**Quality Control of the Data**

This section comes with a health warning: quality control of data is always something that always has an element of subjectivity to it. What one person may choose to include as valid data may be too messy or noisy for another person. There is no explicit “right” way to do things. What is important however, is that the quality control is *consistent* across a dataset. Otherwise, trends could appear that are actually a result of a selection criteria changing across years. With this in mind, it is important to recognise that the quality control you perform will likely be different to what someone in the future decides to perform and as such you should keep as much of the original data as possible; only changing anything that is a very clear mistake that is easy to correct. This is what Jon Shanklin has historically sought to do with the dataset; flagging the original data point, then if possible adding corrections as duplicate readings, rather than changing the original.

In this section we will discuss some of the common errors you may encounter and what to do with them. Spotting and correcting these errors is something that gets easier with experience and also relies on having some knowledge of the stations, the instruments and, to some extent, the individual observers (as some inexperienced observers may make more mistakes or the same mistakes repeatedly). Some of this information is included in the appendix and should help anyone new to the Dobson data processed by the British Antarctic Survey.

**Types of error:**

**Timing errors.** These could be as simple as an incorrect time on the logging PC after an IT update, or more complexly – irregular time offsets.

**Transcription errors.** These could be errors in recording the dial readings or the observation type. How severe the typo is will change how easy these errors are to spot.

**Operation errors.** These are the hardest to attribute a cause to, as an incorrect Q lever setting for a wavelength or not zeroing the microammeter properly during an observation is not something you can adjust for after the fact.

**Performing Quality Control**

**Step 1: Identifying possible errors**

There are broadly three things you can look at in the data to help spot anomalies. The first of these is looking at the final column ozone values:

**Examining Column Ozone Values**

Often when looking at the Column Ozone values for each observation, there can be one or two that stick out compared to the others. This is the simplest case and an example is given below:

Here we can see there are several readings that give “0” ozone, these are measurements where the sun was too low to give a meaningful value. You should also notice three anomalous points, with respect to their closest neighbours. These points have been highlighted in the appendix, so you can try to spot them on your own first.

For most observations, you’ll have pairs of readings, an AD and a CD. If one of these ob types changes with regard to the other, it either could be an indication of an incorrect observation or an instrumental change. If there is no significant change in the standard lamp test results, then it is likely a human cause. Otherwise, it is worth examining the observation more closely.

When plotting the results, try to use a different colour coding for AD and CD results, as this can make it easier to see if the relative difference between the two is changing. A hypothetical example of the usefulness of this technique can be seen below.

In the example above I have swapped an AD and a CD point on the 15th December. It is much clearer to see when AD or CD points are incorrectly attributed using this technique, it can also be seen if the AD and CD points are becoming closer or further apart, which could indicate a change in the instrument.

**Examining Direct Sun Obs**

It is also useful to plot the data by observation type. By plotting each type on their own, we can see trends that would not be apparent if we lump all the data together. For example, we could have a Direct Sun observation that presents a plausible ozone value, but it doesn’t fit the line of the of the other DS values from that day. In this case it may be worth seeing if swapping it with the nearest Zenith observation makes both obs fit better with the day’s data.

In the above example (from real data) it is relatively clear that CDDS gives the highest values, ADDS intermediate values and Zenith (in general) the lowest of the three. There is one clear outlier (the first CDDS), that has seemingly been retried 5 mins afterwards, producing a more realistic value.

**Examining Zenith Obs**

This is very similar to the process for Direct Sun obs but has the advantage that you should have more Zenith obs to compare to, so should also be able to compare over a longer timescale and notice if there are any obvious step changes.

Above is a graph of N value over three days in December 2014, which should present a fairly regular pattern. The other wavelengths should form a similar pattern of “U” shaped curves, but offset slightly from the others:

Just looking at a single wavelength should be a smoothish curve over three days, provided ozone is changing slowly. In the example above, there seems to be a step change in the afternoon of the 2nd December, is this an erroneous measurement? Looking at the Synops from the time, it seems like it is merely a change in the weather, as we go from 8 oktas to 3 oktas of cloud and from heavy blowing snow to light/mod[2]:

89022 41630 80929 11028 21047 39604 49642 52013 73973 88079 333 11028 21052 86348 8/368 92952 96014 91146=

89022 41728 31031 11024 21043 39614 49652 52010 73833 81031 333 81356 83076 92941=

**Step 2: Correction of observations**

Correcting an observation should only be attempted if there is a clear explanation for what has gone wrong. Here are some ways of working out what may have happened to an anomalous observation. If correction is possible, the original should be flagged with a “9” in front of its ob type, giving it a three digit number instead of the usual 2. Then the corrected observation can be added as a duplicate.

**Timing errors:** These can be regular (e.g. all obs offset by an hour), which is often due to an incorrect time on the logging PC, or irregular (a non-fixed offset). This was a real case at Vernadsky of an irregular error and was caused by an interruption in the usual process:

Enter the time -> enter the observation -> submit the observation.

Which became:

Enter the observation -> submit the observation. Enter the time->

After an observer entered the time twice (at the start and end) during one observation. This was only spotted by noticing that there were unusual values of ozone reported and then this was traced back to two obs being recorded as very close to one another. All incorrect times will result in an incorrect value of mu being used, which will affect the column ozone value.

**Transcription errors:**  If two adjacent observations of different types seem anomalous, then it may be worth attempting to swap the obtypes to see if this resolves the problem. Further evidence for doing this can come from a break in a pattern of observations. E.g. if obs appear in the pattern Z,DS, Z,DS, DS, Z then the last two obs could have been coded incorrectly, particularly if they are from the same observer.

Another type of miscoding is numbers (particularly 1’s and 0’s) being swapped around. This can happen whenever or wherever an automatic encoder was not used, particularly at Vernadsky (see Appendix 1). If an ozone value from Vernadsky appear very strange, but look normal when a 1 and 0 are swapped, this is a likely issue.

**Operation errors:** These are difficult to adjust after the fact, so often the anomalous observation can only be flagged and left at that. In the best case, you could change the obtype if only one wavelength is affected, e.g. changing a CDA ob into an AD if C gives strange results. A real example of this is shown below, where the C wavelength is deemed anomalous. The observation is first duplicated, so the correction is next to an original. Then the original is flagged with a “9” in front of its original code. Then the second observation has a third, fictional A observation added, and made to be the average of the original two. This is purely because AD obs have three A dial readings and two D readings. Finally the ob type of the corrected ob is made into an AD only [1].

**Original ob:**

**2020 01 19 09 31 21 95 031E 1859 1853 1047 1046 564 560**

**Flagged and corrected ob:**

**2020 01 19 09 31 21 995 031E 1859 1853 1047 1046 564 560**

**2020 01 19 09 31 21 5 031E 1859 1853 1856 564 560 0**

This type of error can happen for any of the three wavelengths, but is most common for the C wavelengths, as there is no stopper on the Q levers for C, which makes human error more likely.

As a final note on operation errors, Jon does an ozone calculation for the C wavelength by itself for Zenith Blue calculations. If the C is different to AD and CD, that almost certainly implies there's a thin layer of cirrus around, these can then be recoded as a Zenith Cloudy observation.

**Notes:**

Could fabricate a dataset where all the times are offset by 1 hour and therefore all the mu values are wrong as an example?

Graph up the day's readings and look at the curve of N values vs SZA. Look for inconsistencies. Try with vernadksy data – this is more useful here.

All of Jon’s quality controlled data can be found at Z:\cmet\OZONE\data, where Z is mapped to the cmet drive on samba.

**Suggested work:**

It would be useful to have tools that produce graphs for spotting errors automatically. Or tools that produce a list of suspect obs. Either of these would be a good programming challenge and a way to get acquainted with ozone data.

References

[1] The obtypes can be found, alongside other useful metadata about the BAS Dobson record at: <https://legacy.bas.ac.uk/met/jds/ozone/descrip.html> . There are plans to move these pages in the future, so if the link does not work, please contact Stephen Lloyd, Jon Shanklin or Steve Colwell.

[2] See the excellent quick guide to Synops by John Law and Carolyn Graves for more details on these codes. “SYNOP CODE JL\_CG.docx” (Contact Mairi Simms or Steve Colwell if you can’t find this document).

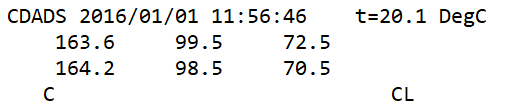
**Appendix 1: Information about the stations and instruments**

Vernadsky (<https://en.wikipedia.org/wiki/Vernadsky_Research_Base>) is a Ukrainian research station at -65.246, -64.258. Formerly known as “Faraday” and run by the British Antarctic Survey, before being sold to the Ukrainians for £1. The Ukrainians agreed to continue the long term record of Dobson measurements there and we continue to process the data that comes from there. As it is a coastal station and much further north than Halley, the weather conditions and the albedo is much more variable than at Halley. Furthermore, the station tends to be on the edge of the Ozone Hole, rather than directly underneath as Halley is, which can lead to column ozone values being much less consistent over days or weeks as bands of low ozone can sweep over the station over the course of a day or even (in extreme cases) a few hours. Finally, as the station is closer to the equator than Halley, Mu values change more quickly over the course of an observation, which could lead to greater errors if the time is incorrect.

At the time of writing (2021), Vernadsky has the Dobson #123 stationed there. Whilst I have never been there, or seen pictures of it, it is my understanding that the Dobson is very close to the ceiling and hence the dial cannot be read in the normal fashion. Instead, there is a mirror behind the dial, from which the readings can be read. The problem being, that mirrors produce a reverse image – so transcription errors are common. Jon has noted that in particular, ones and zeroes being swapped around is commonplace.

As stated above in the section on Vernadsky, Halley is generally more stable than Vernadsky. Small patches of cloud that can impact single observations are fairly rare, and more consistent blankets of stratocumulus/altocumulus/altostratus/cirrus are more common. Halley is normally under the Ozone Hole and as such, changes in column ozone are generally slow.

At the time of writing, Halley has Dobsons #073 (known as Daisy or the Autodobson) and #031 (Daphne/manual Dobson). The autodobson only produces zenith obs and needs less QC than the manual, however it is still important to check it regularly, particularly in the summer season as it may not have been switched back on or be left incorrectly set up (cap over the window, wrong voltage setting etc) after some maintenance or test. The manual Dobson has an encoder on the R-dial, which means transcription errors are rare, but not impossible. What is more likely is a miscoded observation (Direct Sun marked as a Zenith ob or vice versa). In the Halley files on the data drive, there is a text file that includes information such as any present weather and the observer’s initials e.g.:



This can sometimes be useful for diagnosing any observers who are making the same mistake repeatedly or if there is some inclement weather causing an unreliable ob.

**Appendix 2: Some examples of QC**