DATA_SILSO_HISTO
Quality Control Report

Stephen Fay

August 13, 2019

Contents

1 Introduction

1.1 Github repository and project

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

1.2 Brief History et Mise en Contexte

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

The aim of this project is to do a quality control of the data in DATA_SILSO_HISTO - To identify and correct things that are wrong with the data.

2 Setup

2.1 What do the flags mean?

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

Flags key table

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

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

2. If in the Mitteilungen journals there is written a '?' next to one of the data points, I will mark it with a 2, this means that the observer is not quite confident in his/her result. See 3.0.10 - July 3 for speculation on what I think comment '?' means. Under this flag I have also groups the comments labeled 'bad definition of sun-picture' - paraphrasing from German.

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

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

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

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

7. Data who's values have been derived from some formulae, usually because observer noted down area measurements of the total number of millimeters the sun-disk was taken up by sunspots.

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

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

2.2 Equations

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

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

Since we have models where $\bar{x}$ is not the mean but a linear model

$$\sigma\% = 100 \cdot \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\bar{x}}\right)^2}{n-1}} \tag{3}$$

2.3 Python scripts - what they contain

See README.md - ¡PUT HREF HERE¿ it is automatically generated based on what are in the scripts. Generated by the file `create_readme.py`. What follows is an explanation of each of the plotting methods in `graphs_helper.py`.

2.3.1 display_seperate_flags(observer,interval=None,yaxis=Sunspots,save$_as = None$)

Parameters

- observers : the alias of the observer who's data you want to plot
- interval [optional] : tuple of length 2 with the start and end date in string format e.g. interval=("1880-01-01","1890-01-01")
- yaxis [optional] : either 'Sunspots', 'Wolf' or 'Groups', what is to be plotted on the y axis
- save_as [optional] : String, if specified the figure will be saved using this parameter as the name

3 Condensed Log

3.0.1 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.

3.0.2 Friday June 21

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

3.0.3 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`

3.0.4 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

3.0.5 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.

3.0.6 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.

3.0.7 Friday June 28

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

3.0.8 Monday July 1

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

3.0.9 Tuesday July 2

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

3.0.10 Wednesday July 3

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

3.0.11 Thursday July 4

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

3.0.12 Friday July 5

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

3.0.13 Monday July 8

- Finished deriving Carrington
- Backed up databases

3.0.14 Tuesday July 9

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

3.0.15 Wednesday July 10

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

3.0.16 Thursday July 11

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

3.0.17 Friday July 12

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

3.0.18 Monday July 15

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

3.0.19 Tuesday July 16

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

3.0.20 Wednesday July 17

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

3.0.21 Thursday July 18

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

3.0.22 Friday July 19

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

3.0.23 Monday July 22

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

3.0.24 Tuesday July 23

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

3.0.25 Wednesday July 24

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

3.0.26 Thursday July 25

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

3.0.27 Friday July 26

- Sorted out rubrics 684

3.0.28 Monday July 29

- Started investigating Wolf and Schwabe's mysterious holes

3.0.29 Tuesday July 30

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

3.0.30 Wednesday July 31

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

3.0.31 Thursday August 1

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

3.0.32 Friday August 2

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

3.0.33 Monday August 5

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

3.0.34 Tuesday August 6

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

3.0.35 Wednesday August 7

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

3.0.36 Thursday August 8

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

3.0.37 Friday August 9

- Made the histogram plotting methods

3.0.38 Monday August 12

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

4 Tables Figures

4.0.1 The original sql data tables format

Table 1: 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 2: 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 3: 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	

4.0.2 My new sql data table format

Table 4: 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	

4.0.3 Figures - plots and graphs

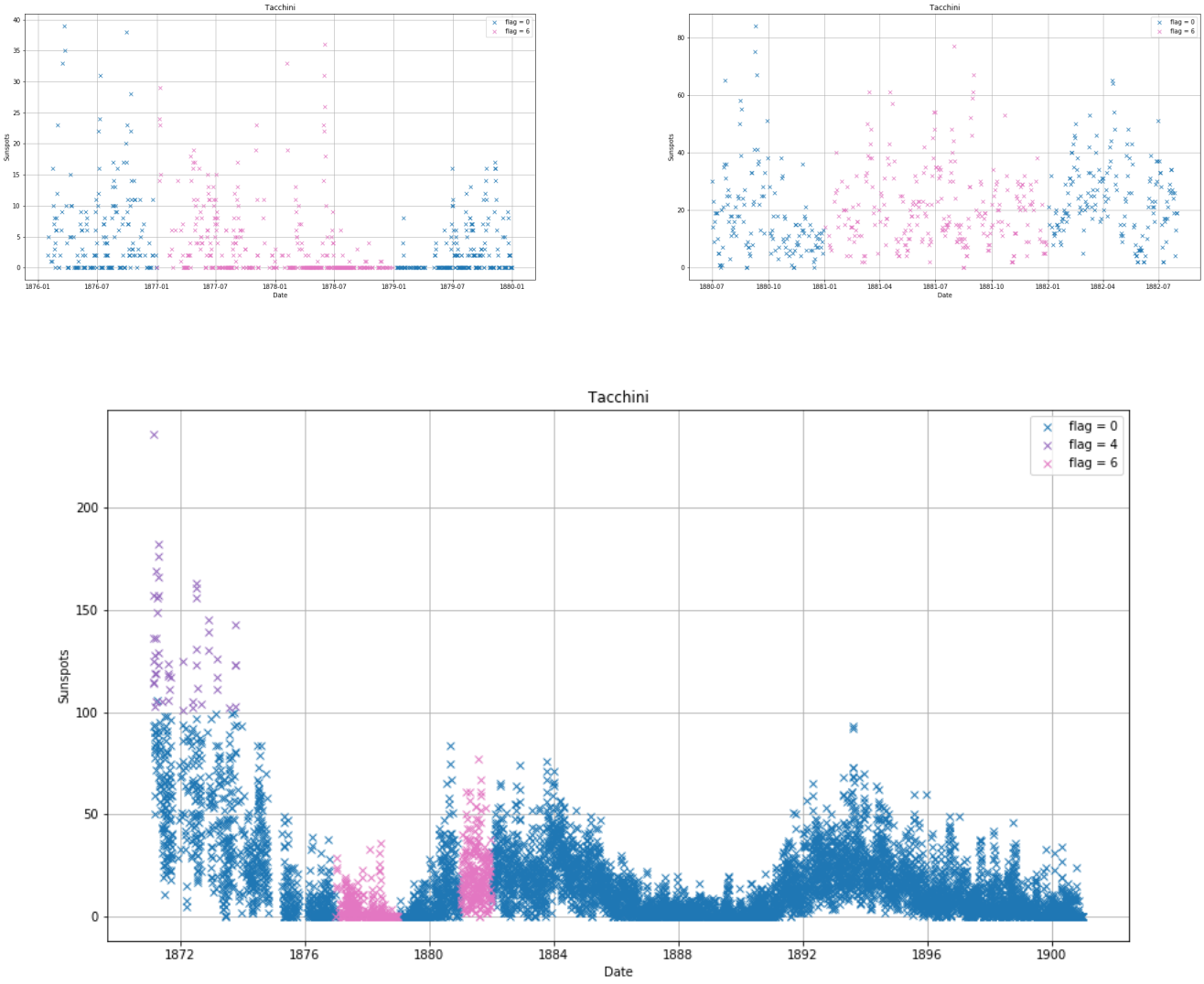
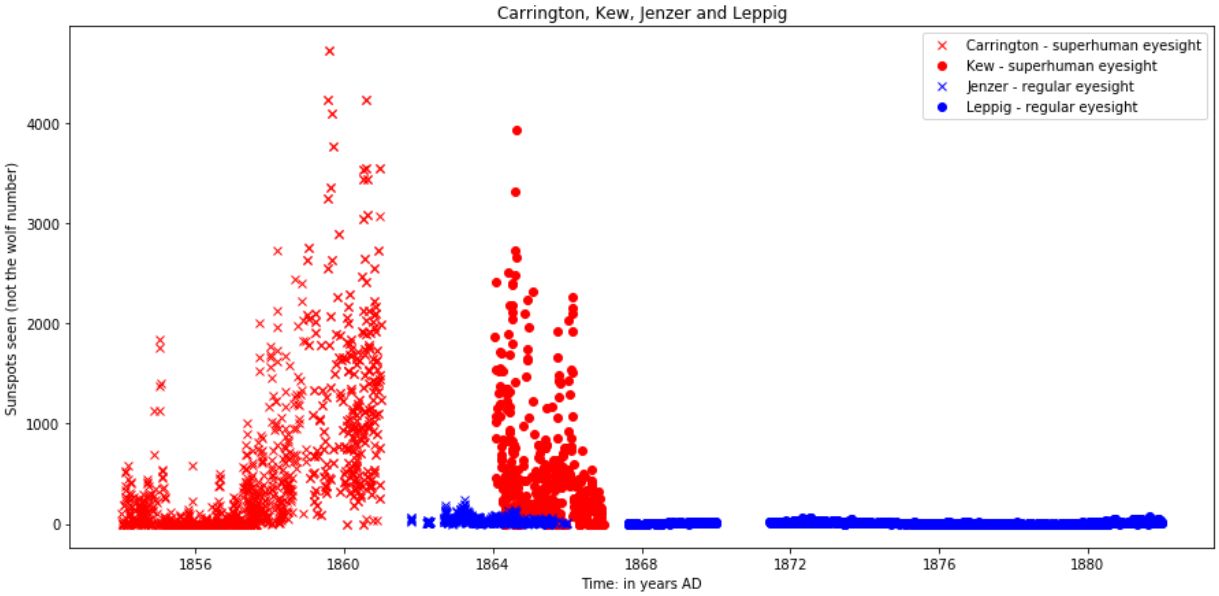


Figure 1: Tacchini



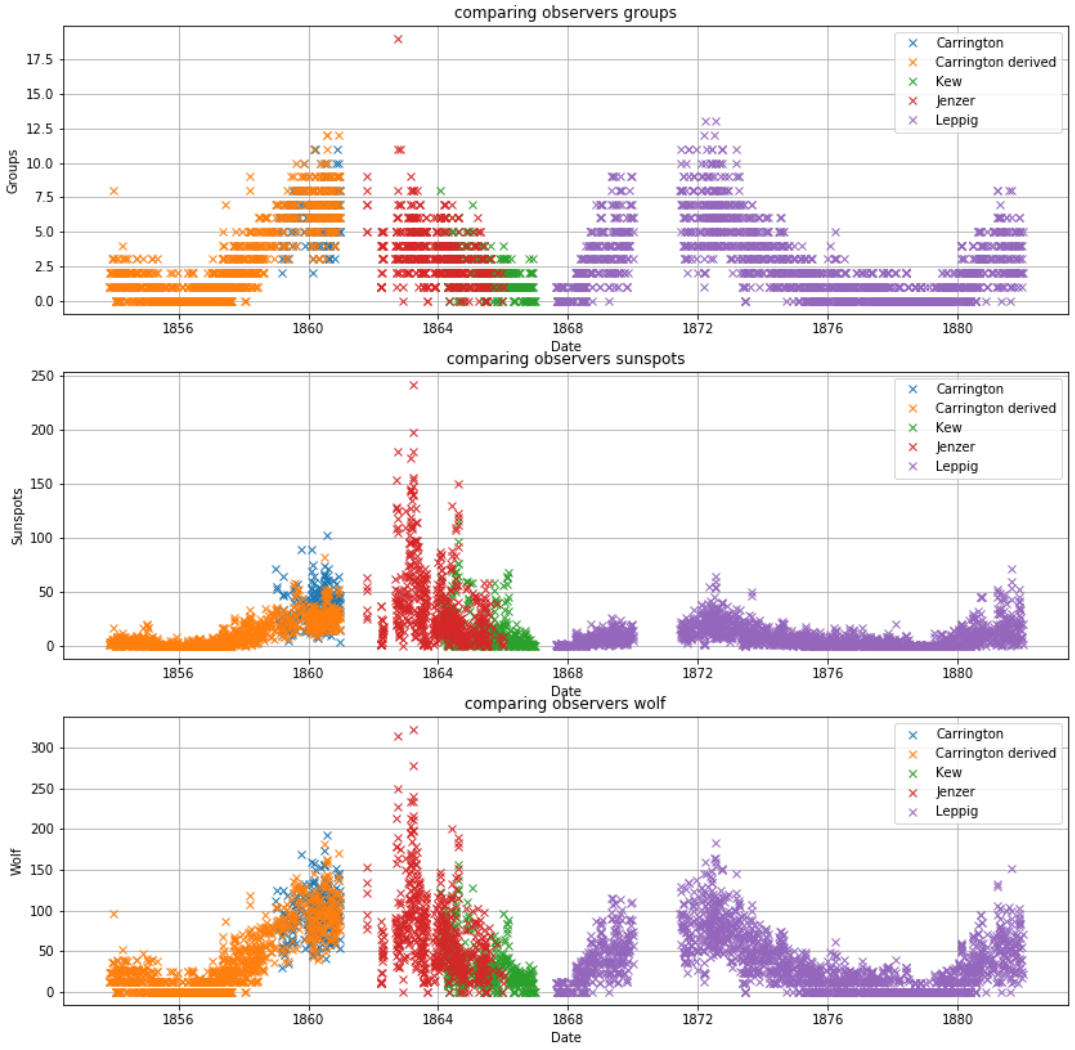


Figure 2: Carrington and Kew - input penumbras instead of sunspots

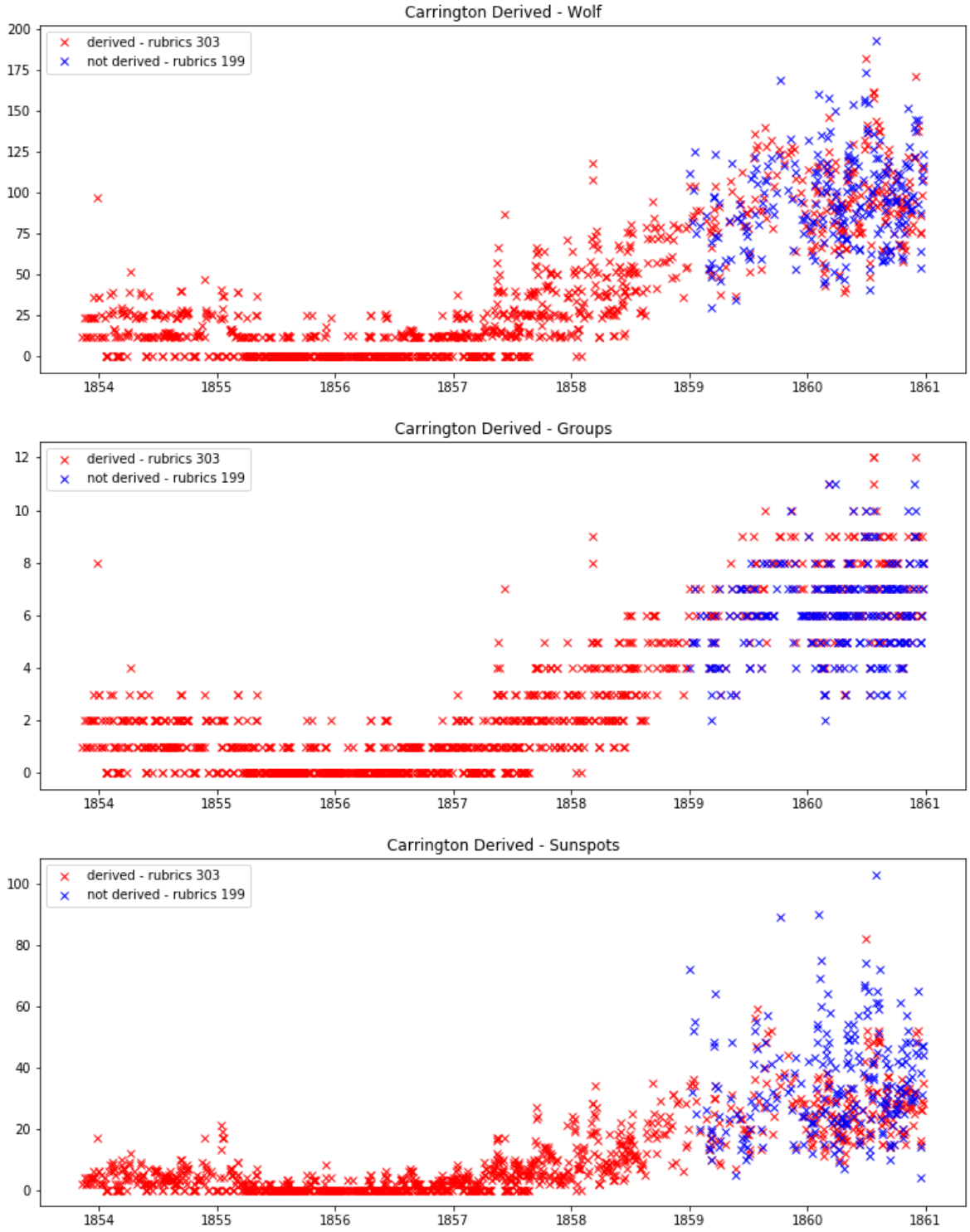


Figure 3: Carrington derived

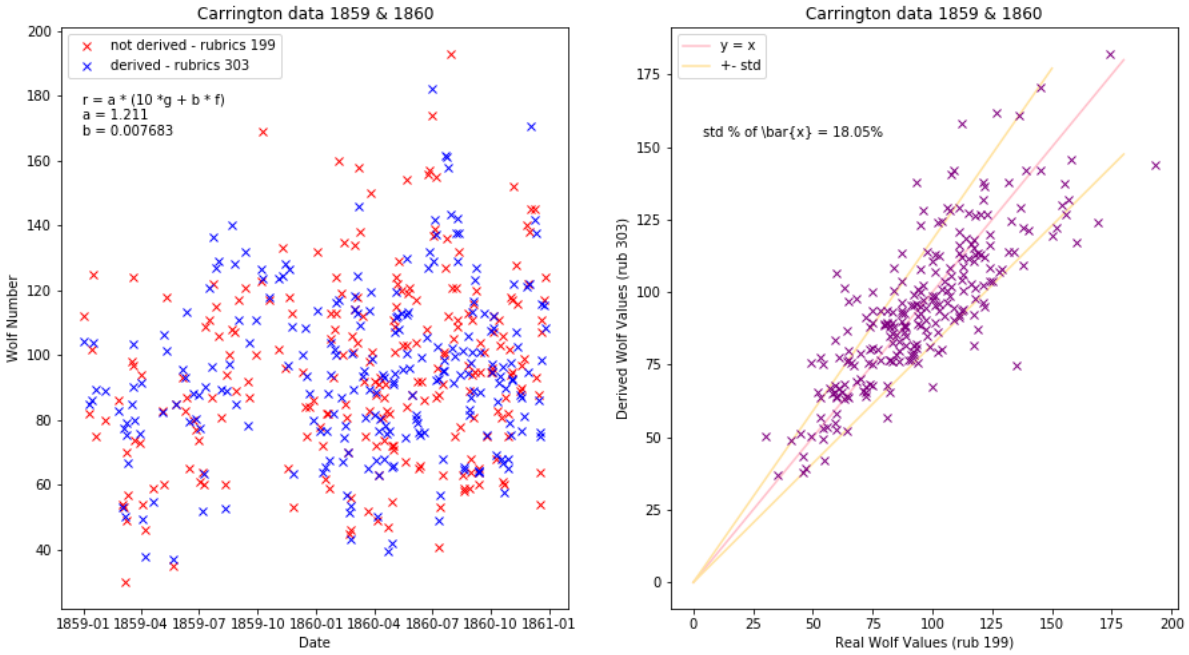


Figure 4: Carrington wolf fit

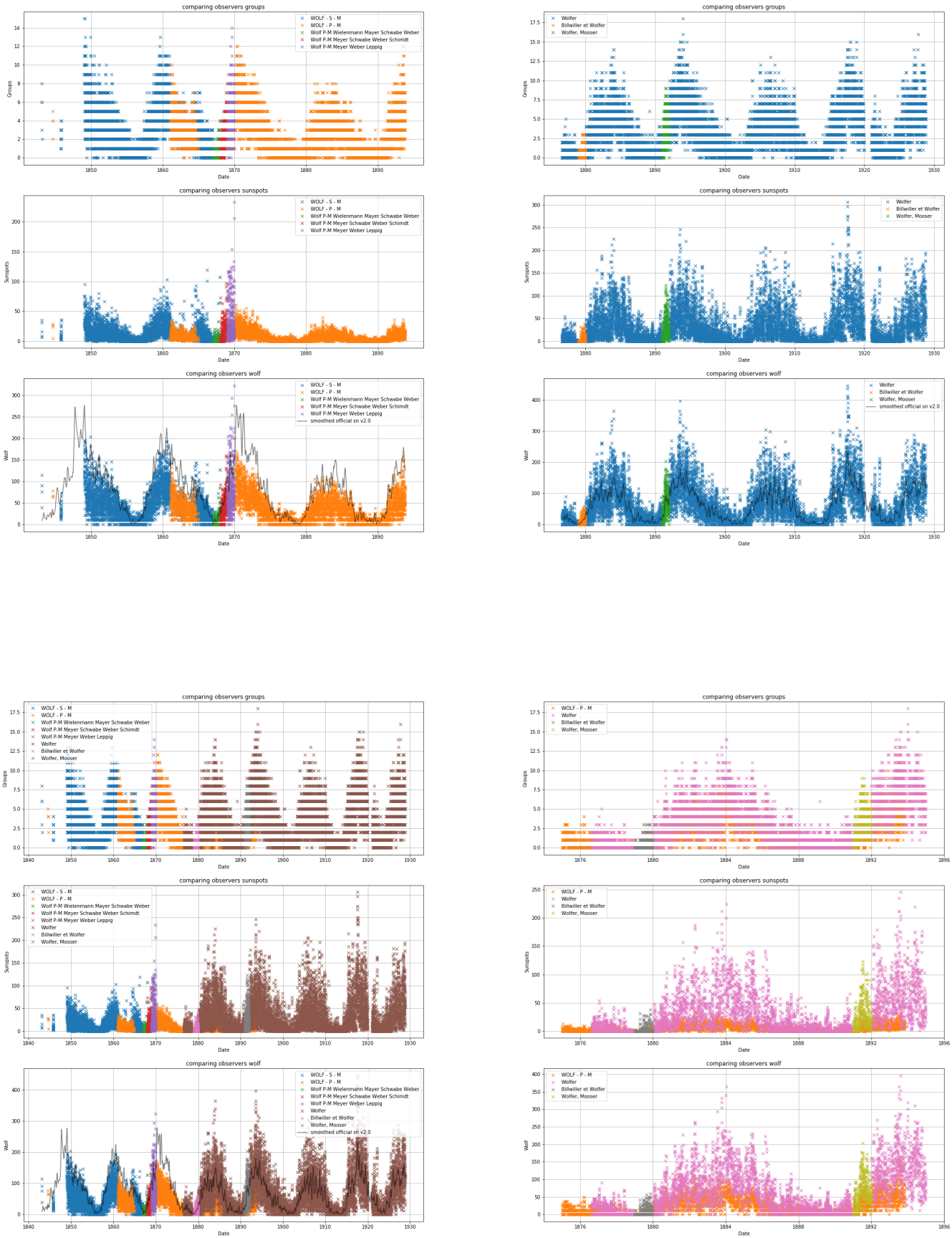
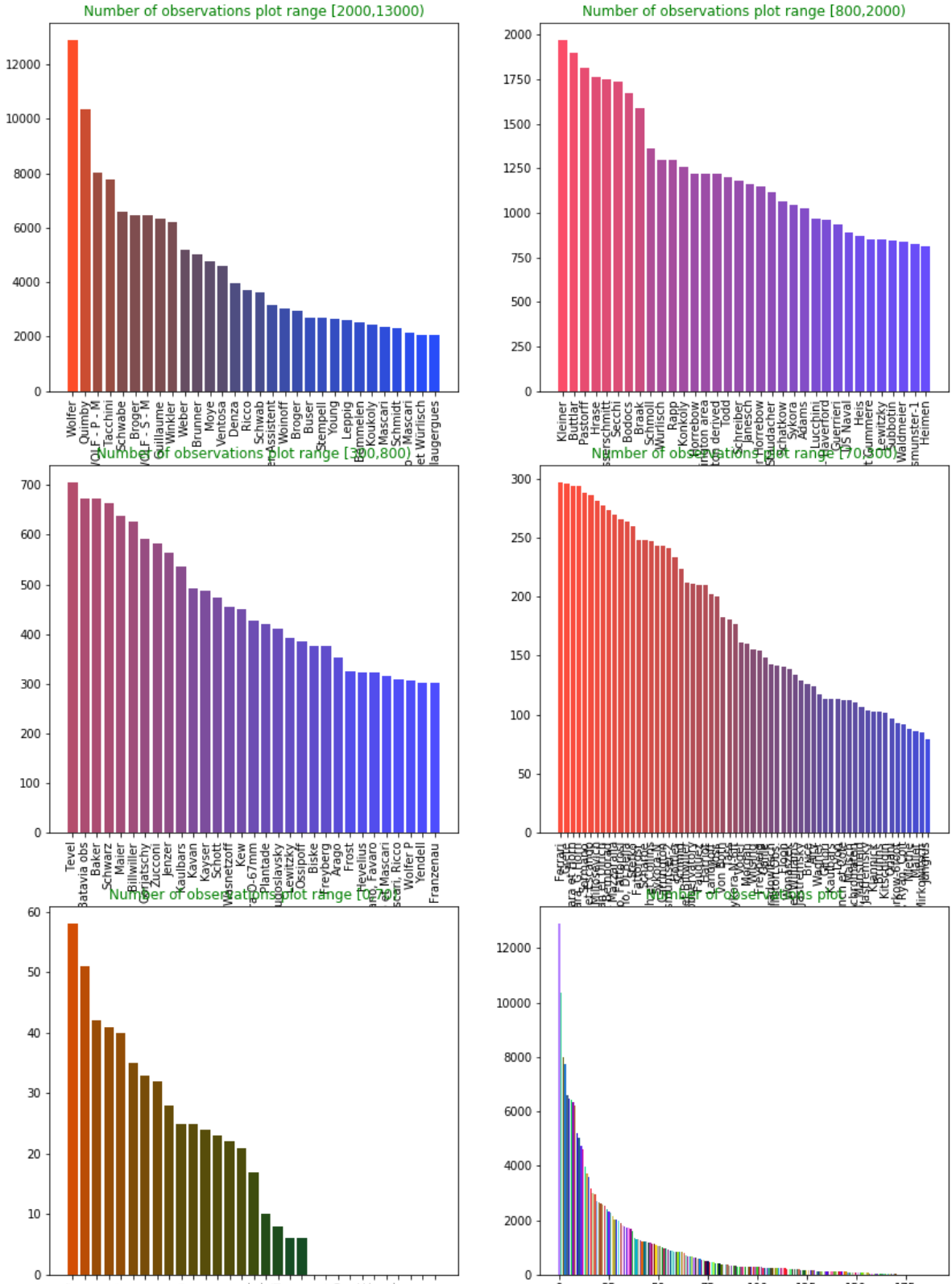


Figure 5: tl = wolf ; tr = wolfer ; bl = wolf and wolfer ; br overlap


Number of observations plot range [2000,13000)
Number of observations plot range [800,2000)
Number of observations plot range [300,800)
Number of observations plot range [70,300)
Number of observations plot range [0,70)
Number of observations plot

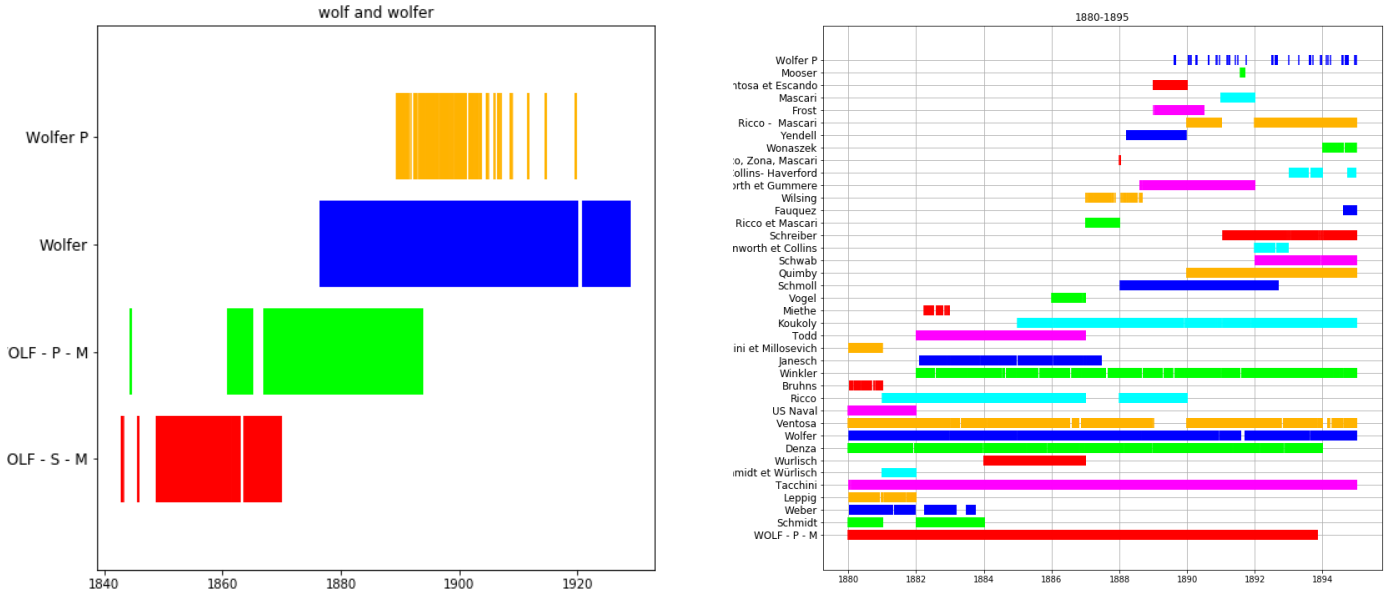


Figure 7: Wolf and Wolfer + 1880 to 1895 event_plots

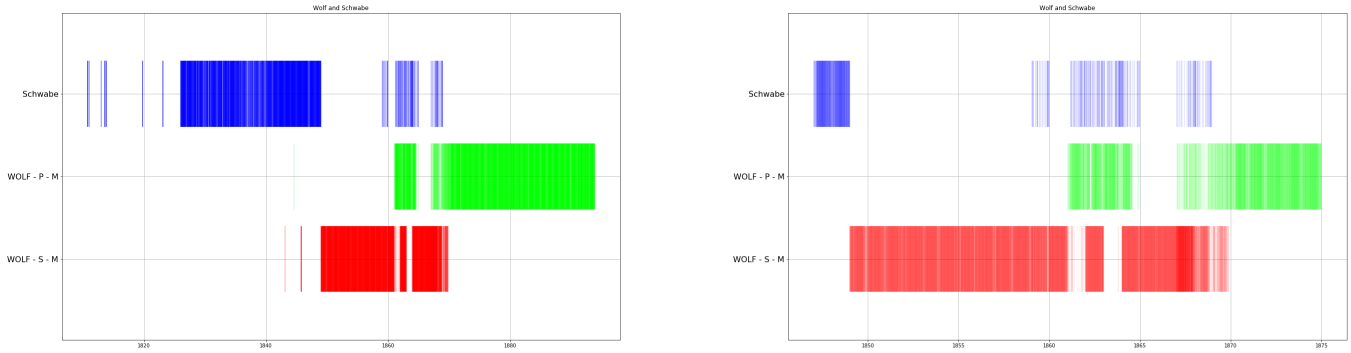


Figure 8: Wolf and Schwabe Eventplots mon 2019-07-29

5 Conclusions

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

5.2 Problems that remain with the database

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

6 Miscellaneous

6.1 Backbone observers

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

https://files.aas.org/astronomy2015/Presentations/DE$_F$r%C3%A9d%C3%A9ric$_C$lette$_H$eliosphere.pdf

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

- make some data visualisations to compare each observer's primary and secondary observing equipment
- for each day / month / year find the highest observation and the lowest observations and add it to the graph so that we have like an upper bound and a lower bound.
- figure out how to smooth graphs with matplotlib and make something nice out off the big mess i currently have
- pie chart of observers with their number of observations
- in the final sunspots number graph cut it into 3 or 4 sections that mark changes in the theory behind sunspots: before wolf ; time where Plato's ideas of the sun being a perfect sphere were still alive and well ; 1908 George Ellery Hale discovers the magnetic link (p14 of nature's 3rd cycle) ; 1955 Eugene parker's theory (p19 of nature's 3rd cycle) ; Nasa send their probe to near the sun

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

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

6.4 Converting the f ('aire')

This section has been moved to the log