DATA_SILSO_HISTO
Quality Control Report
2^{nd} draft

Stephen Fay

August 15, 2019

Contents

1 Introduction

1.1 Github repository and project

https://github.com/dcxSt/DATA_SILSO_HISTO_search
https://github.com/users/dcxSt/projects/2?fullscreen=true

1.2 Brief History et Mise en Contexte

For centuries we have observed the sun and it's ever mysterious sunspots. The 11 year sunspot cycle has long been a subject of debate. Today we wish to have precise quantification of solar activity throughout the previous centuries. This is made possible by the sunspot series. Since the invention of the telescope in the early XVIIth people all over the Eurasian continent have been recording the number of sunspots that appear on the sun's earth facing half.

The aim of this project is to do a quality control of the data in `DATA_SILSO_HISTO` - recently digitized data from the Mittheilungen journals - to identify and correct things that are wrong with the data.

1.3 What do the flags mean?

During the quality control I identified several types of problems with the data that could not be corrected immediately. I added a flags column to the data-table and classified the misbehaving data in accordance to the scheme below:

0 same as Null	1 suspicious	2 Comment in journal = ? uncertain / bad def sun	3 2^{nd} instrument	4 groups > sunspots
5 v. high sunspots	6 misc see comment	7 derived from area-measurements	8	9 null s-spts / grps

Flags key table

0. The default for the flag is NULL, when is estimate that the data-point is perfect and there is nothing wrong with it, I can put it to zero 0.

1. The default flag for fishy looking data. Most of those flagged 1 belong to the category of data-point where the real observer is mentioned in the comment.

2. If in the Mitteilungen journals there is written a '?' next to one of the data points, I will mark it with a 2, this means that the observer is not quite confident in his/her result. See 6.10 - July 3 for speculation on what I think comment '?' means. Under this flag I have also groups the comments labeled 'bad definition of sun-picture' - paraphrasing from German.

3. **New meaning:** secondary telescope / observer commented, specifically this is for those observers who do not take many measurements with their secondary instruments. Sometimes a family member (usually wife) makes a few observations, but not many, these will be flagged with a 3 also. For where it is not realistic to make a new alias out of them... (**Old meaning:** A flag that signifies that this data point is definitely going into the bin ; I used this until the 2^{nd} of august, then I checked that nothing in the databases was flagged with a 3 and changed the meaning)

4. **New meaning:** Data where the groups number is bigger than the sunspots number, and the numbers are not area measurements. (GROUPS > SUNSPOTS) (**Old meaning:** For data that is very dodgy but it is ambiguous as to weather or not it is correct, to determine its validity closer examination is required)

5. **New meaning:** Data where the sunspots number is unusually high, very extremely high - I recon $\frac{1}{4}$ of these data-points are erroneous (very rough estimate). (**Old meaning:** For data that is dodgy, the difference between 5 and 4 is illustrated by example: if i find that a data-point has a groups number of 30 I will mark it with a 4 and comment it, because this is suspicious, if a data-point has a groups number over 60 or above, it will be marked with a 5 (trust me there are some in the hundreds). When it comes to sunspots it's the same but with 100 for 4 and 250 for 5)

6. Miscellaneous data, take a look at the comment, often the comments here will be what is written in the Mittheilungen.

7. Data who's values have been derived from some formulae, usually because observer noted down area measurements of the total number of millimeters the sun-disk was taken up by sunspots. I use flag=7 for both the original un-derived area measurements data as well as the derived data. You can easily tell them appart because the un-derived measurements have different observer aliases than the derived ones, for instance: 'Carrington derived' and 'Carrington area'.

8. (**Old meaning:** Bad definition of the sun picture / the sun was not clear / no sharp image of the sun, perhaps due to cirrus cloud or something... - *this meaning was made redundant because flag 2 means the same thing*)

9. SUNSPOTS IS NULL $\vee$ GROUPS IS NULL - the data is missing in one of these two columns - most of these are copied correctly into the database; often the observer noted the groups number but not the sunspots.

1.4 Equations

Derivation of wolf number from area measurements

$$r' = a \cdot (10g + b \cdot f) = 10a \cdot g + c \cdot f \tag{1}$$

Our model is linear, with Gaussian error. Thus the probability P of the wolf number r being associated to the number of groups g and the total area of flair f is

$$P(a,b) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \cdot \left[\frac{r - a(10g + b \cdot f)}{\sigma}\right]^2\right) \tag{2}$$

So the probability of us obtaining the results that we do is

$$P(a,b) = \prod P_i = \prod \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2} \sum \left[\frac{r_i - a(10g_i + b \cdot f_i)}{\sigma_i}\right]^2\right) \tag{3}$$

We define chi-squared in the following manner

$$\chi^2 = \sum \left[\frac{r_i - a(10g_i + b \cdot f_i)}{\sigma_i} \right]^2 = \sum \left[\frac{r_i^2 + a^2(10g_i + b \cdot f_i)^2 - 2r_i \cdot a(10g_i + b \cdot f_i)}{\sigma_i^2} \right] \tag{4}$$

Chi-squared test. We want find the parameters a, b such that the probability of us obtaining the results we did is maximised. This is equivalent to finding the parameters a, b such that χ^2 is minimised. For this derivation we assume the standard deviation is uniform, that $\sigma = \sigma_i = \sigma_j \ \forall i, j$

$$\frac{\partial^2}{\partial a^2}\chi^2 = \frac{\partial^2}{\partial b^2}\chi^2 = 0 \tag{5}$$

Standard deviation formula

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad\qquad var = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{6}$$

Since we have models where $\bar{x}$ is not the mean but a linear model, the standard deviation given as a percentage of the value x

$$\sigma\% = 100 \cdot \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\bar{x}}\right)^2}{n-1}} \tag{7}$$

1.5 Python scripts - what they contain

See the README it is automatically generated based on what are in the scripts. Generated by the file `create_readme.py`. Below there is a whole section devoted to explaining graphs I made in these files, see section 5.

2 Modifications made to the database

2.1 Duplicated data

This is the data for which a single observer has more than one data-point associated with him on a single day, there were roughly 14000 of these. A fuller description of the types of duplicated data can be found in the duplicates folder - in particular this file. The following is an outline of the types of duplicate data and what I did with them. (see June 25++ in log)

Identical There were some identical data point, that is to say that the fields {groups, sunspots, date, wolf, comment, fk_rubrics, fk_observers} were exactly the same for several data-pts. Here I deleted all but one.

!Year There were entire rubrics where all the fields were entered in correctly except for the year, these ones were easy to correct because the rubrics id gave them away.

!Obs Like with the incorrect year entire rubrics where entered in correctly except for the observer, they were also corrected.

2^{nd} **Obs** There were cases (notably Brunner and his assistant), where several observers share the same observation table and some of their observation days overlap. These observers where separated out.

2.2 Sorting comments

I arranged all the comments into sections according to which rubrics they appeared in and sorted them into different types with the following colour code. Some comments are associated with several of the following colours. Originally there were 8043 commented data-points, (there are now 34875 but this part is concerned with the original comments).

Pink Several comments which mean the same thing. For instance in rubrics 757 there is 1 data-pt with comment '* = Wolfer P' and 30 data-pts with comment '*'. Nothing else is commented in this rubrics. For these types of comment. Having checked the Mittheilungen to confirm my suspicion I changed the short version of each comment and turned it into the long version, in our example we end up with 31 data-pts with comment '* = Wolfer P'.

Orange Commented observer / secondary telescope. Many of the data-pts including the one in the previous example have the real observer commented. In the Mittheilungen journals there are many rubrics where most of the observations are made by one observer, but some are made by another. These rubrics were often typed in under the primary observer only, but with the name of the secondary observer(s) / telescope(s) commented in where indicated in the journals. Often the comments were a single letter. I modified all these comments so that each comment clearly listed either the name of the secondary observer / instrument where indicated.

Red There is one type of comment that appears many times in different forms, these are the comments which turned into the flag 2 (see 1.3). The data-points in question are the ones where the comment looks like one of these {uncertain, Uncertain, ?, mauvaise def img sol, mauvaise definition image du soleil, bad quality sun image}

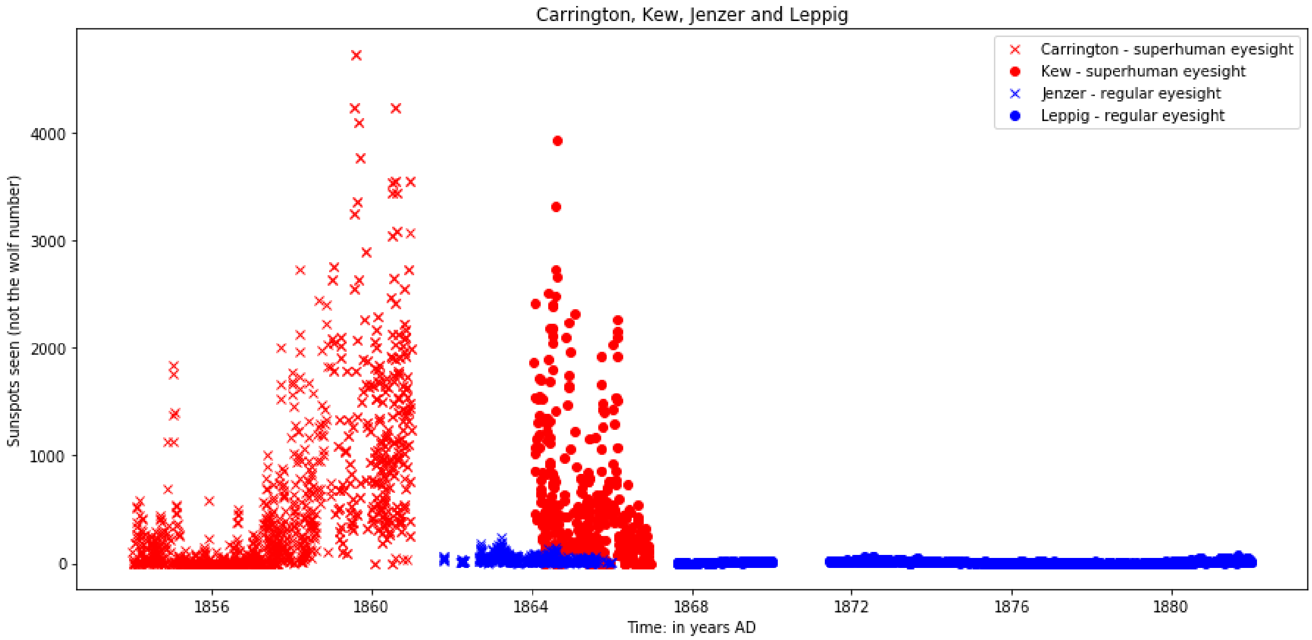
Blue Number comments are other strange comments. These are the data-pts where the comment may look like one of these {0.3 , 0.5 , 2.5 , 1 1 , 14 2 9 , img. 20 cm diameter , 9.5 cm image, derived 29, derived 11}. The numbers sometimes made reference to a secondary observer who observed on the same day as the primary observer of a rubric; in Secchi's case (derived n) the comment indicated the sunspot number derived from his umbra surface area measurements. I modified the data where appropriate.

Other There are also those comments I could not classify easily, these ones I looked up individually (as with the blues) and dealt with them on a case by case basis. 2.3 ; 2.3

2.3 Derivation of wolf number from umbra area measurements and other derived data

By plotting certain observers you can see that there is clearly something wrong with some of the data.

Carrington and Kew umbra area measurements in sunspots field (see 2.3.3)



2.3.1 Secchi

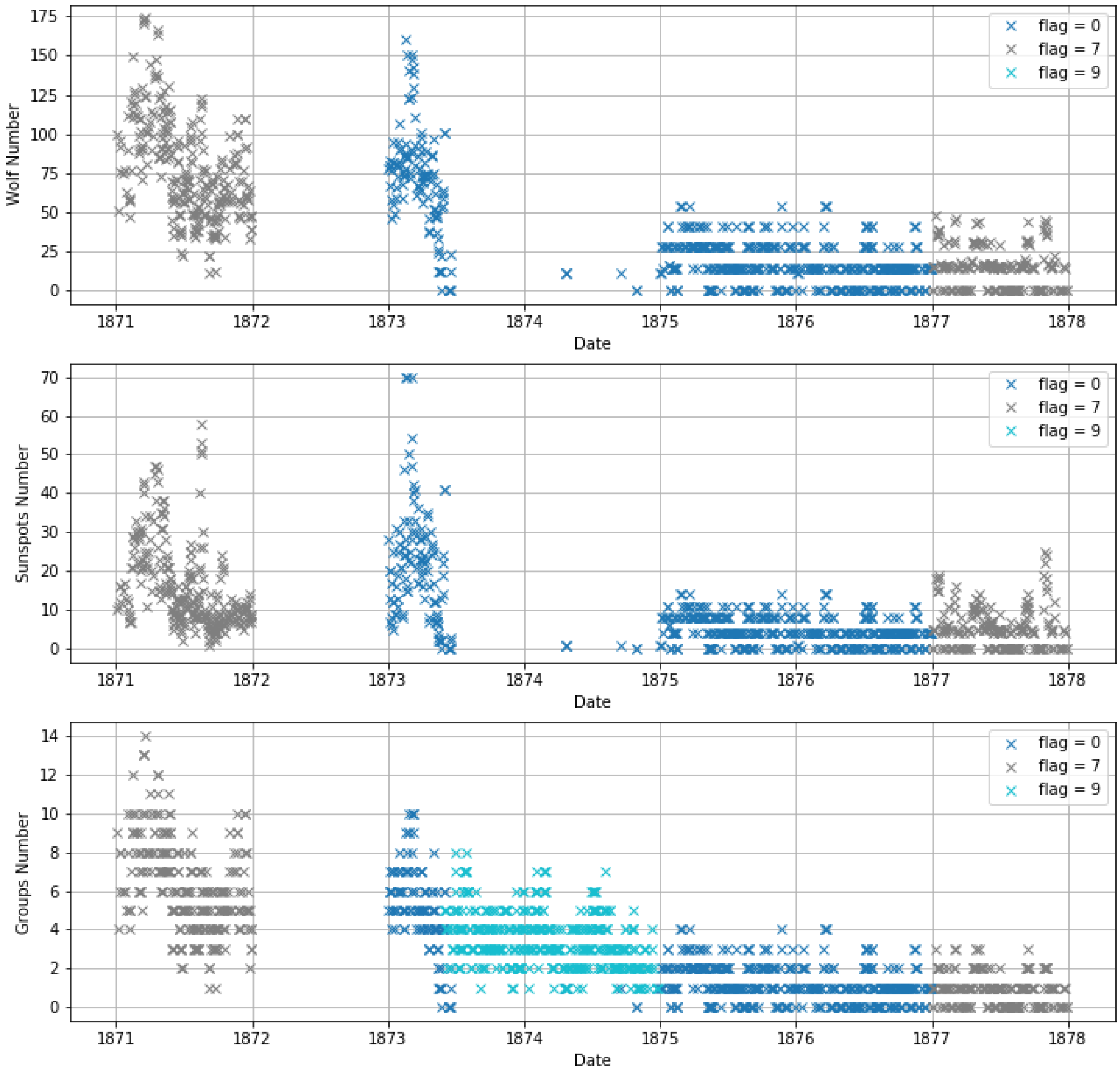
In rubrics 299 the data-table contains Angelo Secchi's group measurements and total area of the sun disk covered by umbra of the sunspots measurements, along-side sunspot number estimates s that he derives $s = f \cdot 0.15$. The factor of 0.15 he derives by least-squares fit. The assumption he makes is that the wolf number r is related to the groups number g and the area measurement f as described by equation 1

$$r = a \cdot (10g + b \cdot f) \quad \text{with } a, b \in \mathbb{R}$$

He obtained the values $a = 1.41 \quad b = 0.15$ by doing a least-squares fit. χ^2 is given in equation 4. I used the data Wolf had derived in the comments for the sunspot number.

Below is a scatter plot of Secchi's data. As you can there are still problems with it. from mid 1873 we only have data for the number of groups observed, and in 1872 there is a hole in the data.

Secchi's data seperated by flag; wolf, sunspots, groups

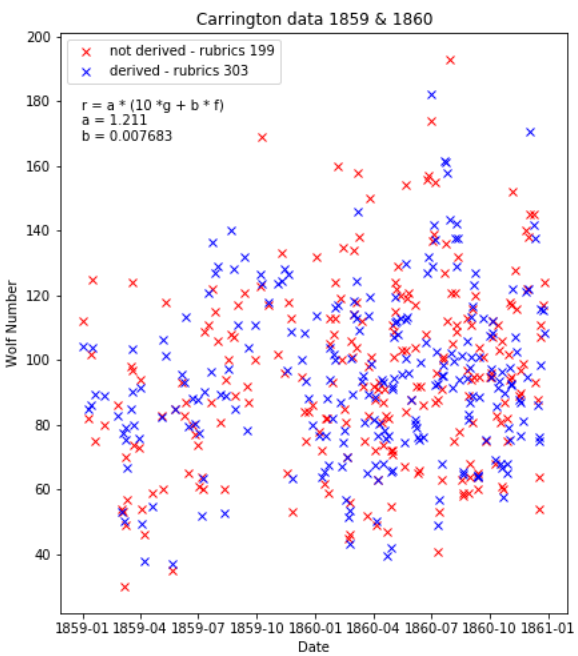


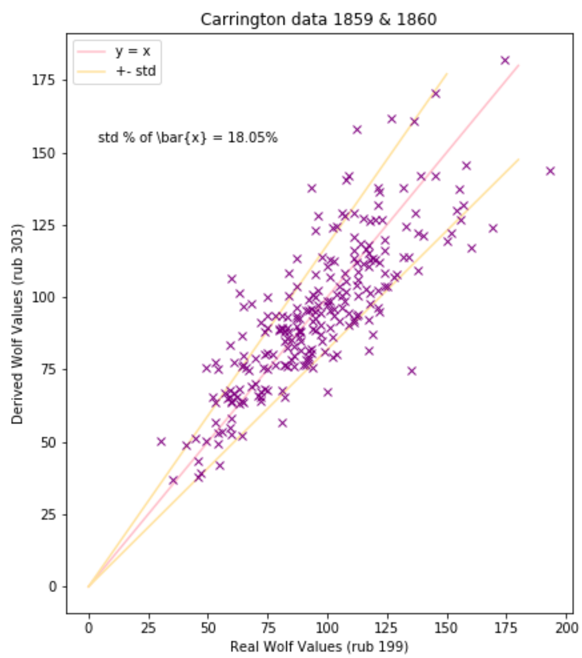
2.3.2 Carrington

Carrington has 7 years worth of observations recorded in the Mittheiulungen journals. For all 7 years he records measurements of the total area of the sun disk that is covered by the umbra of the sunspots as well as the groups number. In 1859 and 1860 Carrington *also* has a table with measurements of the groups and sunspots, it is from this 2 year over-lapping period between the two data-sets that I was able to deduce the value of the parameters a and b for the relationship describe by the author of Rubrics 299. Using equations (2), (3), (4) and (5). In this case equation 5 is merely a (long) quadratic equation. I used `scipy.optimize.curve_fit` to do this calculation.

You may notice a slight hypocrisy - after using an invariant standard deviation to calculate Carrington's best fit parameters, I then display the standard deviation as a percentage of the wolf value. With hindsight I think a more appropriate model for the standard deviation is one which varies with the wolf number r like so $\sigma(r) = \alpha + \beta r$. That said it would have been tricky to implement.

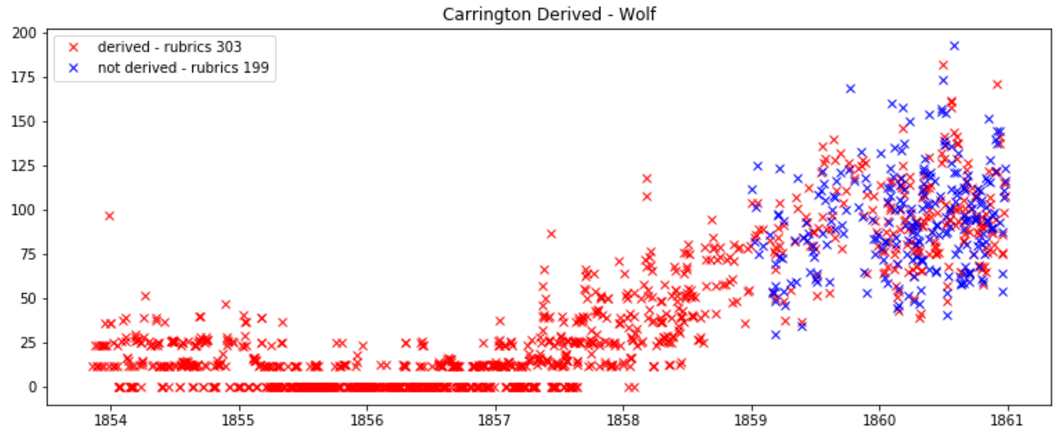
Carrington least squares line of best fit

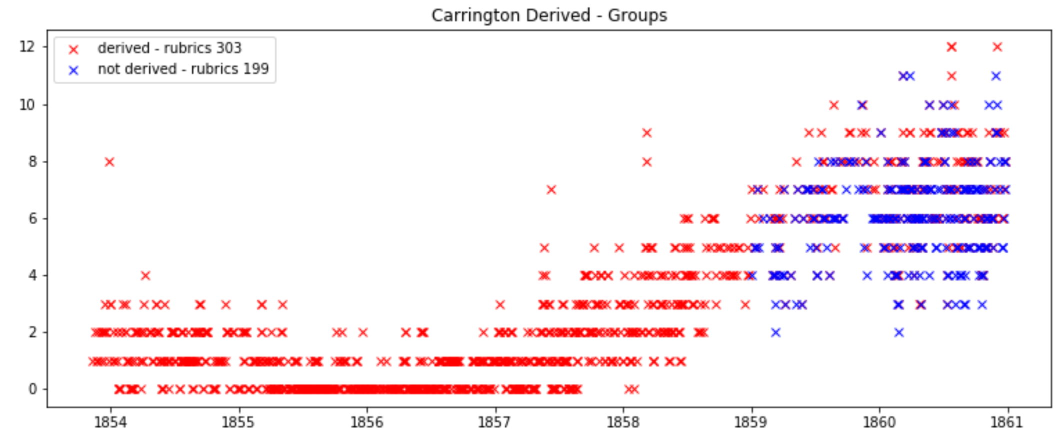


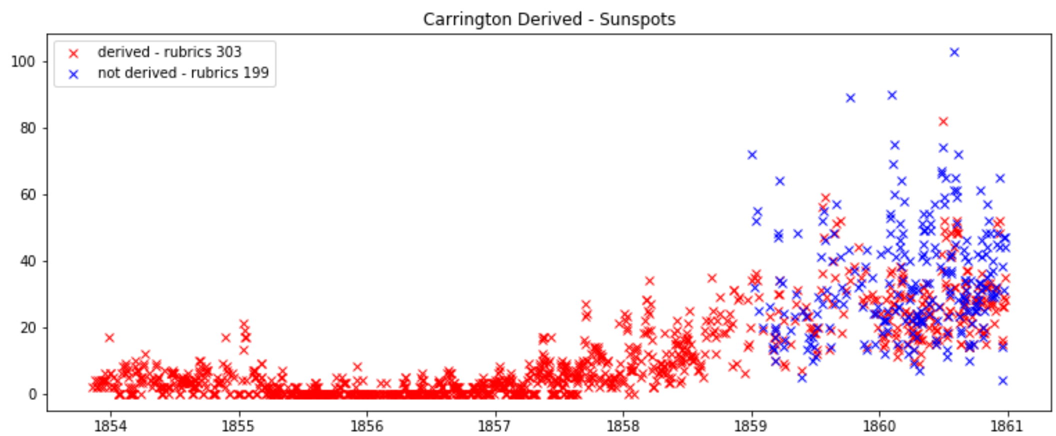


Carrington's data now resides in 3 seperate observer aliases.

- 'Carrington' - this is rubrics 199, the data from 1859 and 1860 where Carrington actually recorded the sunspots and groups number.
- 'Carrington area' - rubrics 303, 7 years worth of umbra area measurements
- 'Carrington derived' - rubrics 303 transformed by equation (1) using the best fit parameters $a = 1.21$ $b = 0.00768$







For further detail on Carrington and Secchi see July 5 to 10 in the log. Further I translated Rubrics 299 with deepl.com/translator - an online translator, the rubrics (in English and German) can be found in section 5 of the log.

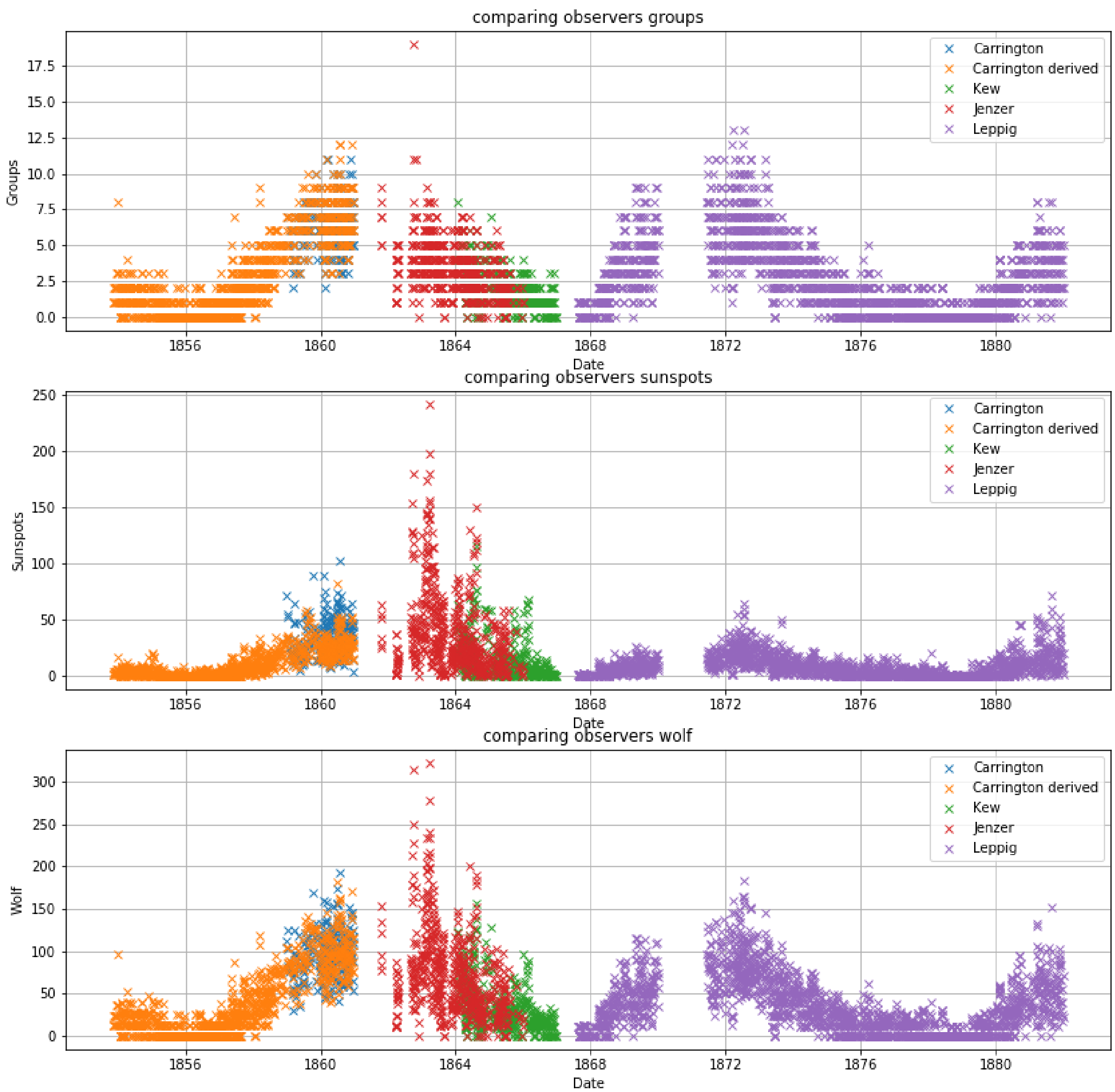
2.3.3 Kew

Applied the same fix to Kew that I did to Carrington. In Kew's case the Mittheilungen did not have any sunspots / wolf number data - only groups and total umbra area of sun-disk. The good news is that the author of rubrics 306 already did the work for me and has found $a = 0.763$, $c = 0.032 \Rightarrow b = 0.042$

$$r' = 0.763 \cdot (10g + 0.042f)$$

Here is a plot of Carrington, Kew, Jenzer and Leppig's data similar to the one above but this time using the values transformed by Wolf's umbra conversion formula. Which looks much more like it should.

Carrington and Kew derived measurements (see fig 2.3)



2.4 Inserting Schwabe's diagram data

Motivation / Why outline the process of insertion? This is important because there is a good chance that it's not exactly right. Provide the links to online data here.

2.5 Kremsmunster observatory

2.6 Other corrections

2.6.1 Tacchini patches

2.6.2 Miscellaneous

- Numerous typos were found along the way, these were corrected individually.
- Rubrics 828 - changed the observer from 'Konkoly' to 'Scharbe', and where there was P commented, the observer became 'Pokrosky'
- In rubrics 167 there are periods where it is written 'None observed' from this date to that date in the Mittheilungen. I entered these in the data-base as 0.0 data-points and commented them 'None observed'. The reasons I think these periods contain data with 0.0 rather than being days that Adams did not observe the sun are
  (a) On the days that observers don't observer there is a lack of data rather than a 'none observed'
  (b) There are long periods where there is no Adams data in the same table, so why would Adams sometimes put None observerd and then sometimes just put nothing.
  (c) It is important to include dates where Adams observed no activity in the database or his monthly / yearly averages will be completely skewed, specially because this is during a minimum.
-

3 Herr R.Wolf and Herr Wolfer - controversial transition period

3.1 Wolf

Table of each of Wolf's rubrics and which telescope he uses from 1870 on-wards. The motivation for this is that after learning about how much of a hotly disputed topic Wolf was I though it was a good idea to really present Wolf's data in as clear a manner as possible.

KEY

‡ Quadruped = *' der entweder von mir oder von Herrn Meyer nach ganz entsprechender Art mit Vergrosserung 64 meines Vierfussers erhaltenen Normalbeobachtungen'* $\longrightarrow$ *normal observations received either from me or from Mr. Meyer in quite appropriate way with enlargement* ×64 *of my four-footer*

‡ $2\frac{1}{2}$ foot = *' einem 2 1/2 Fusser bei Vergrosserung* ×42 *gemacht'* $\longrightarrow$ *a 2 1/2 ft made with enlargement* ×42

‡ Parisian = *'* $2\frac{1}{2}$ *fssigen Pariser-Fernrohr bei Vergrsserung 20 gemacht'* $\longrightarrow$ *the* $2\frac{1}{2}$ *foot Parisian telescope with enlargement factor* ×20

‡ Parisian† = Most probably the Parisian telescope : *‘ Ein beigesetztes * bezeichnet Beobachtungen, welche ich mit dem kleinern Instrument machte, und mit 3/2 in Rechnung brachte ’* ⟶ *A * indicates observations, which I made with the smaller instrument, and charged with 3/2.*

‡ Pocket 1 = *‘wenigen mit * bezeichneten Beobachtungen wurden auf Ausflugen mit einem kleinen Taschenfernrohr erhalten’* ⟶ *observations marked with * were obtained on trips with a small pocket telescope*

‡ Pocket 2 = *‘Die mit * bezeichneten Beobachtungen sind auf Ausflugen mit einem kleinen Taschenfernrohr angestellt, und werden mittelset des Factors* $\frac{3}{2}$ *den brigen homogen gemacht’* ⟶ *observations marked * are made on excursions with a pocket telescope, the observations are homogenised by means of a* $\frac{3}{2}$ *adjustment factor*

‡ A telescope in brackets e.g. (Parisian) means that **no telescope was mentioned** but what is in the brackets is almost certainly the one used.

‡ Wolf + Carr = I could not find a description of the telescopes used by Wolf or Carrington, he just says somewhat cryptically that the unmarked observations were made by him and Carrington

‡ s = observations made by Schwabe that found their way here

‡ var = various observers, a whole host, usually 4 or 5.

Rub_{id}	Rub_{no}	Mitt no	Page	Date	Primary	*=Secondary	Comments
-	0	1	153-9	1849-55	var	-	see footnote **
-	0	3	110	1856	var	-	-
-	0	6	125	1857	var	-	-
-	0	8	67	1858	var	-	-
-	0	11	2	1859	Wolf, Carr, s	-	see footnote ‖
-	0	12	69	1860	Wolf + Carr	Parisian† + s	mostly primary ¶
-	0	14	120	1861	$2\frac{1}{2}$ foot	Parisian†	primary obs = Wolf mostly secondary
-	0	15	134	1862	Quadruped	Parisian†	primary obs = Wolf mostly secondary
-	0	16	164	1863	Quadruped	Parisian†	primary obs = Wolf mostly secondary
-	0	17	194	1864	Quadruped	Parisian†	primary = Weilenmann + Wolf mostly secondary till winter
-	0	21-30	18	1865	Quadruped	Parisian†	prim: Weilenmann, Fretz, Wolf mostly primary
-	0	21-30	74	1866	Quadruped	Parisian†	prim: Weilenmann, Fretz, Wolf mostly primary
218	0	24	104	1867	Quadruped	Parisian†	prim: Weilenmann, Meyer, Wolf roughly half half
219	0	26	208	1869	Quadruped	Parisian†	primary: Meyer + Wolf
158	274	30	403	1870-71	Parisian	Pocket 1	Mostly primary
165	289	33	111	1872	(Parisian)	Pocket 2	Mostly primary §
179	313	36	266	1873	(Parisian)	-	100% primary
186	326	38	394	1874	(Parisian)	-	100% primary
192	335	39	418	1875	(Parisian)	-	100% primary
198	344	42	50	1876	(Parisian)	-	100% primary
205	365	46	185	1877	(Parisian)	-	100% primary
212	385	49	251	1878	(Parisian)	-	100% primary
236	410	50	298	1879	(Parisian)	-	100% primary
220	430	52	50	1880	(Parisian)	-	100% primary
258	453	55	160	1881	(Parisian)	-	100% primary
288	470	59	337	1882	(Parisian)	-	100% primary
247	488	62	84	1883	(Parisian)	-	100% primary
266	505	64	158	1884	(Parisian)	-	100% primary
281	522	67	299	1885	(Parisian)	-	100% primary
295	539	69	349	1886	(Parisian)	-	100% primary
329	563	71	16	1887	(Parisian)	-	100% primary
358	584	73	109	1888	(Parisian)	-	100% primary
386	603	76	226	1889	(Parisian)	-	100% primary
418	624	78	296	1890	(Parisian)	-	100% primary
433	642	80	381	1891	(Parisian)	-	100% primary
312	664	82	53	1892	(Parisian)	-	100% primary
330	685	84	120	1893	(Parisian)	-	100% primary

For more information on rubrics 1 which contains a huge chunk of WOLF - S - M from 1849-55 see **??

‖As indicated the observers are all mixed up, however he does specify outside the table the days where Schwabe is the real observer. This data seems to have been correctly digitized.

¶This year he uses His and Carrington's data from 1859 and 1860 to derive the correction factor $k = 1.5$ for his Parisian telescope, and also for Schwabe but this is less important. Further, he finds that Carrington's k factor is the same as his.

§*Pocket 1* and *Pocket 2* are the same telescope, but only for *Pocket 1* is there mention of a correction factor

Conclusions:

- Things look very good for WOLF - P - M, asides a handful of observations in 1870-72 all the data seems to have been entered in correctly. Everything after 1872 is almost definitely correct (ignoring the fact that there may be typos). Before 1870 it's slightly trickier, I still haven't excluded the possibility that Wolf may have been using the Pariser as far back as the 1850's but just failed to mention it.

- As for WOLF - S - M things are not so good. From 1849 to 1860 we have no way of identifying which observations are Wolf's own, let alone what telescope he was using. The only information he provides us with from 1856 on-wards is how many days of the year he observed, and how many days he used other people observations, but not exactly when - actually he doesn't even tell us that, he says how many days he or his assistants observed though the quadruped $\times 64$ magnification telescope.

- There are 526 flag = 1 observations with WOLF (S-M & P-M) as observer. 36 of these are because they are not his, I haven't moved them because I haven't yet found out whose they are (the observations are attributed to Wolf before 1849). Most of the others seem to have missing sunspots numbers, some of them are marked as having missing groups but I suspect the digitizer just typed the group number into the sunspots column because most of these are $x \leq 5$. These still have to be dealt with. This may involve deriving a formula that guesses a wolf number based purely on the group - perhaps it could also take as a parameter what the average sunspot index is (heavily smoothed) around that time, but I shouldn't get ahead of myself I still have alot to do and not much time.

3.2 Wolfer

4 Problems that remain with the data

- Grouped observers (observer aliases that contain several names) - most of these have been dealt with but I think there is still a small handful of rubrics where the alias consists of several names (see size_data_by_observer_hist() plot to check).

- Commented observers (A's data is under B) [$\approx$ 3000 dta-pts]

4.1 Flags

4.2 Non-derived measurements

Ferrari submitted area measurements (in 2 rubrics: 425 and 398) but no sunspot measurements... The only thing I can think of to rectify this would be to scale his sunspot measurements using the factor Wolf calculated for Secchi $s \approx 0.15f$. For the moment they are flagged 7 and are sitting in the database un-derived.

4.3 Other problems

4.3.1 Schwabe and Wolf event-plot out-liars

Some of Schwabe and R. Wolf's data appears in the database before they started observing anything. I didn't delete these because the data probably belongs to someone and if we throw them away they may be lost forever. But I couldn't find them in the Mittheilungen journals (they are labeled mitt 0 rubrics 0, which could be anywhere really).

4.3.2 Miscellaneous

- Those data-points marked with flag=6 in rubrics 811 are data-points where the sunspot number and dates where somewhat ambiguous and cryptic. It is perhaps worth having a German speaker examine this one.
- In 1869 there is a hole in Leppig's data that needs to be investigated and fixed if possible.
-

5 Visual data - guide to plots and graphs

5.1 Scatter plots

5.2 Histogram

5.3 Bar charts and pie charts

6 Condensed Log

6.1 Before The Solstice

I only started the log on the 21^{st} so I forgot the details of what I was doing before then. The time was spent learning the basics of `SQL` and how to interface with an `SQL` database through the mysql terminal; acquainting myself with the data and with what it is I ought to be doing. This is the period where I wrote some of the basic methods that I now use every day for accessing and connecting with the Mittheilungen.

6.2 Friday June 21

- Started writing the log
- Made 'searching_the_manuals.py'
- Searching database for 'uncertain' comments

6.3 Monday June 24

- Discovered and sorted many duplicated data-points, (duplicate means two dta-pts for one date and one observer)
- Methods used can be found in `searching_the_manuals.py`

6.4 Tuesday June 25

- Backed up the databases and started flagging for the moving process.
- Wrote a new script to deal uniquely with deleting the duplicates (and putting them into '`RUBBISH_DATA`')
- Commented the rubbished duplicate data points

6.5 Wednesday June 26

- Wrote methods for finer mass commenting (in `db_edit.py`)
- Flagged data with abnormally large groups and or sunspots numbers. (FLAG=4 WHERE > 100 ; FLAG=5 WHERE > 250)
- Set 212 flags 3 for putting things in the bin. There are still 4000 pairs of duplicates that need attending to, originally there where 14000
- Scrutinised what I had flagged, reread my scripts, checked that things are in the right place.
- Having doubts about what is reasonable / unreasonable sunspots number.

6.6 Thursday June 27

- Scrutinised flagged data from yesterday
- Turned my attention to the data labeled '*' in the comments
- Moved the flagged duplicates to `RUBBISH_DATA`
- Found entire rubrics worth of data written in wrong year
- Started writing `corrections_needed_handwritten.txt`, to make clear all my tasks.

6.7 Friday June 28

- The notes I took about the duplicates can be found in different_value_duplicates.txt, some things I found interesting so I decided to copy most of the file into this report (see long version of the log)

6.8 Monday July 1

- Using what I did on Friday to bin some duplicated data and modify some other data
- Made a new alias in `DATA_SILSO_HISTO` (and `BAD_DATA_SILSO`) called 'Brunner Assistent'.
- Backed up the the databases to sql files
- Most of this data has been cleaned up, the rest can be done by hand

6.9 Tuesday July 2

- Made some pretty plots in *suspicious sunspots plots.ipynb* in the root directory
- Made a method in the jupyter notebook mentioned above that plots an observer's stuff and color codes the flags.
- Checked some of them in the journals
- Started patching Tacchini's missing holes

6.10 Wednesday July 3

- Continued fixing Tacchini (see figure **??**)
- Went back to searching the manuals for errors from error sheet.
- Looking through for 'uncertain' comments
- Figured out what COMMENT='?' means; blurry image / bad definition of img
- Found some comments where there is both an observer and a question mark at the same time, for these ones I left the comments as they are and changed only the flag from 1 to 2.
- Finished looking at red comments (I still need to change them and move them all with python, I will do it tomorrow.)
- Looking at blue comments (the ones where comments are just numbers)
- Backed up databases

6.11 Thursday July 4

- Looking into Carrington's case.
- I updated the flag 7 to "derived from area-measurement" and flagged all of Secchi's sunspot values that were derived from the penumbra and / or umbra.
- Dealt with Secchi
- Spoke to F. Clette about the possible conversion from the 'aire' to a sunspots number. He gave me some clues as to where to look in the mitt.
- Excitement! I found on page 131 of Mitt 31-40 written after rubrics 299 a description of how the author (I think R. Wolf himself) derived a formula for turning Secchi's 'aire' into a sunspots number
- 8.6 here is what is written in German and Italian, with a translation in English.
- Backed up databases

6.12 Friday July 5

- Continued working on Carrington - Main event = did a least-squares regression fit to optimise the constant values in the equation that transforms 'aire' into wolf number.

6.13 Monday July 8

- Finished deriving Carrington
- Backed up databases

6.14 Tuesday July 9

- Derived Kew's misbehaving data
- Made a new 'README.md' that auto-generates based on what is inside my python scripts
- Tidied the report and added some figures

6.15 Wednesday July 10

- Sabrina gave Arnaud and I a tour of the Observatories facilities
- Separated Carrington into two aliases
- Figured out what to do with Secchi (now I just have to do it)

6.16 Thursday July 11

- Dealt with rubrics 375, see `secchi_derivation.ipynb`
- Dealt with 2 more of Secchi's rubrics, there are still 2 annoying ones
- Continued checking and correcting typos and anomalies from the blue comments sheet

6.17 Friday July 12

- Dealt with the red comments, changed alot of their comment to '?' which was written in the journals and changed the flag to 2
- Put some thought into how to deal with oranges aswell as Wolf / Wolfer

6.18 Monday July 15

- Transferred all the data with flag = 2 into the database `GOOD_DATA_SILSO`
- Upgraded the plotting methods in `graphs_helper.py`
- Plotted Wolf and Wolfer and aliases in which they appear - see `wolf_wolfer_investigation.ipynb`
- Brainstormed how I was gonna tackle wolf / wolfer's data

6.19 Tuesday July 16

- Planned out how I was going to tackle the Wolf - Wolfer problems
- Wrote some methods to smooth data and also to plot the sunspots number
- Corrected typos and errors

6.20 Wednesday July 17

- Corrected Adam's data by hand data-point by data-point
- Made some more plots in the `wolf_wolfer_investigation.ipynb`
- Investigated Wolfer some more
- Accidentally deleted database and lost all the edits I made to the database today... :([luckily I have backups from yesterday]

6.21 Thursday July 18

- Accidentally deleted the data again, re-corrected Adam's data by hand
- Fixed the data and imported the old data into the database `ORIGINAL_DATA_SILSO_HISTO`
- Wrote to Laure and she gave me some good ideas for how to detect drift
- Rereading papers to get better understanding of what I ought to be doing

6.22 Friday July 19

- Made a cool plotting tool in an attempt to visualise the drift of observers, didn't work out brilliantly
- Made a couple new aliases which I populated with data 'Mooser' (for rubrics 122 only)
- Made the alias 'Wolfer P' who now has 307 datapoints.

6.23 Monday July 22

- Wrote a cute little script to automate backing up the databases and committing with git (there will be more frequent backups now).
- Translated a big long rubrics
- Sorted all the data attached to fk_observers IN (60,61,62) - all composite observers from these strange rubrics_number = 0 in the mid to late 1800's - into their proper observers.

6.24 Tuesday July 23

- Made some fancy plots and plotting tools : `size_data_by_observers()` plots a bar chart of of all the observer aliases on the x axis and on the y axis it plots how many data-points are associated with them. `event_plots()` shows you the observer aliases on the y axis this time, and the x axis is the dates, plotted is all the dates each one observed on.

6.25 Wednesday July 24

- Updated the create_readme file
- Did some event plots and investigated 'Ricco, Zona, Mascari'
- Learned a whole lot of things from F. Clette about his work, the current state of solar physics and some interesting things about Burnner and other observers

6.26 Thursday July 25

- Added a descriptor in markdown to the beginning of each jupiter notebook
- Re-plotted some data from the Wolf - Wolfer transition period that has been modified since last time I plotted it and the change is magnificent!
- Finally abolished 'Ricco, Zona, Mascari' and appropriately sorted the data
- Launched an investigation of rubrics 684 which is very confusing. There are many problems with it.

6.27 Friday July 26

- Sorted out rubrics 684

6.28 Monday July 29

- Started investigating Wolf and Schwabe's mysterious holes

6.29 Tuesday July 30

- Doing some archeological excavations on the Wolf mixup and Mitteilungens 1 though 10
- Deleted some erroneous data of WOLF - S - M (same needs to be done for 67)
- Failed to find a suitable correction for certain things see log for details

6.30 Wednesday July 31

- Did some more corrections on Wolf's data (basically data-entry)

6.31 Thursday August 1

- Discovery, R. Wolf uses 3 telescopes not 2. While he is in charge of the observatory in Zurich he uses what he refers to as the '$\times$64 magnification quadruped' (paraphrasing), in 1870 he switches primary telescope to the Parisian $\times$20 magnification, and all the while when he goes on trips he takes with him a pocket telescope. It is still unclear as to weather he uses the Parisian much before 1870, my guess is that while he was still going to the observatory every day all the official measurements would be made with the big one, and the Parisian which most likely stayed in his home was used only for recreational purposes.

6.32 Friday August 2

- Found that R. Wolf might actually use primarily the Parisian as his secondary before his retirement (1867-9)
- Added Schwabe's online data to the databases
- Made a method to make stacked area plots for the frequency of observation

6.33 Monday August 5

- Perfected stacked area plots to point of being fully functional with options
- Started having a go at the orange highlighted comments - am changing aliases based off of comments, after cross checking what the comment says and what is written in the preamble of the rubric each time of course.

6.34 Tuesday August 6

- Cleared out some of the data in BAD DATA SILSO which was flagged weirdly (lots of the data points with flag = 4)
-
- Found some of Wolf's missing data

6.35 Wednesday August 7

- Thoroughly scrutinized Wolf's data in the Mittheilungen journals and arranged my findings into a table with crucial information concerning his observations

6.36 Thursday August 8

- Did some final updates of the data flagged with flag=4, flag=5 and flag=9
- Corrected 'Schwabe Drawing' 's data
- Made some edits to the flag section of the report

6.37 Friday August 9

- Made the histogram plotting methods

6.38 Monday August 12

- Plotted the pie-charts and bar-charts that display the changes that I have effectuated to the database.
- Brainstormed final draft of report

6.39 Tuesday August 13

- Modified some of the plotting functions
- Started writing second draft of report

7 Conclusions

7.1 Before and After - outline of the modifications I made to the database

7.2 Problems that remain with the database

This idea will probably never come to fluition - in the spirit of Herr Wolf who was the first one to estimate π by monte-carlo approximation, find the frequency of random errors in the Mittheilungen by Monte-Carlo approximation (should take about 1 day). Pick 1000 to 2000 datapoints at random using some algorythem, and find each one in the mittheilungen to see if it is entered correctly, do some stats on this and calculate a student error factor. (better to look up if there are known numbers for this kind of task)

8 Miscellaneous

8.1 SQL data-table format

8.1.1 The original sql data tables format

`DESCRIBE DATA`

Field	Type	Null	Key	Default	Extra
ID	int(11)	No	PRI	NULL	auto_increment
DATE	date	YES		NULL	
FK_RUBRICS	int(11)	YES	MUL	NULL	
FK_OBSERVERS	int(11)	YES	MUL	NULL	
GROUPS	int(11)	YES		NULL	
SUNSPOTS	int(11)	YES		NULL	
WOLF	int(11)	YES		NULL	
QUALITY	int(11)	YES		NULL	
COMMENT	text	YES		NULL	
DATE_INSERT	datetime	YES		NULL	
FLAG (i added this)	tinyint(1)	YES		NULL	

`DESCRIBE OBSERVERS`

Field	Type	Null	Key	Default	Extra
ID	int(11)	NO	PRI	NULL	auto_increment
ALIAS	varchar(50)	YES		NULL	
FIRST_NAME	varchar(50)	YES		NULL	
LAST_NAME	varchar(50)	YES		NULL	
COUNTRY	varchar(50)	YES		NULL	
INSTRUMENT	varchar(50)	YES		NULL	
COMMENT	text	YES		NULL	
DATE_INSERT	datetime	YES		NULL	

`DESCRIBE RUBRICS`

Field	Type	Null	Key	Default	Extra
RUBRICS_ID	int(11)	NO	PRI	NULL	auto_increment
RUBRICS_NUMBER	int(11) unsigned	NO		NULL	
MITT_NUMBER	int(11) unsigned	NO		0	
PAGE_NUMBER	int(11) unsigned	YES		NULL	
SOURCE	text	NO		NULL	
SOURCE_DATE	date	YES		NULL	
COMMENTS	text	YES		NULL	
DATE_INSERT	datetime	YES		NULL	
NB_OBS	int(11)	YES		NULL	

8.1.2 New SQL data table format

`DESCRIBE DATA` (the only table)

Field	Type	Null	Key	Default	Extra
ID	int(11) unsigned	No	PRI	NULL	auto_increment
DATE	date	YES		NULL	
GROUPS	int(11)	YES		NULL	
SUNSPOTS	int(11)	YES		NULL	
WOLF	int(11)	YES		NULL	
COMMENT	text	YES		NULL	
DATE_INSERT	datetime	YES		NULL	
OBS_ALIAS	varchar(50)	YES		NULL	
FIRST_NAME	varchar(50)	YES		NULL	
LAST_NAME	varchar(50)	YES		NULL	
COUNTRY	varchar(50)	YES		NULL	
INSTRUMENT_NAME	varchar(50)	YES		NULL	
RUBRICS_NUMBER	int(11)	YES		NULL	
MITT_NUMBER	int(11)	YES		NULL	
PAGE_NUMBER	int(11)	YES		NULL	
FLAG	tinyint(1) unsigned	YES		NULL	
RUBRICS_SOURCE	text	YES		NULL	
RUBRICS_SOURCE_DATE	date	YES		NULL	

8.2 Backbone observers table

Backbone Observer	Main interval	Full interval	Nb Observers
Staudach	1749 - 1787	1740 - 1822	15
Schwabe	1826 - 1867	1794 - 1883	20
Wolfer	1878 - 1928	1841 - 1944	21
Koyama	1947 - 1993	1920 - 1996	36
Locarno	1957 - 2015	1950 - 2015	22

Source : https://files.aas.org/astronomy2015/Presentations/DE$_Fr\%C3\%A9d\%C3\%A9ric_Clette_Heliosphere.pdf$

8.3 Ideas for the detection of more problems that never saw the light of day

In the duplicates section (2.1) there were entire rubrics entered in the wrong year / under the wrong observer. The only reason they were detected by the duplicates finding algorithms was because they just so happened to be entered into a year where there was already data for that same observer, or in the case of the wrong observer, they were written in under an observer who was also observing in at the same time. It is possible that some data was entered in the wrong year / under the wrong observer without coinciding with other data and so went undetected by these methods.

8.4 Thought repository - ideas that may or may not come into fluition depending on how efficiently I work and get things that need to be done done

- make some data visualisations to compare each observer's primary and secondary observing equipment
- for each day / month / year find the highest observation and the lowest observations and add it to the graph so that we have like an upper bound and a lower bound.
- figure out how to smooth graphs with matplotlib and make something nice out off the big mess i currently have
- pie chart of observers with their number of observations

- in the final sunspots number graph cut it into 3 or 4 sections that mark changes in the theory behind sunspots: before wolf ; time where Plato's ideas of the sun being a perfect sphere were still alive and well ; 1908 George Ellery Hale discovers the magnetic link (p14 of nature's 3rd cycle) ; 1955 Eugene parker's theory (p19 of nature's 3rd cycle) ; Nasa send their probe to near the sun

8.5 old preamble - to be edited out, maybe some stuff written here is salvagable

The aim of this project is to do a quality control of the data in DATA_SILSO_HISTO. Once the data is fixed and cleaned up, it will be stored on a new database - temporarily named GOOD_DATA_SILSO in a more user-friendly format to what currently exists. I will also get rid of any useless or redundant columns (such as the observers comment column - there are no comments)':). A third, temporary database will be mad to keep a closer eye on the data that still needs to be examined with more scrutiny : BAD_DATA_SILSO. This database will act as intermediary between DATA_SILSO_HISTO and GOOD_DATA_SILSO. We will effectively be storing 2 databases-worth of information in 3 databases. The original DATA_SILSO_HISTO will have the old data and will be corrected in due course. The intermediary BAD_DATA_SILSO will start as a copy of DATA_SILSO_HISTO and end up empty as the corrected data is removed from it and placed, in the new format, into GOOD_DATA_SILSO.

8.6 Converting the f ('aire')

This section has been moved to the log