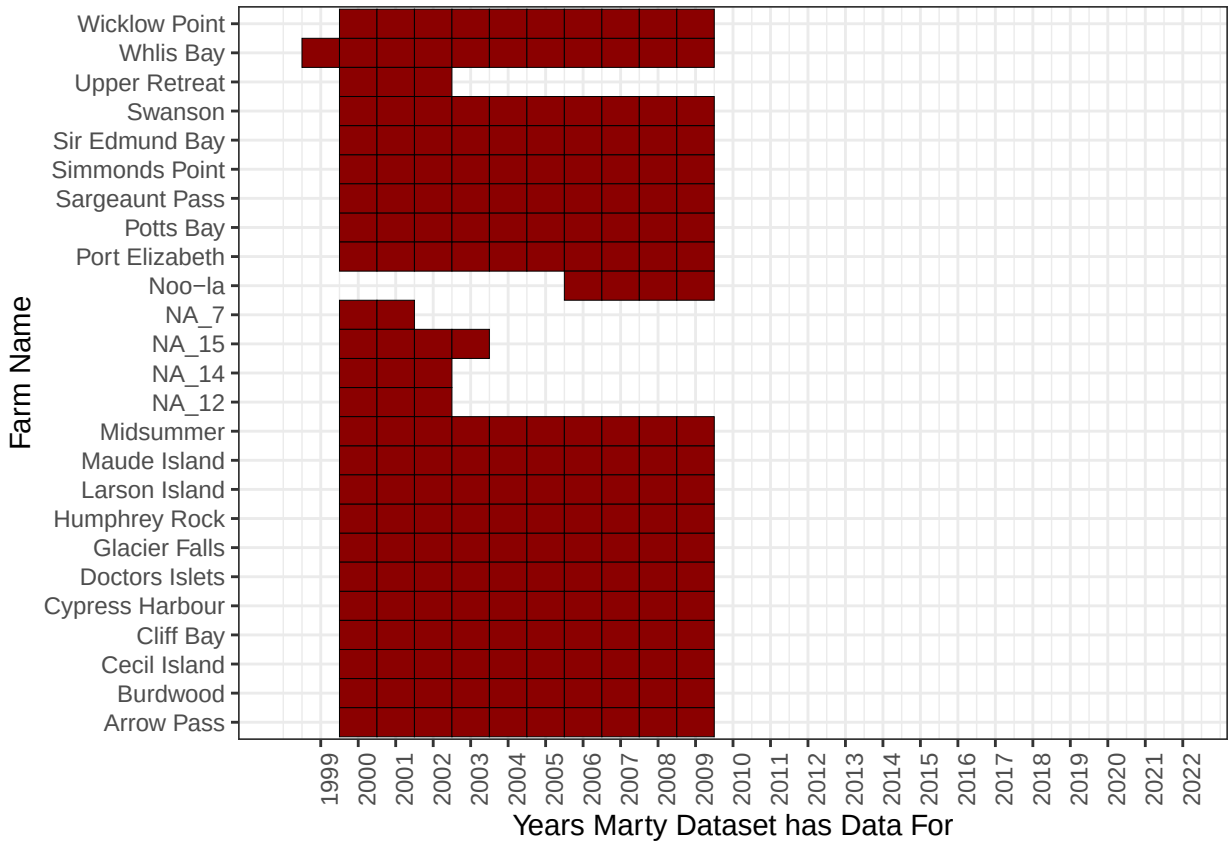



## DFO Coverage:

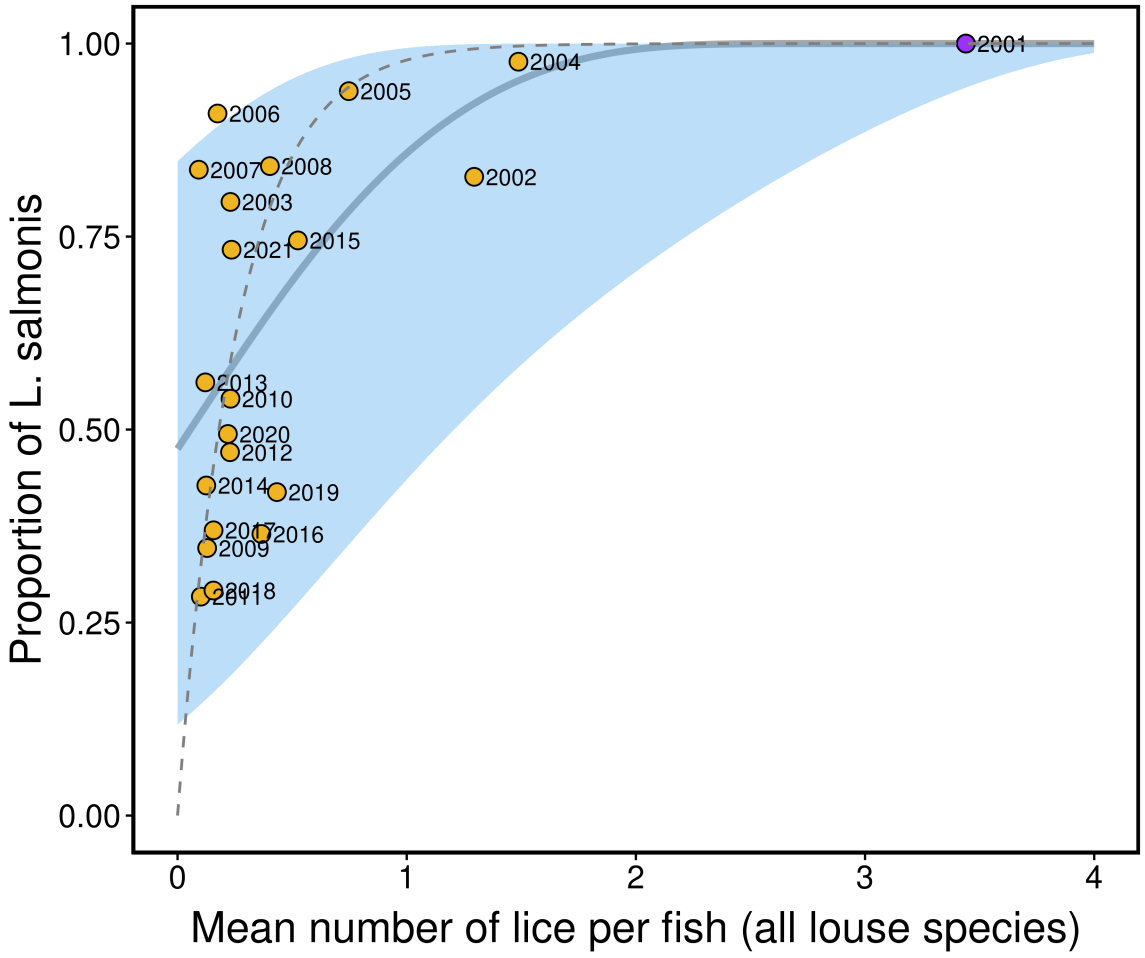
Timeline of which farms and for which years data exists for:


## Marty Coverage:

Timeline of which farms and for which years data exists for:


Wicklow Point
Whlis Bay
Upper Retreat
Swanson
Sir Edmund Bay
Simmonds Point
Sargeaunt Pass
Potts Bay
Port Elizabeth
Noo−la
NA_7
NA_15
NA_14
NA_12
Midsummer
Maude Island
Larson Island
Humphrey Rock
Glacier Falls
Doctors Islets
Cypress Harbour
Cliff Bay
Cecil Island
Burdwood
Arrow Pass
Farm Name
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Years Marty Dataset has Data For

## Figure 2 - Re-done



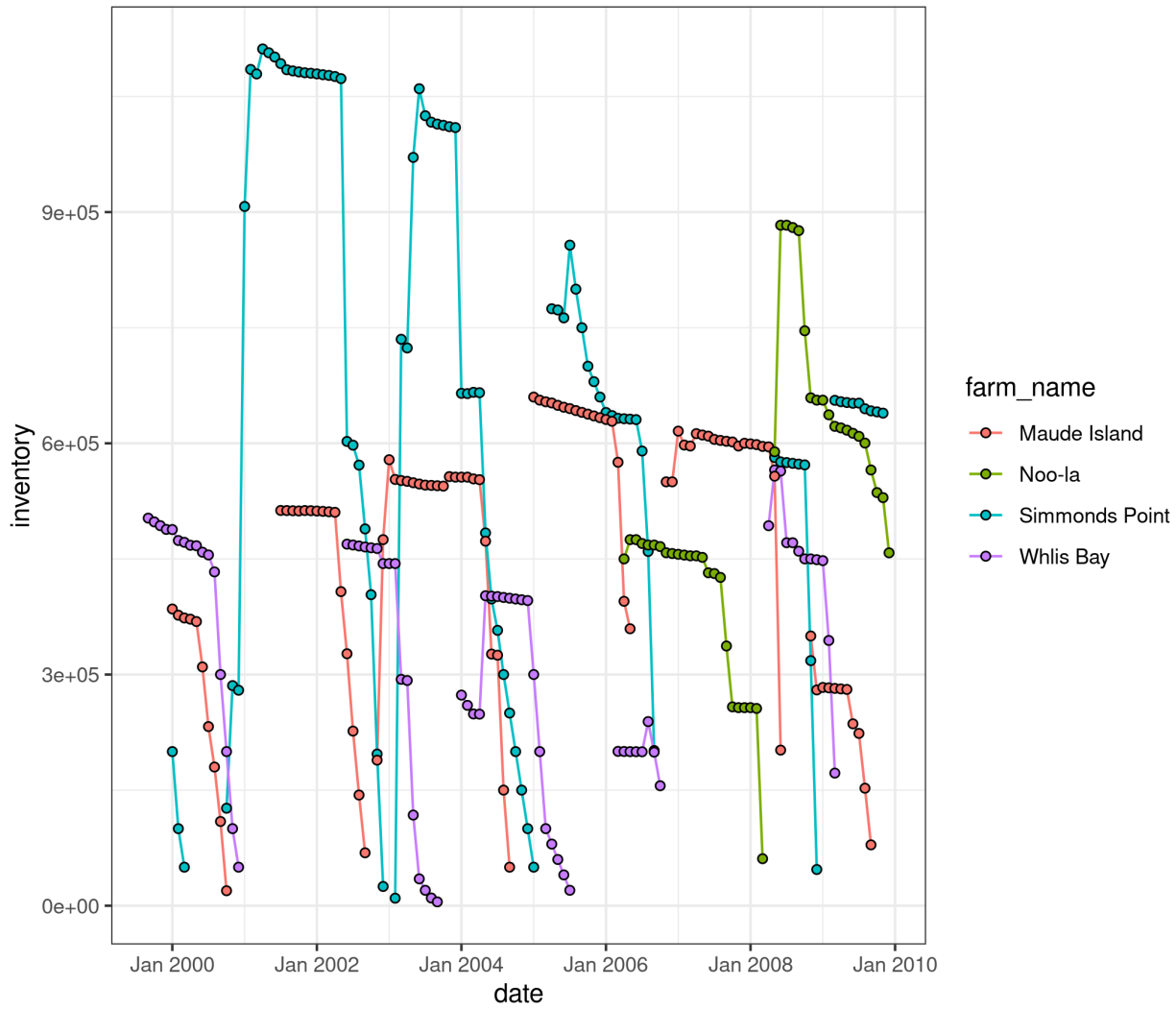

We can see that the predicted value doesn't cahgne, but it does give us a lower bound with more variance to use.

## Inventory Inputation

An important note is that while there are 6 farms outside of BATI control, only four of them exist back into the Marty dataset:

1. Maude Island
2. Noo-la
3. Simmonds Point
4. Wehlis Bay

I went through the inventory we have in the Marty dataset and looked at how it changed through the months and years. The difference between years is actually relatively substantial:

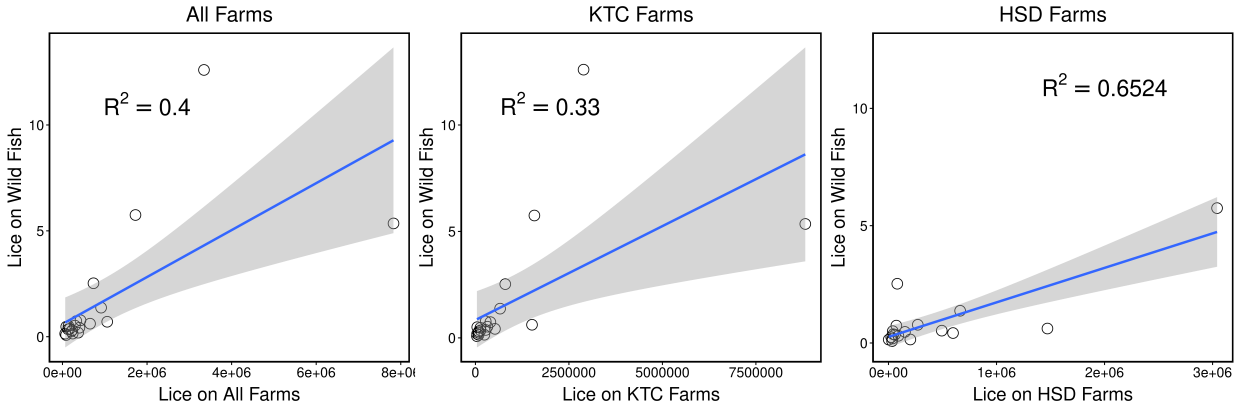



So I took the mean values for the two months of interest (March & April) and used those to move it forwards in time.

## log-scale relationship

So I redid this with a few different approaches. I started by fitting the models with GAMs. **The adjusted $R^2$ values were 0.98 for all farms, 0.901 for KTC farms, and 0.788 for HSD farms**.

I also tried looking with just a linear relationship. This plot is here:




I don't know how useful this is as it's being driven by such small values on both axes. Open to suggestions on fixing this

## Next things to do:

Re-do SR calcs and those models based on discussion today