IQuOD Duplicates – initial steps

Ed King

Ann Thresher

CSIRO

The process so far:

- Download entire WOD database
- Split it into smaller tar files and directories
- For each profile, sum the Z, T and S (if present)
- Sort by number of points in the profile
- For = number of points:
  - If T and S, find all profiles where sums are identical (Z, T **and** S)
  - If only T, find all profiles where sums are identical (Z and T)

- Manually assess outputs by plotting two profiles on same axis, and scanning metadata – Lat, Long, Date and Time
- Judge whether these might be true matches, possible matches or non-matches
- Write the first 2 to files for further assessment
- Look for patterns that will allow further automation

Results: the easy ones:

27 points:

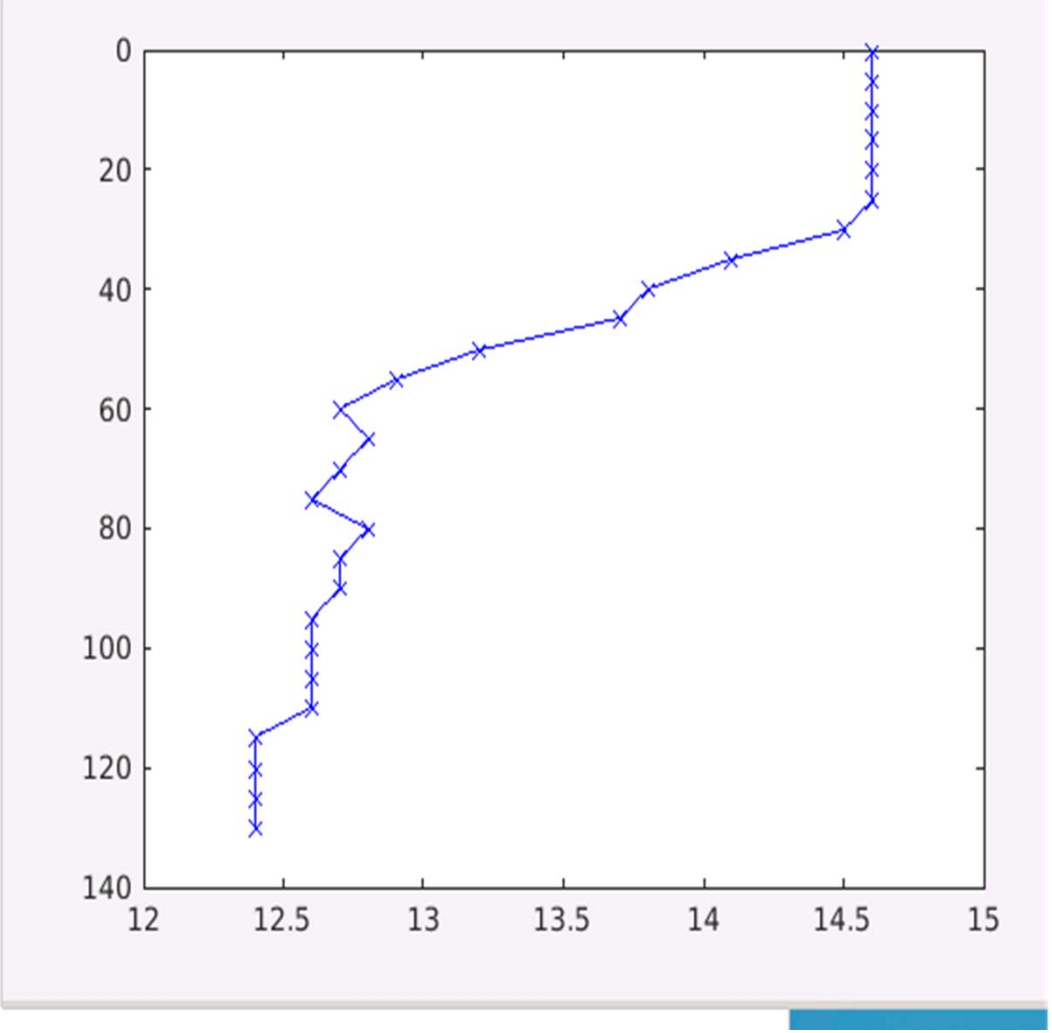
MBT/wod_0016168050.nc

- 52.5 -19.9 1965 07 25 2.0

MBT/wod_0076975200.nc

- 52.5 -19.0 1965 07 25 2.0

Clearly dupes – only keep one



Another easy one:

27 points:

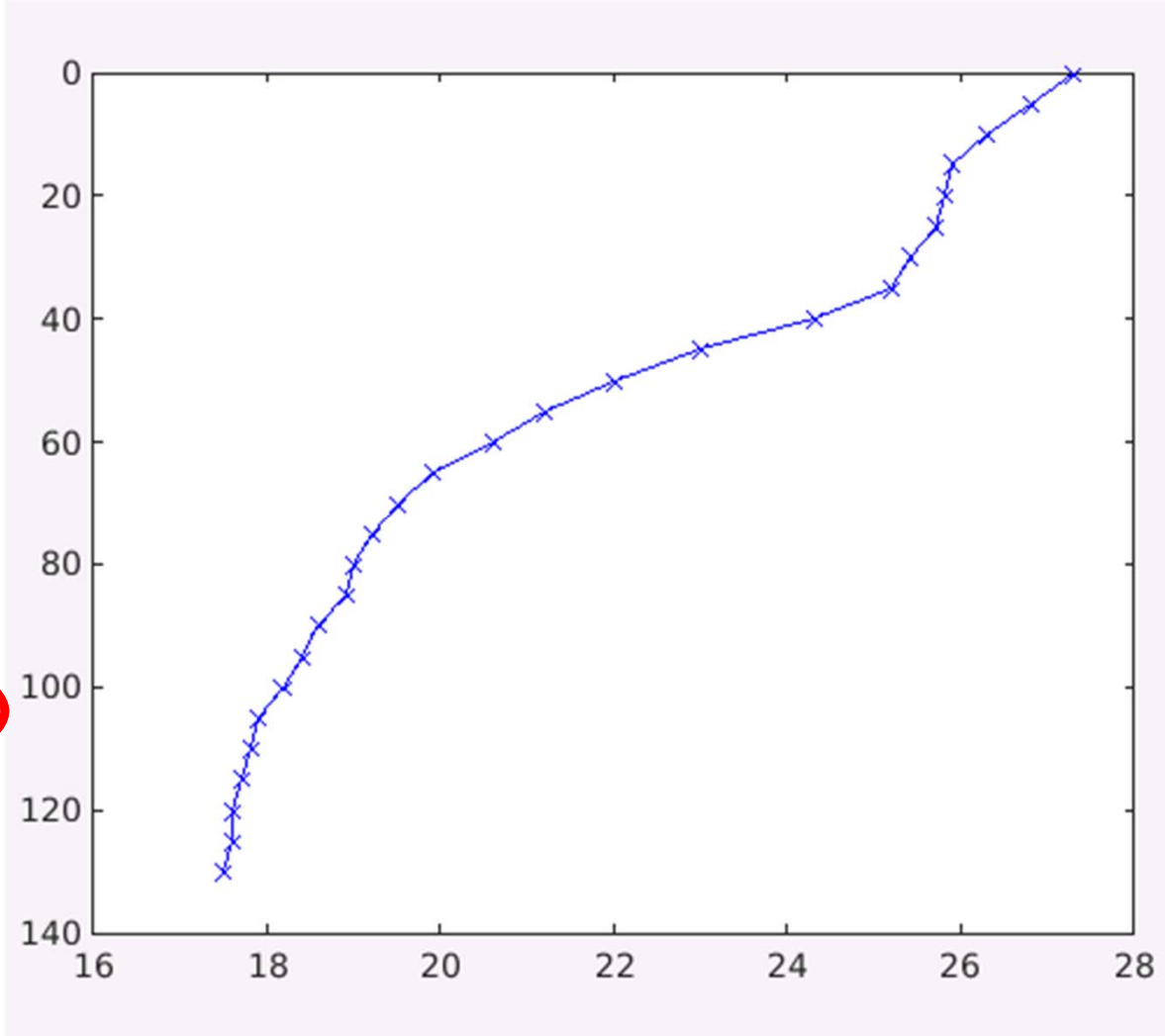
MBT/wod_0043173930.nc

- 35.8 35.4667 1948 07 03 24.0

MBT/wod_0076975200.nc

- 35.8 35.467 1948 07 03 23.9833

Clearly dupes – only keep one



Decimal Degrees vs Degrees/Minutes

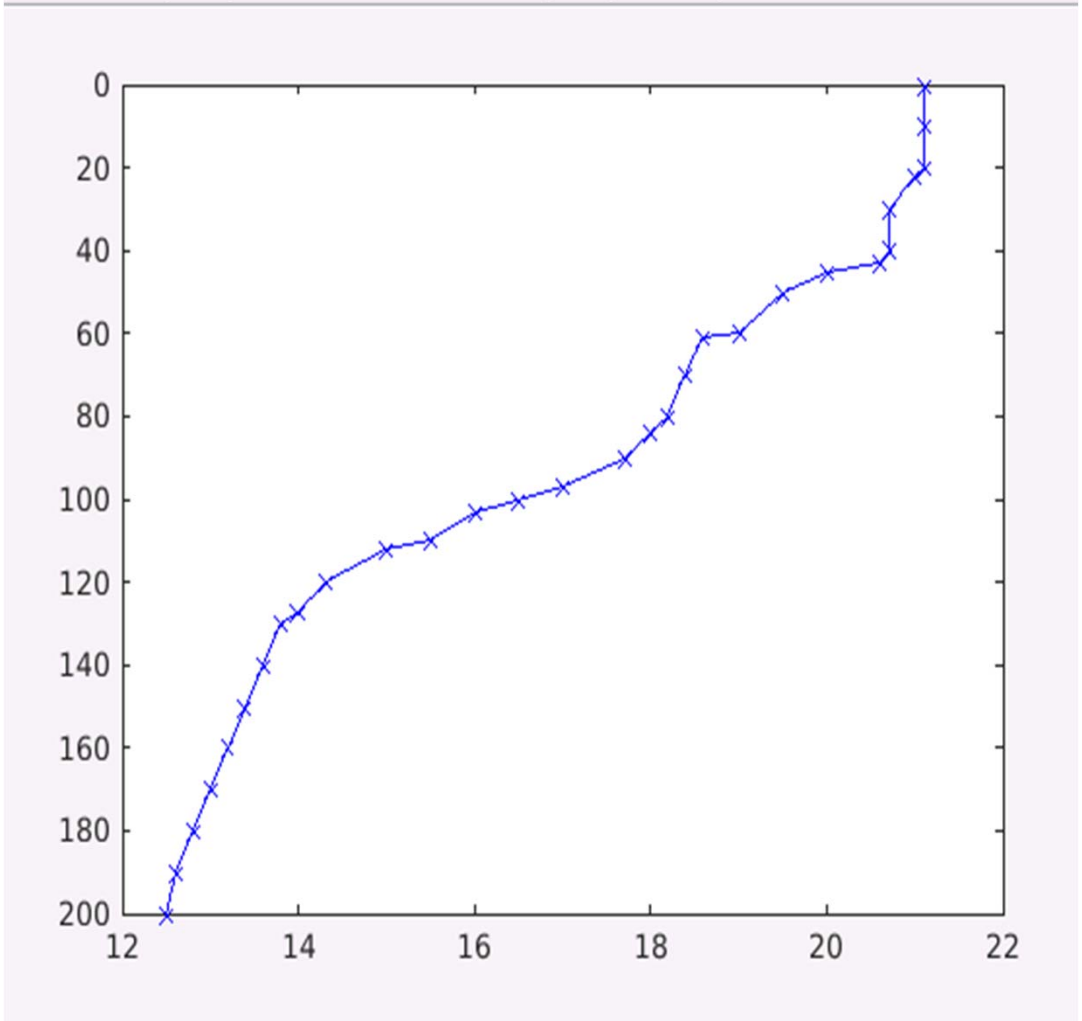
30 points:

MBT3/wod_0050375940.nc

- -13.733 -81.1 1984 05 30 15

MBT3/wod_0050375960.nc

- -13.44 -81.06 1984 05 30 15



A little harder:

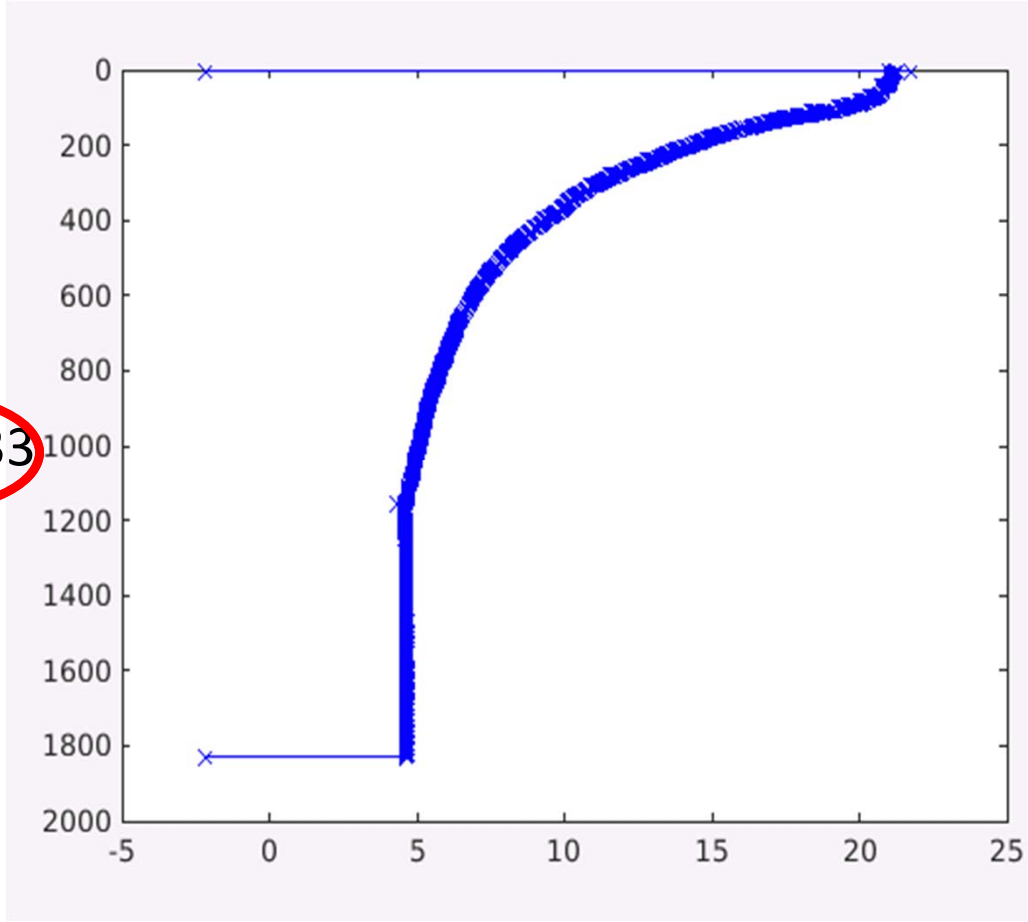
2907 points:

XBT2/wod_0039293390.nc

- 25.3167 -93.4833 1991 03 06 4.5833

XBT/wod_0039296020.nc

- 27.33 -95 1991 03 09 3.733



Clearly dupes – only keep one – but which one???

Harder:

30 points, identical location, date/time, P, T and S:

- OSD3/007887/wod_0078870290.nc
  - P T S
  - Phosphate, Silicate, Nitrate, Ammonia, Chlorophyll, NO2NO3, various weather measurements…
- CTD/003297/wod_0032972620.nc
  - P T S

Why keep the second copy? Should we have a policy of merging these sorts of duplicates?

Even Harder:

27 points:

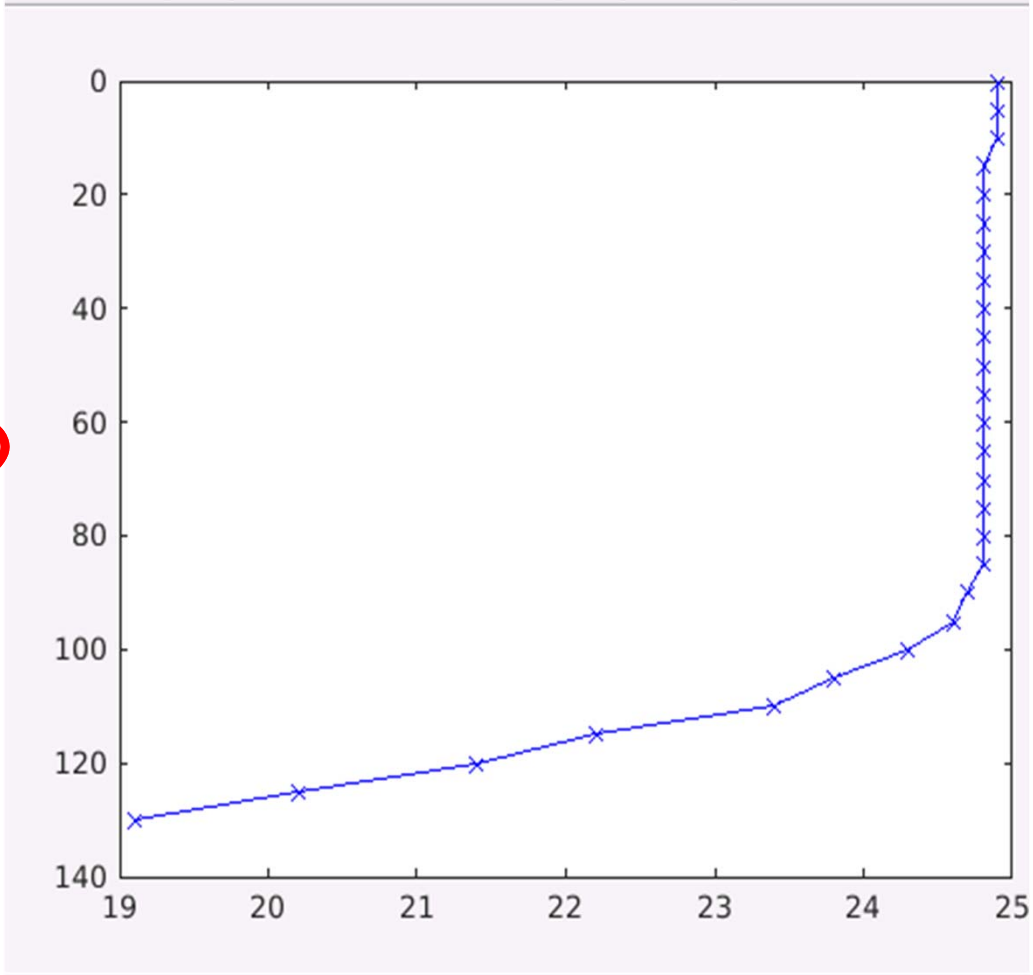
MBT/wod_0044011340.nc

- 24.3667 -81.9 1950 12 04 11.00

MBT/wod_0044902570.nc

- 24.3667 -81.75 1952 2 19 11.83

2 years apart but identical – dupes?



Nearly impossible? (there are a lot like this)

MBT/004233/wod_0042335220.nc

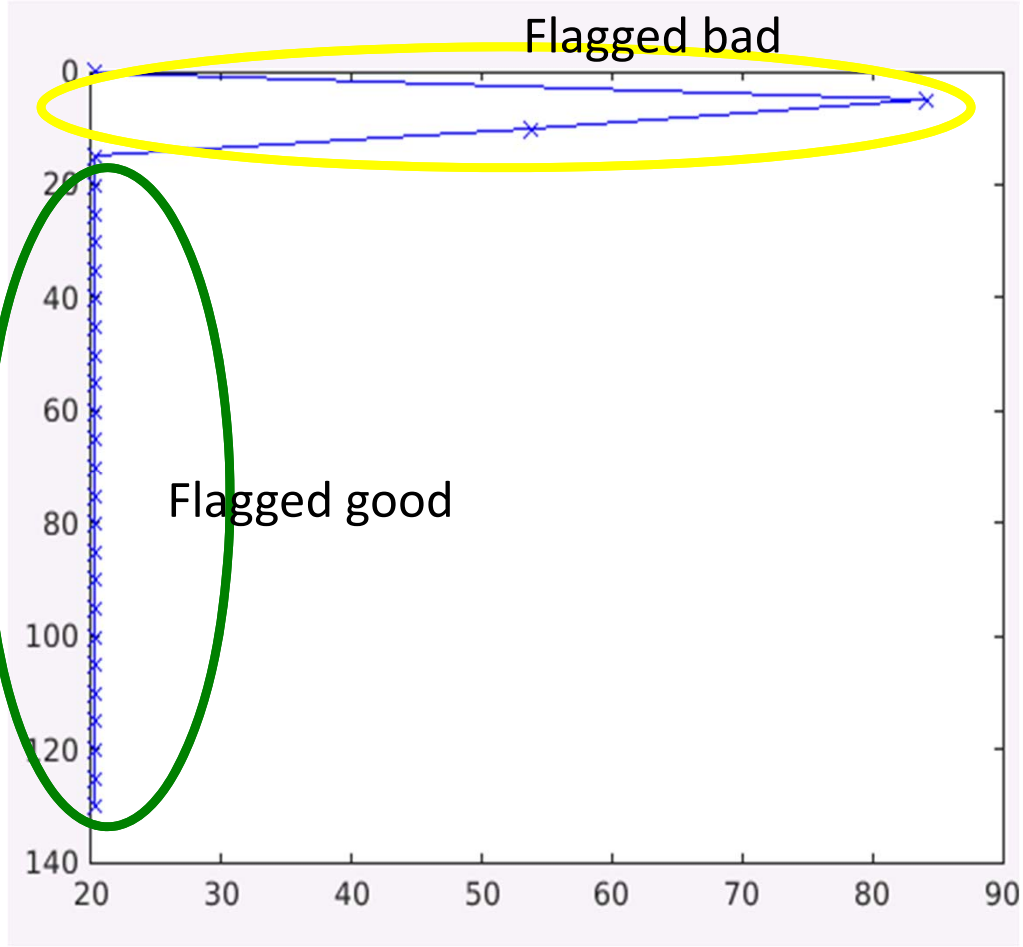
- 34.9667 -62.0333 1945 01 01 16.5

MBT/004405/wod_0044055670.nc

- 35.1 -48 1951 02 11 8

Very different metadata, probably not true dupes but there are a LOT that are physically identical to this one

QC issue: What is the probability that a profile has constant T from 0-130m in the north Atlantic? Flags indicate deeper T is good so didn't fail WOD climatology tests?



Other issues that need to be addressed

4104 profiles with more than 30 points were identical and of constant T with varying metadata. These were ignored (for now)

1747 data points in the profile:

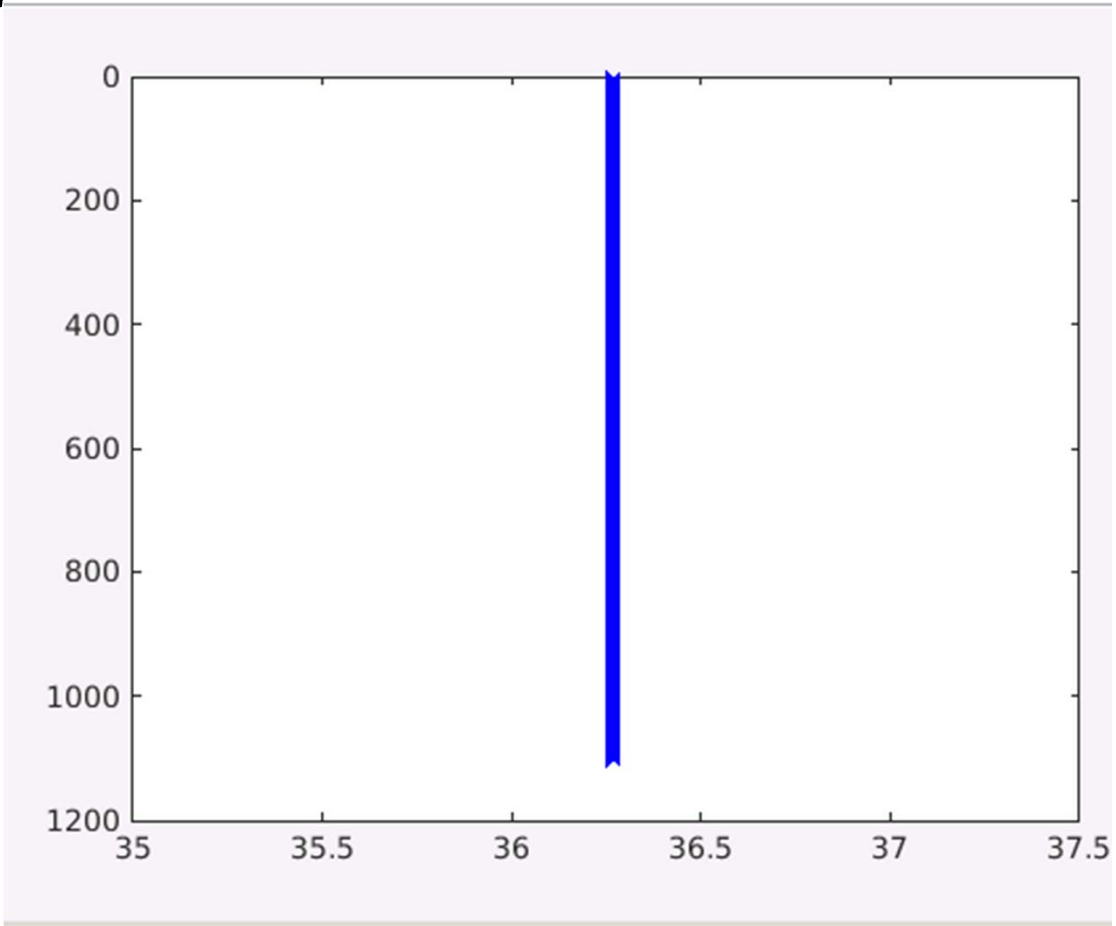
L'Astrolabe from CSIRO:

- -65.06 143.7826 2011 12 30

L'Astrolabe from France:

- -47.866 146.8819 2014 11 23

These will probably both have exact duplicates elsewhere within the 4104 profiles ; data (but not profile) is properly flagged 'bad'; does IQuOD want to include these in the database we serve?



Summary so far :

N points	N dupes	N possible dupes	N const T-dupes	N non-dupes	Total N
4	?	3939^	264,828+	?	875,125
27	208	93	2902+	169	3462
30+	3793^		4104+		7897

+ all T within a profile identical

^ not looked at individually but all points match

Next steps:

- Search for exact duplicates in space and time, adjusting for resolution
- Search for near dupes in space and time +/- 2 degrees lat/long, +/- 1 month(?) adjusting for resolution
- Ignore data type at this stage

Decision-making process:

- If date/time/lat/long identical or within 2^{o} lat/long or 3 months time:
  - If resolutions different, compare closest vertical points using interpolation
  - If resolutions identical, compare T and S if present
  - If 90% of points match, and data_type is identical, automatically declare them duplicates and then Tim will need to compare metadata to decide which is ‘best’ copy

Further rules – data_type:

- Always keep the profile of highest ‘quality’ i.e.,
  - XBT kept when it matches a BATHY
  - CTD kept when it matches a TESAC
  - If two XBTS or CTDs match, keep the higher resolution profile

- If a CTD matches a bottle cast, however, KEEP BOTH

- If we have two profiles of same type, and we are convinced they are true duplicates, can we merge metadata?

Further suggestions?

- if fall within 2^{o} lat/long or within time window
  - If resolutions different, compare closest vertical points using interpolation
  - If resolutions identical, compare T and S if present
  - If 90% of points match (within a tolerance for rounding), consider metadata (see next slide) and select ‘best copy’ to retain.