DATA_SILSO_HISTO
Log

Stephen Fay

July 19, 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

2 Equations

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

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

3 Log

3.1 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.

3.2 Friday June 21

- Today no-one was in the office in the morning so I didn't have access to the Mittheilungen journals and decided to start writing this instead
- at 10:20 I was let into my bit with all my notes and the journals and began 'searching the manuals' part of the project documented in the Github project linked
- been spending time writing in all the pink corrections, including typos
- started writing 'searching_the_manuals.py'
- wrote and executed methods : def_correct_typos_for_pink() ; pink()
- searching the manuals for all comments labeled 'uncertain' so as to figure out what is this word's range of meaning (wishing I had paid attention in German class)

3.3 Monday June 24

- 9.15 picking up from where I left off, I am currently scouring the manuals for any 'uncertain' data
- 10.40 came across some duplicate data, and mysterious comments... there are some stars '*' that signify a change of instrument but nothing is written. The annoying thing about the duplicate data is that it is coming from

- spent the morning making that duplicate finding and sorting algorithm, now I need to analyse the nature of the problem further. For each of the duplicates identify what kind it is, weather it's the same observer with the same instrument; if the duplicated data has for example the same rubrics numbers as each other ; if they record the same information (sunspot groups, sunspots, wolf number) ; if check to see if any clues are hidden in the comments of these duplicated data

- in searching_the_manuals i wrote : find_duplicate_observers() ; find_obs_id_by_date() ; find_observer_alias_by_i ; find_duplicates_data() ; write_greater_duplicates_data_text()

- i'm gonna go and delete some of the data so i will log everything in order to be very careful

3.4 Tuesday June 25

- 9.45 I have decided to start making modifications to the database, this is risky business - I don't want to have the blood of Galileo's data on my hands, in a few seconds I can destroy hours upon hours of one of my predecessors' work. Which would be a shame. This is why I am creating a new table in both the old and the new database that will serve as a rubbish bin, so that I simultaneously copy and delete some data. The data will be copied and destroyed in the same script but the coping will come *before* so if there are bugs nothing will be lost. First I will back up the databases as they are.

- While making the rubbish bids for DATA_SILSO_HISTO if found that DATA_DEV was non-empty, it contains data which claims to be observations made by the grandfather of this series - Rudolf Wolf. Only the observations are dated January 1600 - Galileo's time, 216 years before Wolf's birth! And so I renamed the DATA_DEV to RUBBISH_DATA and added the flag column, leaving those four observations inside where they probably belong...

- Wrote a new script to deal uniquely with deleting the duplicates
  - finished writing move_data_to_bin and delete_entered_twice_duplicates
  - executing delete_entered_twice_duplicates()... done
  - finished commenting these data points in rubbish data in both databases

3.5 Wednesday June 26

- wrote a new method in db_edit for appending comments rather than replacing them

- wrote *unreasonable_sn_flag()* a method that takes a look at the groups number and the sunspots number of each entry and determines if it's realistic or not. I decided somewhat arbitrarily that if the groups number was higher than 30 it would be flagged with flag 4, if the groups number was higher than 60 it would be marked with a 5 this is beyond unreasonable. I did something similar for the sunspots number $sunspots > 100 \Rightarrow flag := 4$ and $sunspots > 250 \Rightarrow flag := 5$ (see the flags section **??**). The method was executed and ran without a hitch (after a bit of debugging)

- just set another 212 flags for putting things in the bin. There are still 4000 pairs of duplicates that need attending to but considering i started with 14000 that's not bad... Some of the duplicates may be left as they are. Also i figured out that i had flagged some which just had a 0 sunspot number and so i went and unflagged them.

- i spent alot of time scrutinising what i had flagged, rereading my scripts, seeing that I've been using really inefficient algorithms, checking things are in the right place. And wondering how on earth many things ended up with the flags they ended up with.

- something that has been annoying me in this search is I can't seem to be able to determine what is an unreasonable number of sunspots that can appear on the sun, because many many observer record having over 250. This is why I will start using graphical tools to help me figure all of this out. I will make the graphs in a jupyter notebook.

3.6 Thursday June 27

- first thing i did today is to go though and look at lots of the flagged data from yesterday on the mysql databases
- Panic! While searching I came across a big problem. Many of the datapoints are labelled '*' in the comments, this corresponds to when there is an asterix in the Mitteilungen journals, but here's the twist: the star is usually a reference to the fact that there is a change in the observer's telescope to his/her secondary lunette. This is written nowhere is many cases in the digital database! This is a new task I must take on am I to accomplish my mission here:
    1. Correct all the comments so that they display useful information i.e. '*' — '* = 8 cm Oeffnung mit 64-facher Vergrosserung und Polarisationshelioscop'
    2. When you tackle the 'Creating New Aliases' part of the project (see my Github - username = dcxSt - project sun)
- I found there I had flagged all the mysterious '*' comment 1 and that none of them have found their way into my pristine database GOOD_DATA_SILSO and so I went through each of them individually and wrote the changes that I implemented in the python method 'correct_asterix_comments()' in script 'db_homogenise_comments.py'. This took quite some time and included translating German with my good friend google-translate. This corrected about 250 data-points' comments (i didn't bother to count)
- after a long search of the data in GOOD_DATA_SILSO with FLAG=3 which have no superior duplicates I found that these were infact correctly flag and that their double had not yet made it into my new database because there were commented (usually with an asterisk *) so i moved them into the bin
- Found a bug in move_flag3_to_bin() which may have been causing some of the perplexing problems I had earlier
- I ran the method 'flag_many_duplicates()' many times using the duplicates text files I'd made earlier for inspiration to change it subtly so as to catch those sneaky no good duplicates!
- IMPORTANT: I just found some data which has been written in in the wrong year. In rubrics 820 students have mistakenly typed in the data for Wolfer in the year 1900 and written it in the year 1899.
- In the folder duplicates/3 2019.06.27 I am writing the file corrections_needed_handwritten.txt which outlines the corrections which are to be made to the data if we want to solve some of these duplicates.

3.7 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.

3.8 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_rub` 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

3.9 Tuesday July 2

- exploring ways of creating visualisations that will help me catch some of the suspicious data-points. Right now 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. Looking closer at the picture, 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)**

3.10 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.

3.11 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`$_{7t}o_good()T$

- 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
- 5 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.

3.12 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 5.2. 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 data-points 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 (5.1). 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.

3.13 Monday July 8

- I have been editing my 2 reports in the morning
- `carrington_investigation_wolf.ipynb` Realised that modifying the groups was useless because I would be dividing them 1.01, so I'm skipping step 2 (last Friday)
- Judging by the visuals I don't need to do step 3, also it would be risky because as Sabrina said I don't have any data close to 0.
- Plotted carrington's data
- Lunch
- Created new rubrics id (fk=1010) in `DATA_SILSO_HISTO`, this one has the same rubrics number as fk=175, `RUBRICS_NUMBER = 303`, but it is for storing the derived carrington.
- Pickled the dictionaries used in carrington_investigation_wolf.ipynb in cause they become inaccessible as I change rubrics 303 in `GOOD_DATA_SILSO`
- My scripts are getting very messy to work with, this is slowing me down alot. Once I have finished with Carrington I will start making that list of all the scripts and what they contain.

- Moved rubrics 303 from DATA to RUBBISH_DATA in GOOD_DATA_SILSO using db_transfers.move_data_to_bin_
  This required some very slap-dash patching. I'm going to look up what sort of diagrams there are out there to keep track of what is in each of my scripts, this is not very professional...
- I realised also that about half the methods I write I only write once, the reason I wrote them as methods is in part to keep track of every operation I performed but it is also annoying because I end up with a bunch of methods that are really in the wrong place. These ought to have been written in a jupyter notebook, or deleted.
- Successfully inserted the derived 303 into DATA_SILSO_HISTO with the rubrics id = 1010 instead of 175 using db_edit.insert_old_format(). No apparent errors.
- Copied all those from rubrics id 1010, rubrics number 303 into GOOD_DATA_SILSO using db_transfers.db_tra
- I realise that there are duplicates now in GOOD_DATA_SILSO, sooner or later I may bin the derived data from rubrics 303 that overlaps with the 199 but I will first seek consultation from Laure. In-fact I think the wisest thing to do might well be to split carrington into two observers : 'Carrington' and 'Carrington derived'. But this creates an imbalance in the sunspots number because we are lending extra weight to Carrington for the years 1859 and 1860.
- I just spent the last half hour cleaning up the databases after a jupyter notebook imported an old version of on of my scripts which put into the rubbish bin all of the data with comment like '%erive%' in GOOD_DATA_SILSO...
- Just realied that I have entered in all the carrington's with the same date!
- I fixed it with SQL commands, it surprisingly didn't take any time at all.
- Saved the databases

3.14 Tuesday July 9

- Before moving back to the comments, I'm going to apply the same fix to Kew that I did to Carrington. The bad news here is that I don't have any real data for Kew that I can compare his penumbras to. The good news is that the author of rubrics 306 already did the work for me and has found $a = 0.763$, $c = 0.032 \Rightarrow b = 0.042$

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

- Opened new jupyter notebook kew_derivation.ipynb
- Imported and changed Kew's data according to the formula provided.
- Pickled Kew's penumbra data kew_penumbra.pickle as well as his corrected data kew_derived.pickle
- Made new rubrics id (fk=1011) in DATA_SILSO_HISTO for differentiating Kew's derived data from the penumbra data which I will not just scrap (alot of hard work was put into getting those).
- Checked that none of Kew's data had found it's way into BAD_DATA_SILSO
- Moved rubrics 306 from DATA to RUBBISH_DATA using db_transfers.move_data_to_bin() in both the databases DATA_SILSO_HISTO and GOOD_DATA_SILSO.
- Inserted the derived 306 into DATA_SILSO_HISTO with rubrics id = 1011 instead of 177 using db_edit.insert_

- Copied all those from `DATA_SILSO_HISTO` with fk_rubrics = 1011 into `GOOD_DATA_SILSO` using `db_transfers.d`
- Saved databases and commit to github
- Lunch
- Wrote `create_readme.py` to automatically write a long readme that tells you what is in each python file as part of an initiative to make my work available to people other than me.
- Cleaned up the log some more so that Laure can hopefully make some sense out of it (fingers crossed)
- Finnished condensing the log
- I just had a great idea: In the spirit of forward thinking start designing the way you are going to present the changes you have made to the data, brainstorm:

  1. Explanation of new format
  2. put the log after the tables and figures
  3. Section 2 should not be called setup call it something that makes more sense, Section 2 should be for these things i am listing here
  4. Pie chart with the number of data points that have remained the same and the ones I have changed and the ones I have added
  5. Pie chart that shows what kind of modifications I made
  6. Review the explination of the flags, make it more complete yet still concise
  7. More figures that demonstrate how I have modified the database
  8. A concise list of all the things that were wrong with the database and the changes I made, this can be put in a conclusions section that I will make hereafter
  9. Should probably get rid of the thought repo

3.15 Wednesday July 10

- Today was bring your Arnaud to work day. Sabrina gave Arnaud and I a tour of the observatory in the morning and I showed him what I was doing and he explained to me some cool quantum stuff as well as some EM stuff and some add cal, divergence and curl stuff...
- Unfortunately I didn't do any work, I will get cracking at lunch time and stay until 18.00.
- After lunch I had a look at the Github project manager and realised that Carrington and Key were only an intermission in searching the manuals, and that before I have another look at some diagrams and graphs I must finish scrutinising the greater comments list 3 (i.e. 'searching the manuals')
- I went back to Sachen's mysterious comment = 'x'.
- I just realised that 'sachen' means 'things' and that 'Niedersachsische neue Zeitungen von gelehrten Sachen auf das Jahr 1730. Hamburg in 8' means 'Lower Saxon new newspapers of learned things to the year 1730. Hamburg in 8.', and that the existence of the observer with alias 'Sachen' is most likely the student who typed this in having miss-interpreted the cryptic German capitalisation of Random Nouns, in rubric 296.

- There is however only data from 1730 in this rubric and I think it best to leave it be. The mysterious x is still just as mysterious, It either means 0 sunspots and groups, or missing observation ; if it's the former they should have put 0, and if it's the latter then why even bother entering in the date... Probably safer to throw away than to keep, this is something I would consult Laure about but I fear she is very busy at the moment.

- After some deliberation I decided to change the flags of these to 6, and let be these ones in `BAD_DATA_SILSO` where they shall remain, and for the ones in `DATA_SILSO_HISTO` to just set flag to 6.

- I came across my friend Carrington while scouring the blue comments (blue is the color I use for miscelaneous comments), and looking at his data I thought more about what I should do with him, and I think more and more now that it is a good idea to separate out Carrington's observations into two distinct aliases 'Carrington' - for rubrics 199 and 'Carrington-derived' - for rubrics 303. As for Kew perhaps renaming him to Kew derived.

- Oh shoot I just remembered that I had planned to make a little summary of what I was doing for Laure yesterday. I will do that now in a short email. Done

- I implemented the changes, there is now an alias called 'Carrington derived' fk_obs=192

- Some of Secchi's data is marked with a 1/2 in the comments

- It seems I have just uncovered some more sunspots numbers that need deriving. These ones appear to have been hiding from me and staying under the radar due to the fact that they were measured around the years 1877, this is near a minimum so the 'the column 'Area mm quadrati' in which one unit should correspond to 21.56 millionths of the area of the solar disk.' could easily be mistaken for real sunspots number. Good thing there was written 1/2 in the comments

- First I cross referenced with rubrics 299 to see if the scales were the same.

- rubrics 299:
$$f = \frac{1}{46352.5} = 2.157 \cdot 10^{-5} \quad \text{solar disks}$$

- rubrics 375:
$$f = \frac{21.56}{10^6} = 2.156 \cdot 10^{-5} \quad \text{solar disks}$$

- Yes they are the same! So I can use the same conversion rates that are derived in rubrics 299 for rubrics 375. Unfortunately, because of the way the wolf number is calculated it is not the case that the correction gets more accurate with smaller wolf numbers, infact I suspect that the inverse is true and that the smaller the aire, the bigger the σ

- Typo in rubrics page number which I corrected, there is not page 319 it goes straight from 290 to 391, then all the way up to page 474 and over to 375 and from there back up the whole way to 474. Anyway I spent ages looking for page 319 for rubrics 319 and eventually found it on page 274, where I found a post-it that I had already placed there... I corrected this in the databases.

- Anyway rubrics 319 seems to be mainly sunspots observations, indeed there are a lot of blank bits.

- Rubrics 334 (also Secchi, all these i'm checking are Secchi) has no sunspots numbers, there are only 6 out of about 120 entries that have a sunspots number and they are all either 0 or 1. All the rest have only groups, for the year 1874. I think for these ones what needs to be done is put them in the database but with flag = 9

- Same for 363
- It's a similar sort of story for rub 339 (1875). However I have just thought of a potential problem. If we are calculating averages for a long period and we decide to only keep Secchi's data when there are no sunspots this will bring the average down from it's real value. The author of the rubrics 339) has the following suggestion, just give them fake wolf numbers, here I literally transcribed from the book:

Eine Reihe von Vergleichungen ergab analog 293 die corespondirenden Werthe:
A series of comparisons, analogous to 293, revealed the corespondirenden Werthe:

g	0	1	2	3	4
r	0	14	28	41	54

- Naturally I looked to see what was up with rubrics 293 and this one seems to belong to a certain gentleman by the name of Johann Friedrich Julius Schmidt who also forgot to observe the sunspots in 1872. So I translated the text from his rubrics to see if there was anything of interest. There was - here is a summary 'what a pity that he didn't measure the sunspots will his small telescope and only did it with his big telescope. Here is a table that we can use to estimate r based off of g' (see rubrics 291 5.5):

Conversion table from rubrics 293

g	0	1	2	3	4	5	6	7	8	9	10	11	12	13
r	0	20	38	54	69	83	97	110	123	136	148	160	172	184

- The data from the two rubrics above needs to be flagged and commented appropriately, perhaps it is wise to deal with the 8 and 9 flags all at once, in which case I will bring that task forward one step and do it right after finishing with the rest of Secchi's misbehaving data.
- The same thing happens for Secchi in rubrics 363, there are no values for the vast majority of the sunspots. This is for the year 1876, I suggest here we do the same as is suggested for 339
- Tomorrow, first thing you do when you come in is deal with all of Secchi's rubrics that require attention. Once this is done, probs by lunchtime if you're being as careful as you should be, look at any other flags 8 and 9. The anoying thing here is that some of secchi's really require 2 flags - derived and missing wolf or sunspots. I guess they will end up being derived, but first flag all of those with missing groups / sunspots appropriately.
- Saved a backup of the sql databases and commited them to the github.

3.16 Thursday July 11

- Dealt with rubrics 375, see `secchi_derivation.ipynb`

1. Starting with `DATA_SILSO_HISTO` - Flagged all from rub 375 flag = 7 (mysql terminal)
2. As a temporary placeholder I flagged all those with comment '1/2' with flag = 6
3. Made a new fk rubrics (fk = 1012) but with the same rubrics number (mysql terminal)
4. Moved the ones in `GOOD_DATA_SILSO` into the rubbish bin using `db_transfers.move_data_to_bin_only_`
5. Made a derived copy of all these from rub 375 but using the formula

$$r = a \cdot (10g + b \cdot f) \qquad a = 1.41 \quad b = 0.15$$

   This required some careful maneuvering in order to incorporate the 1/2 into the derivation.

6. Added to the half comments from rub 375 'aires'

7. Inserted them into `DATA_SILSO_HISTO` using `db_edit.insert_old_format()`
8. Double checked there was nothing in the other two databases
9. Flagged all the non-derived ones 7 and added to the comments so that they would have '1/2 area', so as not to lose the 1/2 information using `db_edit.add_to_comment()`
10. Copied the data into GOOD database using `db_transfers.db_transfer()`

- Backed up databases
- Dealt with rubrics 363, 339 at once as they all need the same treatment
  1. Check each of the rubrics to see if any had groups which exceded the number four as this is the biggest number I am given in the table of conversion derived by the author of rubrics 339
  2. Unfortunately they do in rubrics 319 and rubrics 334 (it goes up to 6), here I translated what the author of the rubrics has to say and the message is merely that of disappointment that Secchi didn't record the data the way he wanted.
  3. I really don't want to do anything rash when handling the data from 319, on the other hand we cannot really do anything with the data without a wolf number... Perhaps the thing to do would be to loop at data from all the observers from that year who have actual data, take the mean for each of the group : wolf correspondences and use those for 319. Or else continue adding 7 each time... No for now I leave 319 as it is
  4. Flagged all from rubrics 363, 399, 334 that have missing sunspots values with flag = 9
  5. changed data in `DATA_SILSO_HISTO`
  6. deleted from `BAD_DATA_SILSO` and added to `GOOD_DATA_SILSO`, 336 datapoints
- What do do with the two remaining Secchi rubrics? (334, 319)
- I kept going with the comments list, there I found that Ferrari's rubrics 425) where he submits data from 1879, the sunspots numbers are in-fact area measurements,
- Found out that Ferrari has only 2 rubrics associated with him
- 425, 389 (1878). I incidentally also translated rubrics 389 and this confirms that those are also areas.
- For now I will flag Ferrari's data, flag = 7.
- Corrected a typo where someone forgot to enter the groups number.
- Corrected a typo for ID = 61974
- Replaced flag=6 where there is strange comment in Mittheilungen, the datapoint is fine but it's ambiguous as to weather it was recorded on the 31st of January or the 1st of February 1885.
- For some reason in the rubrics 811 there are two entries for 1899-07-30, '1.1' and '0.0?'. In the database I flagged it 6 and transferred it to `GOOD DATA SILSO`
- More strange things happen in Kleiner's rubrics 811. The dates start to break chronollogy, the person doing the digitalisation typed them in with his best guess and left a comment, there really isn't anything else to do. I will flagged theses with flag=6 and transfer them. (there are only 3 which I what makes me think that whoever wrote these into the Mittheilungen must have been tired or distracted)
- Verified a some more of the unticked blues
- Backed up databases and pushed to github

3.17 Friday July 12

- Rubrics 828 was mistakenly attributed to 'Konkoly', the actual observer is 'Scharbe' and the mysterious 'P' comment is for the days when 'Pokrowsky' who doesn't yet have an alias. I corrected the primary observer and made the comment more explicit in each database.
- After some thought I changed the meaning of flag 8 from 'null groups'- which I am never going to use to 'bad definition of the sun picture', this brings it's meaning closer to that of flag 2, they may infact be the same, but the Mitteilungen makes this ambiguous.
- Modified some more of the blue ones, ther were a few typose here and there, also including rub 759 where I replaced the '*' with a '?' - it was written - '* means uncertain number' in the Mitt, I didn't want to put in the flag=8 section so I judged that this comment we closer to a flag=2 '?' comment.
- Finished initial treatment of the blue section.
- Though I said yesterday to Laure that I would start fixing Wolf and Wolfer on Monday, I think it may be wise to first deal with the orange section as doing this may clear up Wolf and Wolfer
- Objectives for the day
  1. Finish dealing with all the red ones (already ambitious)
  2. Make a plan of attack for how to deal the oranges (30 mins)
  3. Take half an hour to examine Wolf and wolfer
     - Plot them and examine plots
     - Find all the rubrics they feature in
- Edited `create_readme.py` to make the README have a nicer feel
- Found that Adams in rubrics 167 was missing a load of data, 624 data-points in all. (all of them 'none observed')
- Added them in using the method `add_to_adams()`
- Found some rogue data-points (typos) which I corrected by hand from rubrics 451
- Dealt with red ones
  - The ones marked 'mauvaise def img du sol' or similar have already be dealt with earlier (on wed 3 jul see 3.10)
  - Made a list of tuples (fk_rubrics,comment) of reds that need a flag=2 and comment='?' in the method `red_uncertain.question_mark()`
  - Applied these changes using `flag_and_comment_question_marks()` (there are now 927 data marked 2)

3.18 Monday July 15

- Continuing on from last friday
  - Transferred all flag = 2 data from `BAD_DATA_SILSO` to `GOOD_DATA_SILSO` using method `db_tranfers.tra`
  - Did a little check through the terminal that everything had transfered properly
- Wrote some methods in `graphs_helper`

  - `data_by_obs_alias_histo()`
  - modified `get_data_by_obs_seperate_flags()` so that it can access any database
  - wrote `display_compare_observers()` and tried to make a smoothed line but ended up just wasting time, I couldn't figure it out because there was some tricky stuff going on with the weird date formatting imported from mysql...

- I've been looking at some graphs of Wolf and Wolfer (which I invite you to admire here on the report 5) and there are several things I can say about them, reading off them things I can investigate in further detail tomorrow.
  - Wolf
    * 6 observers include wolf, some of them need merging (orange) and some need separating
    * Investigate the miscellaneous data to the left of the bulk in the mid 1840's. Is it real data or is it bad, should there be data between it and the rest?
    * I suspect that some of the Wolf - S - M observations are in-fact Wolf - P - M and vice versa, investigate the rubrics
    * Figure out if some of the data is duplicated especially when it comes to Wolf - S - M and Wolf - P - M, my duplicate detection algorithems only looked at the alias to tell if they were from the same observer so there might well be some hidden duplicates here
    * Do some more plotting involving the people mentioned in the wolf-aliases to see where they come in to the picture. They are definitely important specially because the point of seperating out the aliases is to find some way of using the 1890's overlapping period to callibrate the wolf number and wolf indices.
  - Wolfer
    * Investigate the 1920 hole
    * See if the rubrics from the two extra aliases can be separated out using the comments
  - Both / other
    1. Turn this brainstorming into a big todo list on github
    2. Plot all the aliases of all the other guys who share things with wolf and wolfer
    3. Make a list of every rubric these guys appear on

- Saved databases before commit and push.

- Reminder, tomorrow send a mail to Laure asking some of the questions I have about wolf and wolfer and specially about the calibration technique used and why the wolf-wolfer overlap is so important

3.19 Tuesday July 16

- Expanded on yesterday's brainstorm of things to look into for wolf and Wolfer

- Made a new alias for Carrington called 'Carrington penumbra' only in `DATA_SILSO_HISTO` because I judged it to be best. Then I moved all that was not in rubrics 199 but was still under obs_alias 'Carrington' into this one.

- Make a list of observers involved in Wolf-Wolfer affairs (and plot them), especially during the 1890s overlapping period

  - Just found mistake, (while making a list of aliases) I plotted 'Schwabe' and 'Schwab' and confirmed in the mitt my suspicion that these were indeed two different observers judging by the fact that 'Schwabe' has most of his data in the early to mid 1800s and there two rubrics mistakenly attributed to him in the 1900s 25 years after his death.
  - So I fixed this error in all the databases (478 data-points affected in each `histo` and `good`)
  - Curiously I found that the alias 'Wolf P-M Meyer Weber Schimdt Leppig' (yes I this is the correct spelling, for some reason it's Schimdt instead of Schmidt) only has a single data-point ID=46098. I wanted to look up the rubrics in the mittheilungen but the data-point has no rubrics. However it is commented 'Wolf'. So I looked at other data with the same date and there was another data-point ID=25534 with exactly the same values but attributed to the observer 'Wolf - P - M'. I deleted the data-point `..._HISTO` and in `BAD_...`. And then I deleted the observer since it had no more data.
  - A quick search revealed that 85 observers (or at least 85 aliases) made observations between 1850 and 1900. And 44 observers made observations between 1876 and 1895 - the period of overlap between wolf and wolfer. Maybe it's a good idea to plot them?? First I will focus on finding out what is wrong with wolf and wolfer.
  - Upon examination of the wolf-wolfer transition period you can see that the wolf P gets bigger proportional to the sunspot number, my guess is that the improvement my superiors are trying to make has to do with this effect. If anything the wolf p m should go down as wolf looses his eyesight toward the end of his life, not up.
  - You can really see that 'WOLF - S - M' and 'WOLF - P - M' see differently in the wolf graph (5)

- I downloaded a copy of the online sunspot data to use for comparison's with Wolf's allegedly degrading eyesight etc.

- Been looking into smoothing functions, I was at first interested in the Savitzky-Golay filter and tested it out a bit but I realised this was not ideal because it didn't take into account the fact that the data was not evenly spaced out... So I decided to go for the lowess function from the `statsmodel.api`. Found it on this stack overflow page. Really there is no point in being picky here because I'm not doing any statistical analysis, the aim of smoothing the sunspots number is to be able to better identify stray data, so so-long as it looks nice I'm happy.

- Wrote some functions and helpers (mainly in `graphs_helper.py`) to implement this.

- Backed up databases as always

3.20 Wednesday July 17

- There was an (non-alcoholic) appero and football today at lunch-time

- Corrected Adam's data from rubrics 167 - some of it had the year wrong, '1882-...' instead of '1822-...', I corrected these ones manually (20 data-points in each good and histo)

- I have been scouring the data using all sorts of plots, now I have an idea of how many observers are there and how chaotic everything is. I will focus on Wolfer's data first.

- 'Billwiller et Wolfer' summary of findings

  - There are 4 rubrics in the Mittheilungen which are shared between Wolfer and Billwiller: [366,345,386,411]

  - rubrics numbers 386, 366 and 345 have been typed into the database correctly - that is, where Billwiller's observations are entered under alias 'Billwiller' and Wolfer's observations (marked with a * in the mitt) are entered under alias 'Wolfer'
  - rubrics 411 the Mittheilungen is ambiguous I have to asked a German friend for help and am pending his reply. The unmarked data is either a combination of Wolfer and Billwiller or just Wolfer, the reason this is unclear is because the data marked * is Billwiller but it is phrased in such a way that makes me think he was using a secondary telescope. Also in all the other rubrics it was the opposite - Billwiller's data was unmarked.
  - The good thing is that there is nothing else wrong. If it turns out that 'Billwiller et Wolfer' was in-fact 'Wolfer' all along then the change required is easily done, and we can be rid of 'Billwiller et Wolfer'
  - I heard back from my German friend Theo on the 18th and this is what he had to say 'If its an old text Hlfsmittel is Hilfsmittel so not sulfur medium ^^ and yeah the text doesnt explicitly say that they made any observations collectively, so i think the unmarked obs are just from Herr Wolfer and the marked ones from Herr Billwiller'. This is great, see Thursday July 18 for the changes I made

- 'Wolfer, Mooser ' (there is a space after Mooser) - summary of findings
  - Mooser doesn't even seem to exist. I will correct tomorrow

- Something really bad just happened, I over-wrote the database I was working on with an old version of the database about 2 hours ago and just realised now because the old version doesn't have flags. This means all the editing I did today will go to waste, luckily I have been backing them up every day.

- Dropped each database

- Loaded yesterday's backup into each database

- Luckily also I only spend about 30 mins editing the database today so that's only about 1h lost work...

3.21 Thursday July 18

- Modified elements from the rubrics 411 to make the Wolfer ones belong to Wolfer, I left in the comments of the ones observed by Wolfer comment = 'Billwiller et Wolfer'

- Transferred those with flag=0 from `BAD` to `GOOD` databases after having changed the rubrics 411 ones. (they were in `BAD` because they were missing a rubrics source apparently, but this doesn't strike me as very important)

- Corrected Adam's data from rubrics 167, now that this is done (again) I will backup the sql databases before messing with them again.

- Composed and sent an update / ask-for-help email to Laure about what I am currently doing

- Sql databases saved in `2019-07-18_moorning` in the sql_backups... directory, commited and pushed

- Olivier kindly copied the tables from the original database for me onto a universal serial bus flash drive which I will

- Upon inspection of the .sql files I figured out why I lost all my data yesterday, I had tried to import the old .sql file into a database called `ORIGINAL_DATA_SILSO_HISTO`, but this didn't work because the .sql backup when run executes a bunch of commands to create the info in `DATA_SILSO_HISTO`. To solve this problem I went into the sql files Olivier had given me and changed the front sections so that they would insert all into `ORIGINAL_DATA_SILSO_HISTO`. (the reason I want the old database is to compare it with the new ones I am making and modifying to track the changes I am making and also be able to present my work to other people by showing some nice visualisations of how the data is different)

- I correctly imported the old data, but I am kicking myself now because I just re-deleted the databases that I am working on and also found out that I hadn't backed them up properly 3 bullet-points ago when I thought that I had. So I'm not feeling too good because this means I will have to correct all of Adams's data all over again as well as all of the wolfer and billwiller changes To give you an idea, corrected Adams data involves typing in 40 sql queries that are each about 20 words long, so once I correct them for the third time I will have spent $40 \cdot 20 \cdot 3 = 2400$ words on Adams. This is a massive pain in the neck. I guess some things you gotta learn the hard way...

- Reloaded the backups from Tuesday 16 into the my pc. (where would I be without them)

- Okay here I go once more:

  - Corrected Adam's data from rubrics 167 - some of it had the year wrong, '1882-...' instead of '1822-...', I corrected these ones manually (20 data-points in each good and histo)
  - I am not doing that again so I just saved a backup and checked that it was legit in a directory named with today's date +'_NOON'
  - Modified elements from the rubrics 411 to make the Wolfer ones belong to Wolfer, I left in the comments of the ones observed by Wolfer comment = 'Billwiller et Wolfer'. And un-flagged all the data from rubrics 411 in `BAD_DATA_SILSO`
  - Moved them using `db_transfers.transfer_flag_0()`. Had to patch `db_transfer()` method in order for it to ignore the fact that the data does not have any rubrics_source_date information

- I got a reply from Laure which was very good, she gave me some wonderful ideas including ideas to measure Wolf's drift which sounds like good stuff.

- I've been rereading the Clette et al. 2015 paper about the recallibration of the time-series

  - In section 3.2 it explains the 2-step iteration technique for calculating more accurate k factors for each of the observers works as I understand it like this:
    * (Step 1) Choose an initial 'pilot' station A
    * Compare all the data that station B has in common with the pilot station and see by what factor k it is on average smaller or bigger than A's observations for all the dates where they both make measurements
    * Correct all such stations B with this the k factor
    * With all these initial correction factors build an average S_n and G_n number
    * (Step 2) Use this fabricated S_n and G_n numbers to calculate new and more optimised k factors
    * use these new and optimised k factors to generate a new S_n and G_n
  - I had a quick brows of the 2014 paper as well as the 2013 paper where the backbone method is explained and I obviously don't understand it as well as I should before starting to suggest things but I have a question which is why do we stop at 2 iterations, why not continue iterating until we reach some stable equilibrium point.

3.22 Friday July 19

- Wrote `graphs_helper.display_wolf_drift()` which was non-triveal to write but I'm not sure how useful it is
-

4 Figures

This section is a subset of the section in the log

Figure 1: Tacchini

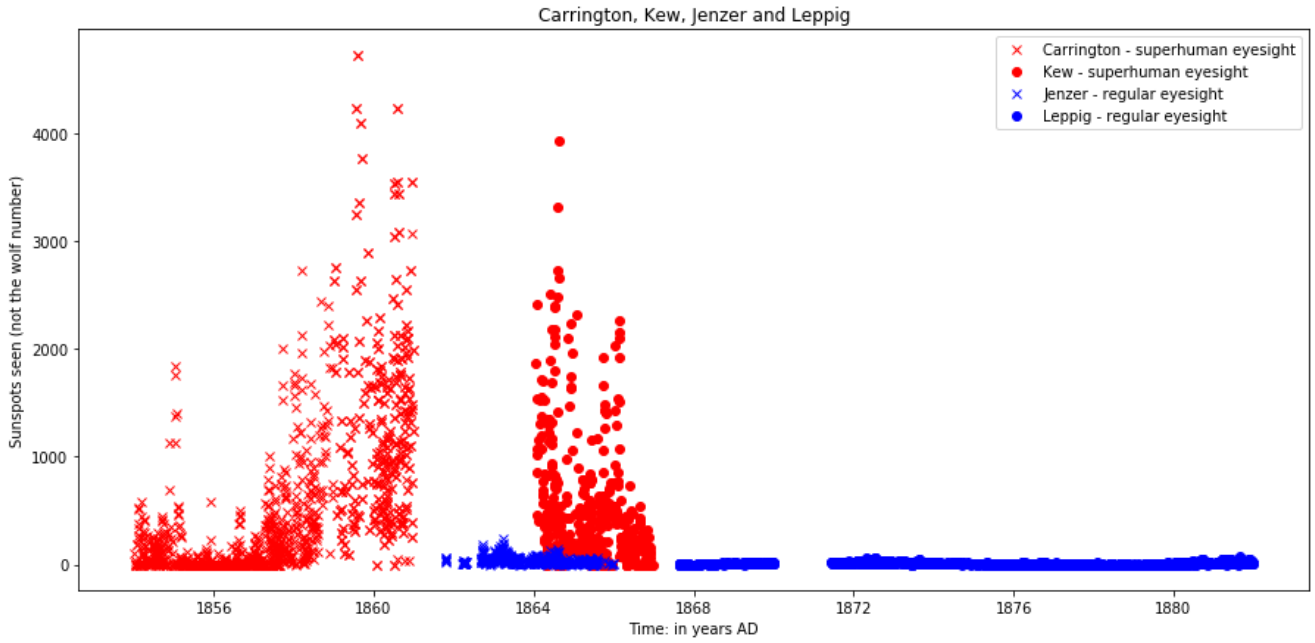

Figure 2: Carrington and Kew - input penumbras instead of sunspots

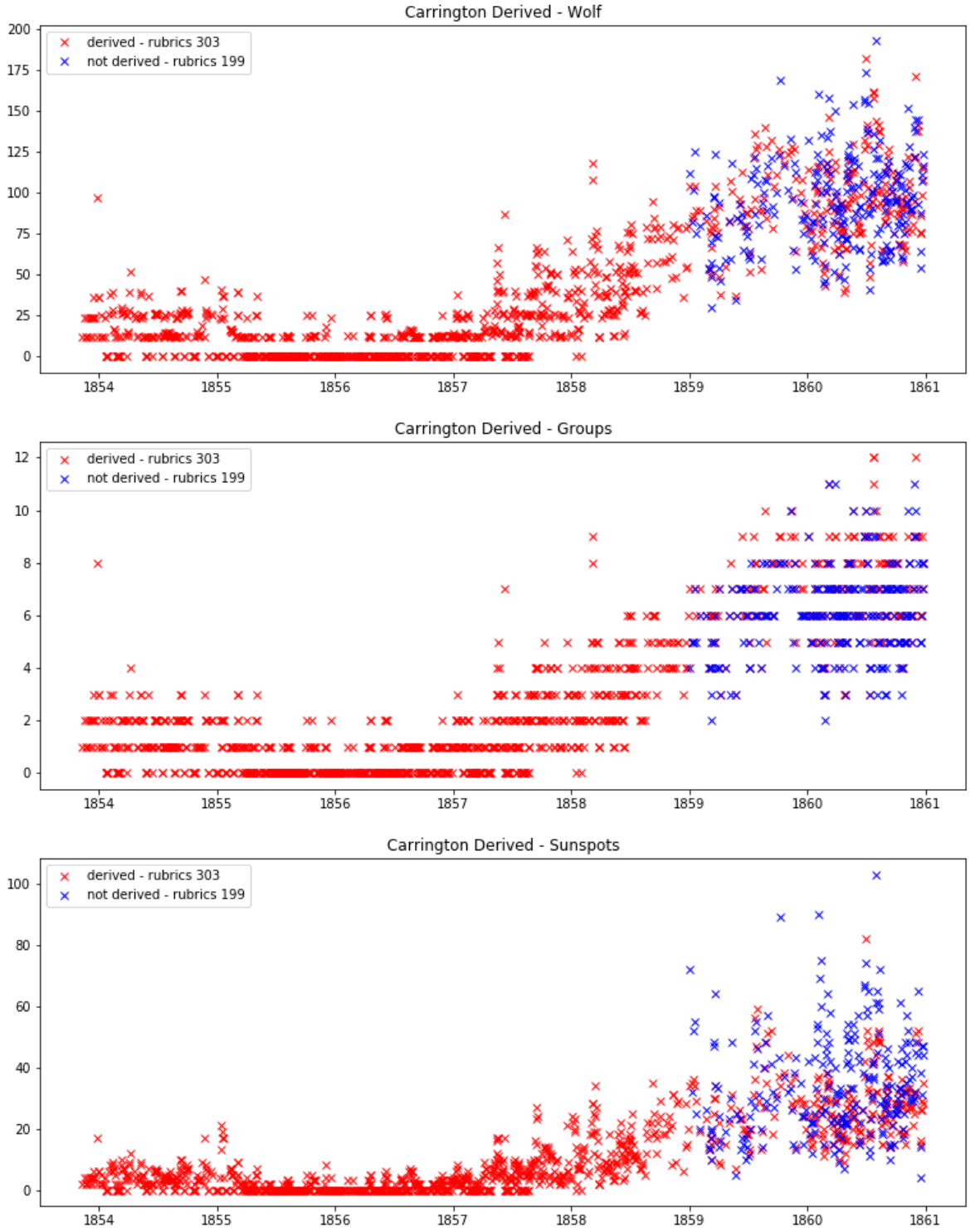

Figure 3: Carrington derived

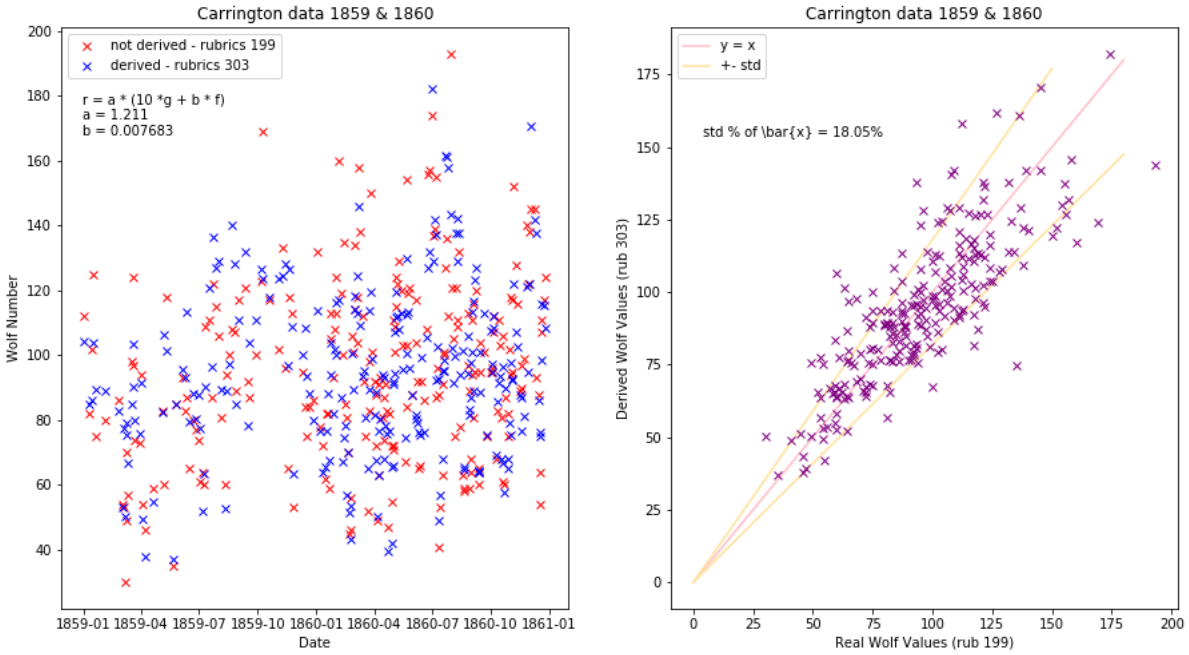

Figure 4: Carrington wolf fit

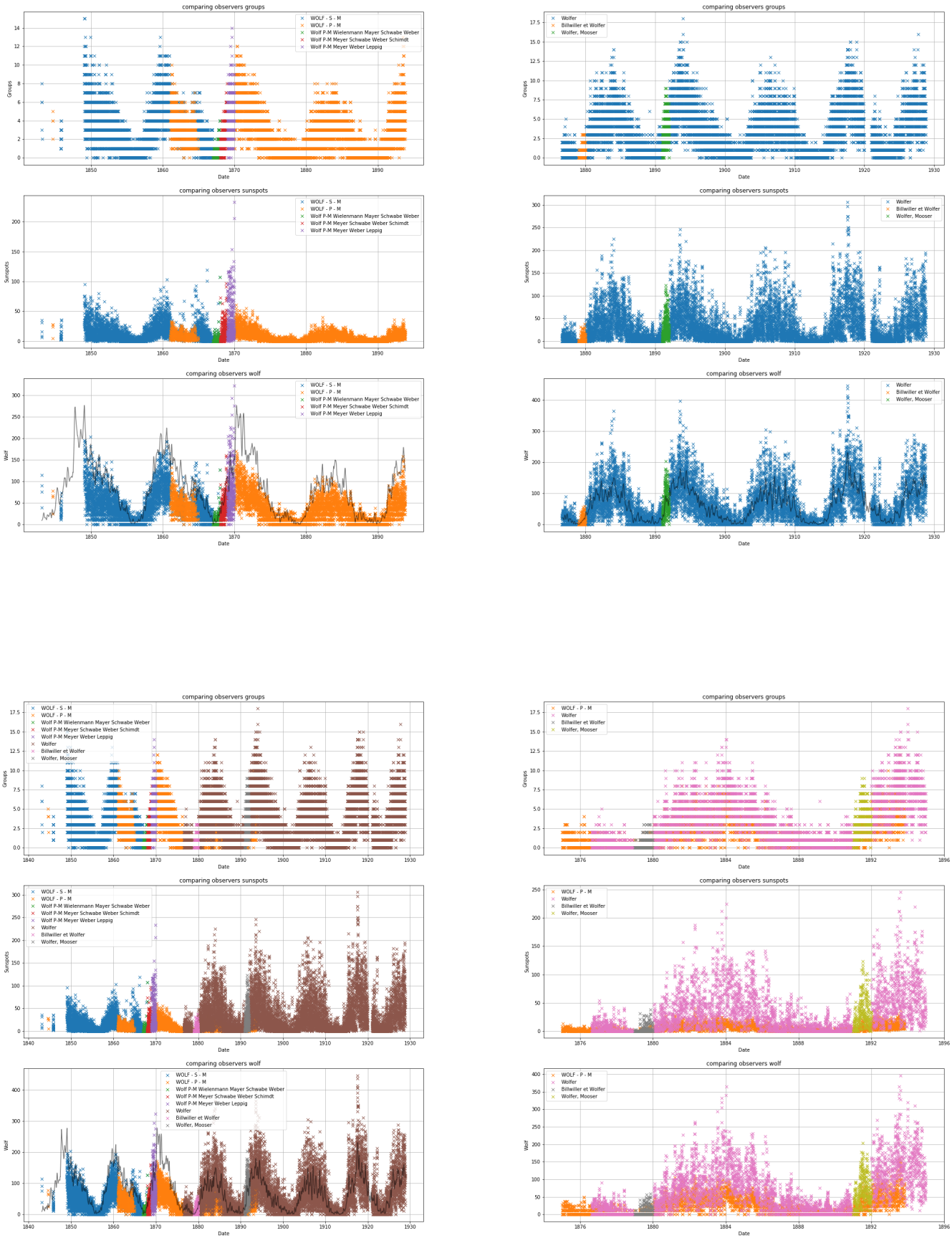

Figure 5: tl = wolf ; tr = wolfer ; bl = wolf and wolfer ; br overlap

5 Converting the f ('aire') - Important Rubrics Translated

5.1 Rubrics 299, Mitt 33, p 128 - 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 geben, die Flache der Sonnenscheibe zu 46352,5 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 an appendix is given a "[It] Register of sunspots observed in the mirror of the Roman College during the year 1871" Collegio Romano durante l'anno 1871", which was held on a series of days of Rome. Remiddi paid groups, instead of the number of spots but contains numbers which give the area in square millimeters inscribed by them, the area of the solar disk is assumed to 46352.5 square millimeters. 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.

5.2 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.

5.3 Rubrics 199, mitt 11-20, p224 - Carrington

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:

5.4 Rubrics 375, mitt 41-50, p244 - Secchi

German Herr Professor Secchi in Rom und sein Adjunkt Ferrari haben 1877 folgende Gruppen-Zahlungen und Flachenbestimmungen erhalten, und theils in ihrem Vulletino theils in gefalligster Weise direct mitgetheilt:

Die zweiten der hier gegebenen Zahlen geben nicht, wie bei den ubringen Serien, die Anzahl der Flecken, sondern sind der Rubrik 'Area mm quadrati' entommen in welcher eine Einheit 21.56 Millionstel der Flache der Sonnenscheibe entsprechen soll

English In 1877 Professor Secchi in Rome and his adjunct Ferrari received the following group payments and area determinations, and some of them in their Vulletino in the most pleasing way directly informed:

The second of the numbers given here do not give, as in the ubringen series, the number of the spots, but are taken from the column 'Area mm quadrati' in which one unit should correspond to 21.56 millionths of the area of the solar disk.

5.5 Rubrics 293, mitt 31-40, p114 - Johann Friedrich Julius Schmidt

German Sonnenfleckenbeobachtungen in Athen im Jahre 1872. Aus einem Schreiben von Jul. Schmidt, datirt: Athen 1873 I 2. Herr Director Jul. Schmidt hat im Jahre 1872 folgende Fleckenzahlungen erhalten:

durch Herrn Prof. Heis in Munster auf meine Bitte vor dem Abdrucke im Mss. mitgetheilt.

*) Die Beobachtungen von Augus bis Ende Jahres wurden mir ducrch Herrn Professor Heis auf meine Bitte vor dem Abdrucke im Mss. mitgetheilt.

Die vollstandigen Beobachtungen sind mit dem sechsfussigen Refractor der Athener-Sternwarte gemacht, - die ubringen mit einem zweifussigen. Wie schade, dass versaumt wurde auch mit dem kleinern Instrumente

die Fleckenzahlungen auszufuhren, und so diese, namentlich fur die Wintermonate an Beobachtungstagen so reiche Serie, ohne viel grossere Muhe noch viel brauchbarer und werthvoller zu machen, als sie dadurch werden, dass man die AthenerGruppenzahlen g nach folgender aus vielen Vergleichungen festgestellten mittlern Scale in Relativzahlen r umsetze. Es entsprechen sich namlish im Mittel

wovon ich fur die wenigen Tage, welche ich in meiner JahresTabelle fur 1872 noch auszufullen hatte, wirklich Gebrauch machte

Conversion table from rubrics 293

g	0	1	2	3	4	5	6	7	8	9	10	11	12	13
r	0	20	38	54	69	83	97	110	123	136	148	160	172	184

English Sunspot observations in Athens in 1872. From a letter by Jul. Schmidt, dated: Athens 1873 I 2. Mr. Director Jul. Schmidt received the following spot numbers in 1872:

by Prof. Heis in Munster at my request before the impression in mass.

*) The observations from August to the end of the year were sent to me by Professor Heis at my request before the impression in Mass.

The complete observations are made with the six-foot refractor of the Athens Observatory, - the ones with a two-foot refractor. What a pity that it was neglected also with the smaller instruments to carry out the spot payments, and so this series so rich, especially for the winter months on observation days, without making much greater effort still much more useful and valuable than they become by converting the Athenian group numbers g according to the following mean scale determined from many comparisons in relative numbers r. The following correspond namlishly on average

which I really made use of for the few days which I still had to fill out in my year table for 1872.