DATA_SILSO_HISTO
Quality Control Report

Stephen Fay

July 5, 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. For the past 3 to 4 hundred years 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. 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.

2 Processus de filtration / corigee du data (log) / quality control

2.1 Everything wrong with the data

First, it's important to note that note that though I am doing a quality control I do not wish to die of boredom. I will not be verifying each of the 205003 data-points by hand in the Mittheilungen journals, in any case this if I went about it this way I would probably miss most of the errors.

2.2 Annotation keys

2.2.1 What do the flags mean?

0	1	2	3	4
same as Null	suspicious	Comment in journal = ?	move to bin	suspiciously high
5	6	7	8	9
very suspicious	misc see comment	derived from area-measurements	null groups	null sunspots

Table 1: Flags key

0. The default for the flag is NULL, when is estimate that the datapoint is perfect and there is nothing wrong with it, I can put it to zero 0.
1. If the data looks fishy but I'm not quite sure either what is wrong with it or how wrong it is this is flagged with a 1 - the default.
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 2.4.10 - July 3 for speculation on what I think comment '?' means.
3. A flag that signifies that this data point is definitely going into the bin
4. 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. For data that is definitely wrong, the difference between 5 and 4 is illustrated by example: if i find that a datapoint has a groups number of 30 I will mark it with a 4 and comment it, because this is suspicious, if a datapoint has a groups number over 60 or above, it will be marked with a 5 (trust me there are some in the hundreds).

2.3 Search and correct.

2.3.1 Outline

For the first week and a half or so, I spent the bulk of the time acquainting myself with the Mittheilungen journals, and with the software that is used to store and access the database. I also developed the tools in python to facilitate my access to them and to perform the tasks that I need to perform for the filtration process.

2.3.2 Equations

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

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

2.3.3 Python scripts - what they contain

2.4 Condensed Log

2.4.1 Before The Solstice

I only started the log on the solstice 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.

2.4.2 Friday June 21

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

2.4.3 Monday June 24

- Discovered and sorted a bunch of duplicate data, two data-points for one date and one observer
- Methods used can be found in `searching_the_manuals.py`

2.4.4 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

2.4.5 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.

2.4.6 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 many instances of data written in wrong year
- Started writing `corrections_needed_handwritten.txt`, to make clear all my tasks.

STOPPED CONDENSING HERE, WHEN YOU GET BACK START AGAIN

2.4.7 Friday June 28

- Continued what I was writing yesterday, looking through the manuals and identifying. 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

- Carrington datapoint id=31460 needs to go in the bin this is clearly a penumbra

- All observations in rubric number = 808 with fk_rubrics = 461 were made by Konkoly not Wolfer

- All observations in rubric number = 820 were made by Wolfer in the year 1900 and not 1899, many observations are written correctly except for the year of the date, I found that the months and days correspond perfectly with what is written in the journals but not the year!

- The Broger duplicates for 1899-03-17 and 1899-04-18 make very little sense, they are broger under two different aliases but in the same rubrics 801. There are only two of them so they can be done by hand. My guess is that they were punched in twice by different people and my identical duplicates algorythem has already dealt with the ones that were exactly the same, so these are the 2 instances where one of the people got it wrong.

- Observations in mitt 129 rubrics number 02 (or 12902) were all made by Broger in 1931. Manny if not all of them were typed in wrong, it says they were made in 1908. Thankfully everything else about them seems correct.

- For Rubrics 1057 and 1058 there is a serious problem - Both are labled with the wrong observer!
  - Rubrics number 1057, the observations are made in the Capodimonte Observatory by Dr. E. Guerrieri,
  - Rubrics number 1058, the observations are made in Floreze by Robert Lucchini. I checked and both of these boyz are real observers with aliases, so this needs to be corrected.
  - The reason there is confusion about these two is because Herrn J. Sormano in Turin is mentioned since the observations come from letters of correspondence between Sormano and the two observers mentioned above. It is quite possible that this has been going on in other places under the radar. The only reason these two were detected was because they just so happened to be attributed to the same guy for observations on the same date. What I propose is that we have someone (perhaps me, Asside: maybe an AI could do this if we scanned every page of the journals and then trained a neural net to read the rubrics descriptors and figure the observer based on that... This might be hard for this task but I can see how some machine learning could come in handy for error detection and quality control) go through each rubrics number and make a list where every rubrics number is ascribed to an observer, specially for the rubrics descriptors that include the names of several people, come to think of it a German person would proably do a much better job at this then me.
  - To speculate further it is possible that this issue runs deep and that many of the holes in the data are infact a concequence of the kind of error described above (by the way I can cut out alot of my rambling from the final draght of the report, or make a condensed version, I'm just in the habit of writing everything down so that future me can follow the thought process, sometimes it helps ok!)

- Observations from rubrics number 1279 were incorrectly labled as comming from 1919 when infact they are from 1920. The observer (Prof. Anne Young) and all other info is correct.

- Something annoying happens in 1929, Both Brunner - 'Wm. Brunner' and his assistant - 'W. Brunner, Assistent' seem to be observing at the same time [BOTH OF THEM ARE DENOTED WITH THE ALIAS 'Brunner'] with the same 8 cm aperture 64x magnification polarised helio-telescope. Rubrics 1624 (fk_rubrics=840) is Brunner and his assistent's observations are from Rubrics 1675 (fk_rubrics=842). I don't know what we should do, perhaps create another alias 'Brunner Assistent' (pink page marker)
  - Same thing happens in rubrics 12501 (fk=844) is the real Brunner, and rubrics 12503 (fk=846), this is Brunner Assistent (yellow page marker).
  - Same again in 1931. Rubrics 12901 (fk=848) is the real Brunner, and rubrics 12903 (fk=962), this is Brunner Assistent (orange page marker)
  - ...This goes on until (see the file for details)...
  - And in 1944. Rubrics 14401 (fk=1006) is the real Brunner, and rubrics 14402 (fk=1007) is Brunner Assistent. (green page-marker on page 112)
  - Comments on Brunner: I'm annoyed that the assistant(s) doesn't have a name because we now have no idea how many there where. Also (s)he deserves credit for those 10 odd years of commited observation! Because this assistent has been observing with Brunner from 1929 to 1944 he at least deserves an alias.
- Messerschmitt and Wasnetzoff
  - Something strange happens with Messerschmitt, there are two different sunspots values written in for 1908-02-15, one of them has no rubrics number and values 3,17,47 The other has a rubrics number 1028 and values 2,7,27. So I looked in the journals under rubrics 1027 and for this date I found the values 3,17,47 which are the values written in where no rubric is specified. Very strange. There are a total of 4 duplicates for Messerschmitt and this is the only one that has a rubrics number. I find this very strange... I don't really know what to do.

2.4.8 Monday July 1

- Using what I did on Friday (wrote that list of things that were wrong with specific duplicates) to bin some data and modify other data
- Started writing methods flag3_from_correction_txt() and change_rubric_observer() and change_date_rubric() in dealing_with_duplicates.py, and then I realised that all this was much simpler and could be done faster and less error prone if I just punched in the queries to through mysql directly, so this is what I did.
- With care and delicacy I changed the observer aliases / and for the old and bad database I changed the FK_OBSERVERS with the terminal. Now I move onto changing all the faulty dates, this requires a bit of coding because I need to loop through the dates and change each date individually.
- Wrote change_dates() and it's helper change_date_rubric() (in dealing_with_duplicates.py) and executed them in more time than it should have taken. My head is not clear today, but I was rigourouse, there should not be a mistake.
- I am changing the databases quite a bit so I saved a new backup file
- made a new alias in DATA_SILSO_HISTO (and BAD_DATA_SILSO) called 'Brunner Assistent'. I know the name lacks imagination but hey (I couldn't find who it might be online - future me : this is a reminder to ask Frederic if he might know, be ready with dates...).

- I just realised that when I was changing some of the data in GOOD_DATA_SILSO I only changed the aliases, to correct this I will either write now or once everything is finished a method that goes through each data-point finds the alias and corrects the observer data (things such as the country of observation, the observer comments and the instrument etc.)
- wrote and executed the command in dealing_with_duplicates.py change_alias_to_brunner_assistent(), this corrected the brunner assistent problem we had
- backed up the the databases to sql files
- I realised that I had missed correcting the alias for rubrics 195 from Franzenau to Weber so I did that just now
- Ran the method that shows me what is wrong with the duplicates to see how effective my cleanup has been : interesting, the duplicates that were Weber's but marked Franzenau had already been entered under Weber but with rubrics_number=0 and no references in the Mittheilungen, so I reran the flag_many_duplicates.
- Some of the 'Tacchini and Milesovich' are missing sources but the duplicates are identical, so I moved one from each to the rubbish bit by calling the method delete_entered_twice_duplicates() from dealing_with_duplicates.py. There was also the issue that Broger had some identical data typed in for him in two using two different observer ids that both refer to him... so i added an elif statement into the delete entered twice duplicates method to deal specifically with this.
- Most of this data has been cleaned up, the rest can be done by hand

2.4.9 Tuesday July 2

- exploring ways of creating visualisations that will help me catch some of the suspicious data-points. Right not my task is hunting down those ones which I labelled 'suspicious sunspots'
- made some pretty plots which you can find in *suspicious sunspots plots.ipynb* in the root directory
- Using the plot I was able to check some of Weber's suspicious sunspots in the Mittheilungen. And strangely enough Weber's observations are correct! It seems he really did see 476 sunspots on the sun on the 25th of september 1870. I mean the patter fits it's just unsusually large. And you need to be dedicated to count hundreds of spots every day.
- Made a method in the jupyter notebook mentioned above that plots an observer's stuff and color codes the flags. I investigated Tacchini's green (flag=4) datapoints and they are infact correct, they appear in the journals. There is one datapoint from Tacchini which I corrected by hand, this one had a wolf number of 61, I found it in the journals, the correct value was 6 - typo. I corrected it manually.
- In the rubric 279 mitt 30 page 409 Tacchini observes a bunch of sunspots without observing any groups. This is annoying because it means we have no wolf! Other than that they seem to have been entered in correctly, I have not yet transfered these to the GOOD DATA SILSO database, they are still only in DATA SILSO HISTO and BAD DATA SILSO
- I have an idea that we could do some stats and deduce a wolf number even without the groups, and give it a special flag, based on some probability we estimate a wolf number. There might be some complications with this tactic, here are 3 possible ways of doing it and their weaknesses:

1. For each sunspots number possible {1,2,3...} find the corresponding most likely wolf number by going though each data point with that number and doing a distribution (probs normal) to find which is the most likely wolf number associated with this sunspots number. Essentially we define a function term-by term $f \stackrel{\text{def}}{=} \mathbb{N} \to \mathbb{R}$. The weakness : if we sample everything we might find that Tacchini has his own idiosyncratic way of doing sunspots and wolf, and so the data we add to his entriee would not fit well with his methods of observation.
2. For each sunspots number do the same as above but go through the groups number! Again for each sunspots number find sift through all the data of everyone to find the best groups number $g \in \mathbb{R}$, i.e. the g s.t. it sits on top of the fitted normal distribution.
3. For each sunspots number, methods 1) and 2) above but fitted with only Tacchini's previous and future data. The weakness of this method is that there is not much data to go on... (relative to the first two)
4. One thing that worries me in the case of Tacchini is that this guy observers a lot of sunspots. I am looking at his entries right now and in 1870 he regularly observers over 200 sunspots, on 5^{th} of April 1870 he sees 302 sunspots. That said cross-referencing his data with other observers' from the same time seems to support this hypothesis. But the more sunspots he observers, the less data we have to make a nice normal distribution for each sunspots number, there may be sunspots numbers that only appear like 3 times in the whole database, how are we to do any stats on those. The answer is the following method for tying sunspots to groups and wolf number: for each sunspots number we link it to a wolf number by finding the best fit normal distribution for the likeliest wolf number to be associated with it. Then we give this value an error bar which is bigger the less data we have. Then we do a plot x-axis = sunspots number, y-axis = wolf numbers with vertical error bars. Then we do a line of best fit through the whole lot, try several models and do a chi-squared test. It might be worth getting rid of the small ones and only look at data where $s > 20$. Again we could do this for wolf directly and for groups then wolf.

- Before embarking on this adventure I will first endeavour to plug both of Tacchini's hole's that appear in the plots I made: 3 entire years are missing From Tacchini's data 1877, 1878 and 1881

- In 1881 the rubrics number 465 contains two data sets, one of them is entirely Ricco's and the second is from Tacchini and G. Millosevich. Nowhere int the data set is it indicated who saw what, so I looked and found that there was no Alias for Millosevich. This makes me think that he must be Tacchini's assistant or observing partner. Anyway we have a big gap in Tacchini's observations, what I will do is comment all of these observations "Tacchini and Millosevich". I did that and gave them a flag, then regenerated the Tacchini graphs and there is no more gap in 1881, what's more the data looks almost identical. There is one only slightly worrying difference and maybe I'm inventing patterns where I am trying to see them but the 1881 observations are on average a tiny tiny bit higher than the surroundings. Actually I will put a picture ://learn.freecodecamp.org/responsive-web-design/basic-html-and-html5/add-images-to-your-websitebfhere (put link to picture I will include in this report), I think this is infact a seasonal effect, there is more atmosphere...

- I improved the jupyter notebook "suspiciou sunspots plots.ipynb", made a cool colour scheme for the various flags, and you should definitely **check this out (link of tacchini pink patches with flag=6)**

2.4.10 Wednesday July 3

- Precisely the same thing happens to Tacchini in 1877 and 1878 but this time it is "Tacchini und G. De Lisa". I did the same as yesterday: changed the information so that it was no longer in the observer and alias but commented instead, and flagged it with flag 6

- Found outlier for Tacchini, ID=46145, Mitt 30 Page 410 Rubrics 279 date 1871-04-12. Error type = typo. For groups wrote 112 instead of 12. Okay I'll admit, I found it cause I was playing around with the graphs I was generating.
- I wrote some methods in graphs_helper.py mainly for helping me to display data in jupyter notebooks while avoiding over-saturating my jupyter notebooks.
- The reason I've gone off track from the github project objectives is frankly because I was getting board of scouring the manuals for ages, but my enthusiasm for this task has regenerated now and this is what I will do. Once the errors from the `sorted_greater_comments_list3` have been fully dealt with I'll throw myself back into hunting for errors via graphic visualisation. And perhaps implement that idea I had yesterday about doing some stats on the relationships between sunspots, wolf and groups numbers (if I do this I estimate it will take around 4 days, which is quite a lot considering I usually underestimate these things, since it is not crucial to what I am doing I am considering doing this in my free time on the weekend perhaps...)
- Modified some Quimby comments, there was two stars int the rubrics 706 which had no explanation, I modified these two comments '*=secondary telescope' and flagged them with flag 6. There is no point in making a new alias because so few of his measurements are made with his secondary telescope, perhaps we should just get rid of them...
- Realising that I was making inefficient use of my time, I am now going through all the data with COMMENTS='uncertain' where I have manually checked in the Mittheilungen manuals and changing them with the following query: `UPDATE DATA SET COMMENT='?',FLAG=2 WHERE FK_RUBRICS=XXX AND COMMENT='uncertain';`. Since all of these were flagged initially because I didn't know what to do with them, they all find themselves in the BAD_DATA_SILSO so I also execute `UPDATE BAD_DATA_SILSO.DATA SET COMMENT='?',FLAG=2 WHERE FK_RUBRICS=XXX AND COMMENT='uncertain';`. Once I have finished with these they will all be moved to the good database that selects things that have flag=2 and moves them.
- It may well be written at the beginning of one of the Mittheilungen, but I have not yet found what the '?' comment means. I have two theories.
  1. I noticed the question-marks appear in Mittheilungen where there are several observers, in-fact I have not yet found a rubric where there are question-marks but no second observer / telescope. So I suspect it might be that Mr Wolf (or whoever wrote the journals de Mittheilungen) is unsure as to who took the measurement. CORRECTION: I found one where there is no secondary telescope or observer - rubrics 779, observer = Winkler, mitt 90, page 326. There is only one telescope and one observer, yet there are still question marks.
  2. My initial suspicion before I noticed 1. is that the question-marks denoted observations that were made on a cloudy day - or the measurement was somehow obfuscated. Oooh! I think this is right, but still not sure for every observer... There is written rubrics 1081 Herm. Kleiner writes "? bedeutet schlechte Definition des Sonnenbildes" - *? signifie mauvaise dfinition de l'image du soleil*.
- On second thoughts there are alot of red comments, I think it will take the rest of today (the next 4 hours) to verify every one of them in the journals and take note of them. I will do this and then write a script that changes their comments and flags all at once.
- I updated DATA in all databases with `UPDATE DATA SET COMMENT='bad definition of sun image',FLAG=2 WHERE COMMENT LIKE 'mauvaise d%';`
- 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.

– In rubrics 1037 mitt 100 page 359, observer ‘Ricco - Mascari’, there are two data-points with comment ‘0 0’. I checked the values in the journal and they were wrong! There are corrections I made

  1. 1908-10-17 groups 0, sunspots 18, wolf 18 (impossible) → groups=0, sunspots=0, wolf=0, comment=”, flag=0
  2. 1908-11-17 groups 0, sunspots 18, wolf 18 (impossible) → Deleted - there is no observation made on this date in this rubric

– I spotted a missing value in rubrics 12904 observer Buser for the observation on 1931-03-10 whilst I was correction the previous day’s which was incorrect, so I entered it in. There may be more here in this rubrics.

– Now I finished looking at the red comments (I still need to change them and move them all with python, that should only take about 30 mins to copy down all the rubrics numbers into a list and write the algorythem that changes them appropriately, I will do it tomorow.) Right now I am looking through the last two pages of the comments sheet I printed and there are alot of blue ones where the comments are just numbers, these require my attention.

– I looked up all the blue comments on those last two pages in the journals here is a summary of my findings:

  1. Some of the numbers actually pointed to the real values of the data i.e. the data had incorrect sunspot values and the comments had the correct ones. I modified these appropriately. In these cases often there was no groups but there were sunspots.
  2. some of the number were the correct values but the data was also already correct, here I just deleted the comment and removed the flag.
  3. some of the numbers were very perplexing and I have no idea what they were doing there. I just removed the flag here.

– I am going home now but I leave on a cliff hanger: the blues are almost done and I am currently investigating the mystery of ‘x’ which appears in the data section in the journals of mitt 33 page 120 rubric 296. It was called to my attention by a comment also denoted ‘x’

– Actually I will save a new backup of the sql databases before I leave since I edited them quite a lot today.

2.4.11 Thursday July 4

– I did not find an explication for the mysterious ‘x’... Sachen has really got me here. I translated all the text surrounding it and nowhere do they explain why there is an x. Fow now I will leave it. It will probably stay in the `BAD_DATA_SILSO` database

– I have been looking into Carrington’s case and here is what I found. Everything from rubrics 303 is the total area of either ther penumbra or the umbra. I will investigate further.

– 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.

– Moved Secchi’s derived from `BAD_DATA_SILSO` to `GOOD_DATA_SILSO` using the method `derived.move`$_{7t}o_{g}o$

– For some reason the script `dealing_with_duplicates.py` has in it a bunch of method that are really more general that just dealing with duplicates. Because other methods inside the script depend on these I don’t want to delete them, so I copied the tree following methods into `db_transfers.py` : `mod_data_to_bin()` with helpers `transcribe_info_old()` and `transcribe_info_new()`

- I modified `db_transfer` so that you have the option to copy instead of swap
- Spoke to F. Clette about the possible conversion from the 'aire' to a sunspots number. Told me 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
- 4.1 here what is written in German and Italian, with a translation in English.
- Tomorow I will fix Carrington's data, but before doing that investigate all of these : `SELECT * FROM RUBRICS WHERE RUBRICS_ID IN (SELECT FK_RUBRICS FROM DATA WHERE FK_OBSERVERS IN (36,49));`
- Made a new backup of databases.

2.4.12 Friday July 5

- I was searching all of Carrington's data from different rubrics in order to see if I could find an overlap in time from data where there is recorded the 'air' and data where there is recorded the sunspots number. Some of the groups number seem to conflict...
- This is very perturbing, the groups number for rubrics 199 (Carrington) does not seem to agree with the groups number for rubrics 303 (also Carrington), however they are very similar. The only explanation I can think of is that in 303 Carrington writes number_of_big_spots.surface_area and in rubrics 199 he simply write groups.sunspots. I am still hunting for clues in the text 4.1.2. This theory is evidenced by the fact that the number of big spots is always bigger than the number of groups which you would expect were it true.
- The thing to do now is graphs. I will make graphs to try and figure it out.
- Found one typo in the data id=31372 and corrected it
- Made the notebook `carrington_investigation_groups.ipynb`
- Missing data: while trying to analyse carrington's strange behaviour I cross reference every date from rubrics 199 with dates in 1859 and 1860 from 303 and found 7 missing datapoints which I looked up in the journals and 4 of them were in there. So I hand-typed them into `DATA_SILSO_HISTO` and then copied them over into `GOOD_DATA_SILSO` with the method `db_transfers.db_transfer(dont_delete=True)`. They are :
  - rub 303 1860-08-07, ID = 206774
  - for 1860-10-07 and 1860-11-16 there is nothing in rubrics 303, perhaps Carrington only did the sunspots number those days
  - rub 199 1860-09-15, ID = 206775
  - for 1860-10-08 there is nothing in rubrics 199, perhaps Carrington only did the penumbra for this day
  - rub 199 1860-10-28, ID = 206776
  - rub 199 1860-12-10, ID = 206777
- I did a comparison of the groups number from rubrics 199

- I have a problem, in order to implement the modifications I would like to make to Carrington's rubrics 303 data I need the groups, sunspots and wolf all to be floating point numbers, right now in the database they are integers and I cannot modify them as such.

- I found a relationship between what is labelled groups and the actual groups number for rubrics 303, see `carrington_investigation_groups.ipynb`
- These are my suggested modifications : do as the author of rubrics 299 did for Secchi (4.1.1). However I can do better, he did not have the power of computers and he makes many approximations that are not needed today. For instance, he takes the mean from each 20 equations out of his set of 120, there is no need for this today. I will on the other hand draw inspiration from him on his model:

$$r = a \cdot (10g + b \cdot f) = 10a \cdot g + c \cdot f$$

  according to him it fits quite well.
- In order not to disrupt the database I have decided to sacrifice a smidgen of accuracy in order to keep using integers for the r, g and s (wolf, groups, sunspots). (f is the 'aire'). I will do the following things (see `carrington_investigation_wolf.ipynb`):
  1. Basing myself off of the relationship described by equation (1) I will do a least squares fitting as I did for the groups, to find values for a and b
  2. I will then do the same least squares fitting using a modified g which I have multiplied by a factor of $\frac{1}{1.0915}$ to see if the standard deviation is at all better for this group
  3. If the fit is not as good as hoped I will try several other models to see if I can find an equation (with maximum 3 or 4 degrees of freedom) which fits the data well
  4. From my model I will the deduce s, the corresponding sunspots number which, when combined with g gives r
  5. Using my newly found equations and constants I will then apply modifications onto the rest of rubrics 303
  6. I will then round g, s and r to the nearest integer and enter these into the database.
- Something Sabrina brought to my attention was that my training group is only 2 years long, during a maximum, this might mess with the results a little bit.

3 Comparaison du data avant et apres + visualisations

3.0.1 The original sql data tables format

Table 2: 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	

Table 3: 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	

Table 4: 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	

3.0.2 My new sql data table format

Table 5: 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	

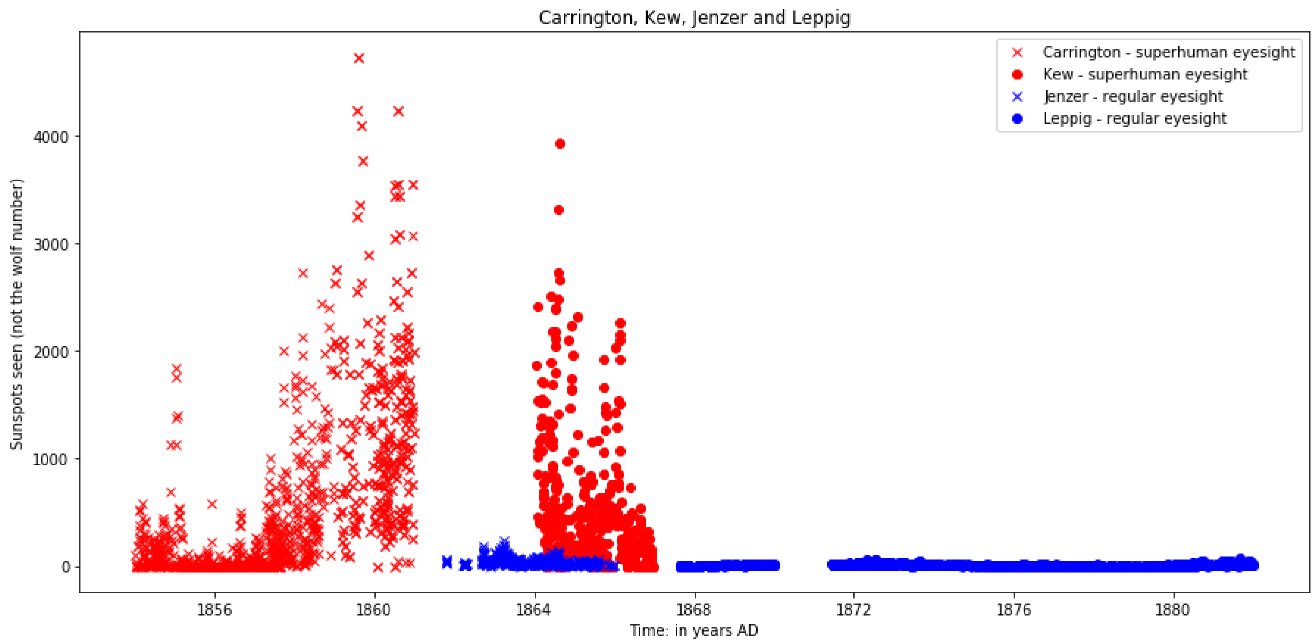

Figure 1: Carrington has great eyesight!

3.0.3 Graphs and visual representations

As you can see, there are some famous legends in solar science such as Carrington and Kew who were able to see over 4000 sunspots at once on a single face of the sun with their crappy telescopes that they had in the 19th century!

3.0.4 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 still were around ; 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

4 Miscellaneous

4.1 Converting the f ('aire')

4.1.1 rubrics 299, mitt 33, p 128 observer Secchi

German p128) Als Anhang ist ein "[It] Registro della macchie solarie osservate alla specola del Collegio Romano durante l'anno 1871" gegeben, welches die an einer Reihe von Tagen von Rom. Remiddi gezahlten Gruppen, anstatt der Anzahlder Flecken aber Zahlen enthalt, welche die von ihnen eingenomene Flache in Quadrat-Millimeter angenommen. Ich gebe dieselben in der gewohnten Weise, d. h so, dass die erste Zahl wie immer der Anzahl der Gruppen, die zweite aber jener Flachenzahl entspricht, - die der letztern gleichgesetzte Zahl endlish eine aus ihr nach untenstehender Formel berechnete, der Fleckenzahl moglichst entsprechende Zahl.

Secchi's Data from rubrics 299

p 130) Meine Relativzahlen basiren bekanntlich auf der Annahme, dass die Fleckenthatigkeit zunachst in der Anzahl der Gruppen, in untergeordneter Weise aber auch in der Grosse derselben ein Maass finde, und es wurde dieser Grosse von mir nur darum die Gesammtanzahl der Flecken substituirt, weil ich einerseits durch viele betreffend Vergleichungen gefunden hatte, das mit der Grosse der Hauptflecken meistens auch die Anzahl ihre Begleiter zunehem, also die Anzahl der Flecken annahernd jener Grosse proportional sei, - und es anderseits nicht nur zu zeitraubend fand diese Grosse fortwahrend zu messen, und (was bei den obigen Beobachtungen, welche nur die scheinbaren Flachen geben, wenigstens vorlaufig unterlassen wurde) auf ihr wahres Mass zu reduciren, sondern namentlich auch ein fur altere Beobachtungsreihen (denen sich gewohnlich die Anzahl der Flecken mit ziemlicher Sicherheit, die Grosse dagegen selten auch nur irgendwie annahernd entnehmen lasst) ebenfalls brauchbares Verfahren einfuhren musste. - Die in der obigen Reihe fur viele Tage, and welchen ich selbst Fleckenzahlungen gemacht, und daraus die Relativzahlen r berechnet hatte, gegebenen Flachen haben mir nun die Moglichkeit verschafft die Richtigkeit meines Verfahrens neuerdings zu prufen, und zugleich eine bestimmte Regel aufzustellen, um zur Erganzung meiner Register fur einzeln Tage ause den bestimmten Flachen die fur mich nothigen Fleckenzahlen annahernd zu berchnen: Bezeichne ich namlich die Anzahl der in Rom gezahlten Gruppen mit g, die bestimmte Flache aber mit f, so muss unter Voraussetzung der Richtigkeit meiner Annahme annahernd fur jeden gemeinschaftlichen Beobachtungstag eine Gleichung

$$r = a \cdot (10g + b \cdot f) = 10a \cdot g + c \cdot f$$

bestehen, wo a, b und c constante Factoren sind. Ich bildete nun 120 solcher Gleichungen, ordnete dieselben nach r, nahm je aus 20 das Mittel, und erhielt so die 6 Normalgleichungen

	$r = 10a \cdot g + cf$	r'	$r - r'$
1	$60 = a \cdot 37 + c \cdot 46$	62	-2
2	$80 = a \cdot 46 + c \cdot 61$	78	$+2$
3	$100 = a \cdot 60 + c \cdot 112$	109	-9
4	$120 = a \cdot 63 + c \cdot 117$	114	$+6$
5	$140 = a \cdot 78 + c \cdot 155$	143	-3
6	$160 = a \cdot 85 + c \cdot 187$	159	$+1$
	Mittlere Abweichung		± 5

aus welchen ich nach der Methode der kleinsten Quadrate

$$a = 1.41 \quad c = 0.21 \quad \text{sodann } b = 0.15$$

und somit fur die romischen Beobachtungen die Reductionsgleichung

$$r' = 1.41(g \cdot 10 + f0.15)$$

fand. Setze man in die Normalgleichungen diese Werthe fur a und c ein, so erhalt man die ihnen beigeschriebenen r', deren Vergleichung mit den r eine unerwartet gute Uebereinstimmung zeigt. Es hat also diese kleine Untersuchung die Berechtigung des von mir fur die Berechnung der Relativzahlen aufgestellten Principes in schonster Weise bestatigt, und mich anderseits ermuthigt in der obigen Beobachtungsreihe jeder Flache die nach der eben aufgefuhrten Formel berechnete Fleckenzahl beizuschreiben, - wobei ich naturlich in den paar Fallen, wo eine ganz geringe Flache eine Fleckenzahl ergab, welche kleiner als die Gruppenzahl war fur sie diese Gruppenzahl substituirte.

English p128) As appendix is given a "Register of sunspots observed in the mirror of the Roman College during the year 1871", which is the one on a series of days of Rome. Remiddi paid groups, instead of number spots but contains numbers which assume the area in square millimeters inscribed by them. I give them in the usual way, i.e. in such a way that the first number corresponds, as always, to the number of groups, the second, however, to that number of flats, - which endlessly calculates from the latter equated number a number which corresponds as closely as possible to the number of spots, according to the formula below.

Secchi's Data from rubrics 299

p 130) As is well known, my relative numbers are based on the assumption that the number of spots finds a measure first of all in the number of groups, but in a subordinate way also in the size of the same, and this size was only substituted by me for the total number of spots, because on the one hand I had found by many comparisons that with the size of the main spots mostly also the number of their companions increases, so the number of spots is approximately proportional to that size, - and on the other hand not only too time-consuming was it found to measure these large ones continuously, and (what was at least temporarily omitted in the above observations, which only give the apparent flat ones) to reduce them to their true measure, but especially also to introduce a procedure useful for older series of observations (from which usually the number of spots is quite certain, but the large ones, on the other hand, can seldom be taken out even approximatively) likewise. - The surfaces given in the above series for many days, on which I had made spot payments myself, and had calculated the relative numbers r from them, have now given me the possibility to check the correctness of my method recently, and at the same time to establish a certain rule in order to approximate the numbers of spots necessary for me to supplement my registers for individual days from the certain surfaces: If, for example, I designate the number of groups paid in Rome with g, but the certain area with f, then, assuming my assumption is correct, an approximate equation must be given for each common observation day

$$r = a \cdot (10g + b \cdot f) = 10a \cdot g + c \cdot f$$

where a, b and c are constant factors. I now formed 120 such equations, arranged them according to r, took the mean from each 20, and thus obtained the 6 normal equations

	$r = 10a \cdot g + c \cdot f$	r'	$r - r'$
1	$60 = a \cdot 37 + c \cdot 46$	62	-2
2	$80 = a \cdot 46 + c \cdot 61$	78	$+2$
3	$100 = a \cdot 60 + c \cdot 112$	109	-9
4	$120 = a \cdot 63 + c \cdot 117$	114	$+6$
5	$140 = a \cdot 78 + c \cdot 155$	143	-3
6	$160 = a \cdot 85 + c \cdot 187$	159	$+1$
	Mittlere Abweichung		± 5

from which I can draw the least squares

$$a = 1.41 \quad c = 0.21 \quad \text{sodann } b = 0.15$$

and therefore for the Roman observations the reduction equation

$$r' = 1.41(g \cdot 10 + f0.15)$$

found. If one enters this value for a and c in the normal equations, one obtains the r' attributed to them, whose comparison with the r shows an unexpectedly good agreement. So this small examination confirmed in the best way the validity of the principle I had established for the calculation of the relative numbers, and on the other hand it encouraged me in the above series of observations to attribute to each surface the number of spots calculated according to the just listed formula, - whereby I naturally substituted this group number in the few cases where a very small surface resulted in a number of spots which was smaller than the group number for it.

4.1.2 rubrics 303, mitt 35, p 241 observer Carrington

303) Warren De La Rue, Balfour Stewart and Benjamin Loewy, Researches on Solar Physics. Second Series: Area measurements of the Sun-Spots observed by Carrington during the seven Years from 1854 - 1860 inclusive, and deductions therefrom. London 1866 in 4.

German Ich ziehe aus dieser Abhandlung unter fortwahrender Berucksichtigung der unter 199 besprochenen Werkes von Carrington und der unter 129 aufgeguhrten schriftlichen Mittheilung derselben folgende Beobachtungen in der altgebohnten Form, nur dass die der Gruppenzahl folgende Zahl (analog wie bei den unter 299 aufgeguhrten Beobachtungen Secchis) nicht die Anzahl der Flecken, sondern die in Millionsteln der sichtbaren Sonnenhemisphare ausgedruckte Flache derselben Bezeichnen:

Durch Vergleichung der fur 1859 und 1860 gegebenen Flachenzahlen mit den in Nr. 199 von Carrington selbst fur dieselben Jahre und Tage mir mitgetheilten Fleckenzahlen, erhalt man, dass durchschnittlich 1000 Flacheneinheiten 24 Flecken entsprechen, und es darf dieses Verhaltniss ohne Anstand benutzt werden, um fur die wenigen Tage, wo das Fleckenregister durch Carrington'sche Beobachtungen erganzt werden kan, die Flachen in Flecken umzusetzen.

English I deduce from this treatise, taking into account the work of Carrington discussed under 199 and the written communication of the same discussed under 129, the following observations in the old-bored form, only that the number following the group number (analogous as in the observations of Secchi examined under 299) does not denote the number of spots, but the area of the same printed out in millionths of the visible solar hemispheres:

What follows is observations made by Carrington for the years specified with 'aire' numbers instead of sunspots numbers. [it seems someone had the same idea as me]

By comparing the surface numbers given for 1859 and 1860 with the spot numbers given in No. 199 by Carrington himself for the same years and days, one obtains that on average 1000 surface units correspond to 24 spots, and this relationship may be used without decency to convert the surfaces into spots for the few days when the spot register can be supplemented by Carrington's observations.

4.1.3 Rubrics 199, mitt 11-20, p224

Observations of the Spots on the Sun from 1853 XI 9 to 1861 III 24 made ad Redhill by R. Chr. Carrington. London 1863 (248 Pag., 166 Plat.) in 4.

German Dieses Ausgezeichnete, erst kurzlich nach Verdienen von der Pariser-Academie mit dem Lalande-Preise bedachte Werk meines verehrten Freundes erlaubt nach seiner Natur kaum einen Auszug, sondern ist zunachst als eine unerschopfliche Fundgrube zu betrachten, in der diejenigen Astronomen, welche sich speqiell mit der Vertheilung der Sonnenflecken, ihren Ortsveranderungen etc. Befassen, ein reiches Material an Zahlen unde Zeichnungen erheben konnen, - wie ja bereits oben eine darauf gegrundete Studie von Herrn Fritz mitgetheilt worden ist, wahrend eine die 'Concluding Section' betreffende Arbeit von mir in einer der nachsten Mittheilungen folgen wird. Dagegen mogen hier anhangsweise zur Erganzung der Nr. 129 der Litteratur die Fleckenzahlungen in den Jahren 1859 und 1860 nachgetragen werden, welche mir Herr Carrington seiner Zeit mittheilte, und die ich in der letzten Zeit neuerdings bei Ermittlung der mehrfach erwahnten 5 taggigen Mittel benutzte. Es sind Folgende:

English This excellent work of my esteemed friend, which was awarded the Lalande Prize by the Paris Academy only shortly after it had been earned, hardly permits an excerpt by its nature, but is first to be regarded as an inexhaustible treasure trove in which those astronomers who are particularly concerned with the distribution of sunspots, their changes of place, etc., can be found. As already above a study based on it has been shared by Mr. Fritz, while a work of mine concerning the 'Concluding Section' will follow in one of the next communications. On the other hand, to supplement the No. 129 of the Litteratur, the stain payments in the years 1859 and 1860, which Mr. Carrington informed me of his time and which I recently used in the determination of the repeatedly mentioned 5-day means, may be added here as an appendix. They are the following: