DATA_SILSO_HISTO
Quality Control Report

Stephen Fay

July 10, 2019

Contents

1 Introduction

1.1 Github repository and project

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

1.2 Brief History et Mise en Contexte

For centuries we have observed the sun and it's ever mysterious sunspots. The 11 year sunspot cycle has long been a subject of debate. Today we wish to have precise quantification of solar activity throughout the previous centuries. This is made possible by the sunspot series. For the past 3 to 4 hundred years people all over the Eurasian continent have been recording the number of sunspots that appear on the sun's earth facing half.

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

2 Setup

2.1 What do the flags mean?

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

Table 1: Flags key

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

1. If the data looks fishy but I'm not quite sure either what is wrong with it or how wrong it is this is flagged with a 1 - the default.

2. If in the Mitteilungen journals there is written a '?' next to one of the data points, I will mark it with a 2, this means that the observer is not quite confident in his/her result. See 3.0.10 - July 3 for speculation on what I think comment '?' means.

3. A flag that signifies that this data point is definitely going into the bin

4. For data that is very dodgy but it is ambiguous as to weather or not it is correct, to determine its validity closer examination is required

5. For data that is definitely wrong, the difference between 5 and 4 is illustrated by example: if i find that a datapoint has a groups number of 30 I will mark it with a 4 and comment it, because this is suspicious, if a datapoint has a groups number over 60 or above, it will be marked with a 5 (trust me there are some in the hundreds).

2.2 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 - it is automatically generated based on what are in the scripts

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
- 6.1 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)

4 Tables Figures

4.0.1 The original sql data tables format

Table 2: DESCRIBE DATA

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

Table 3: DESCRIBE OBSERVERS

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

Table 4: DESCRIBE RUBRICS

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

4.0.2 My new sql data table format

Table 5: DESCRIBE DATA (the only table)

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

4.0.3 Figures - plots and graphs

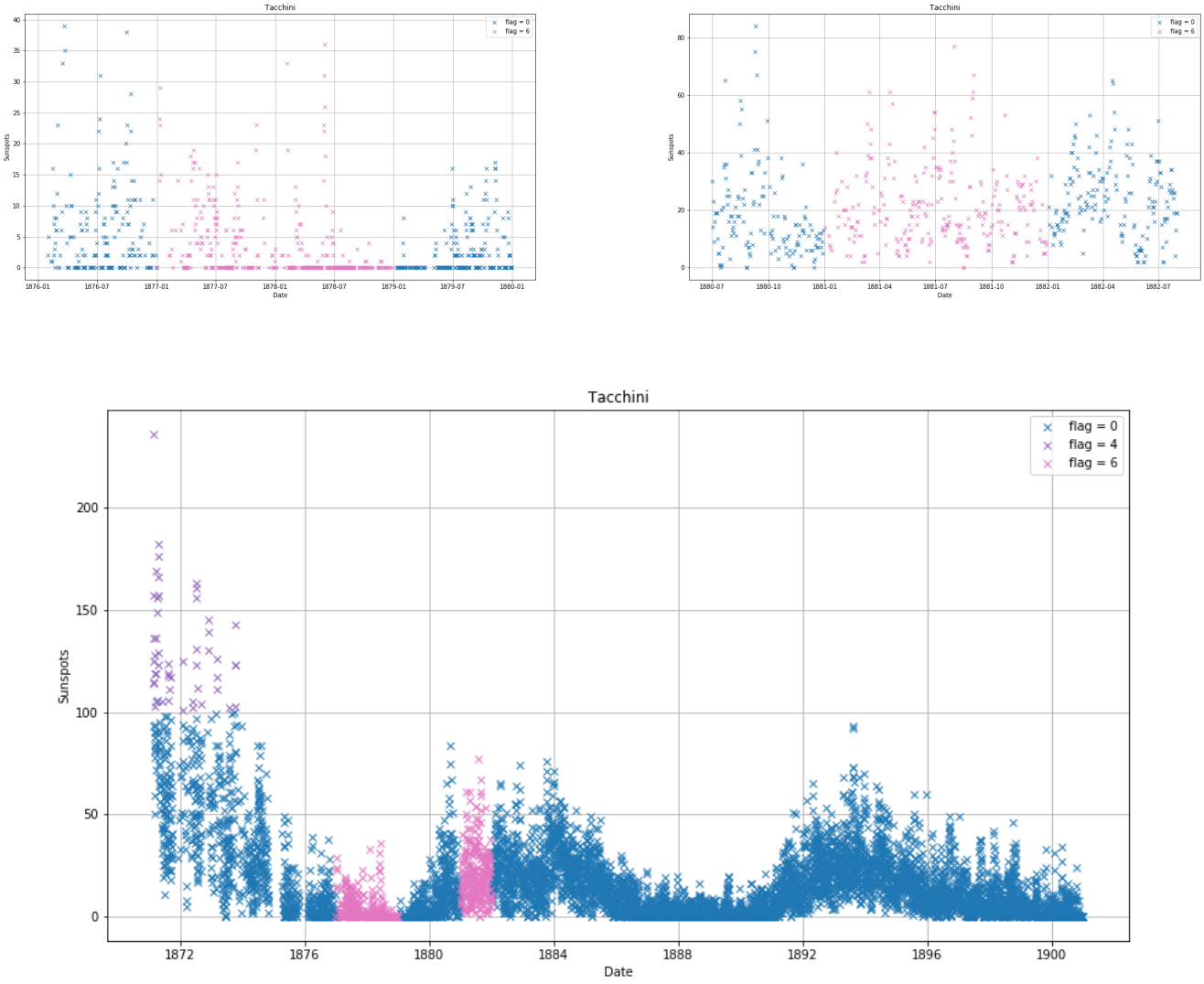

Figure 1: Tacchini

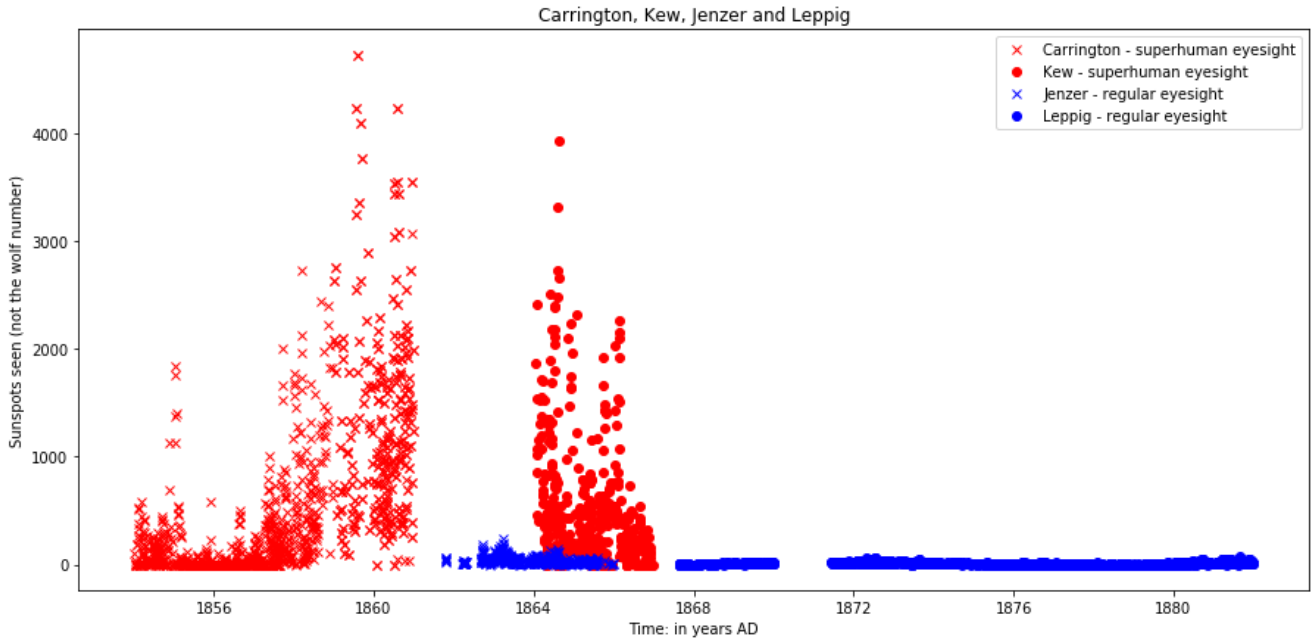

Figure 2: Carrington and Kew - input penumbras instead of sunspots

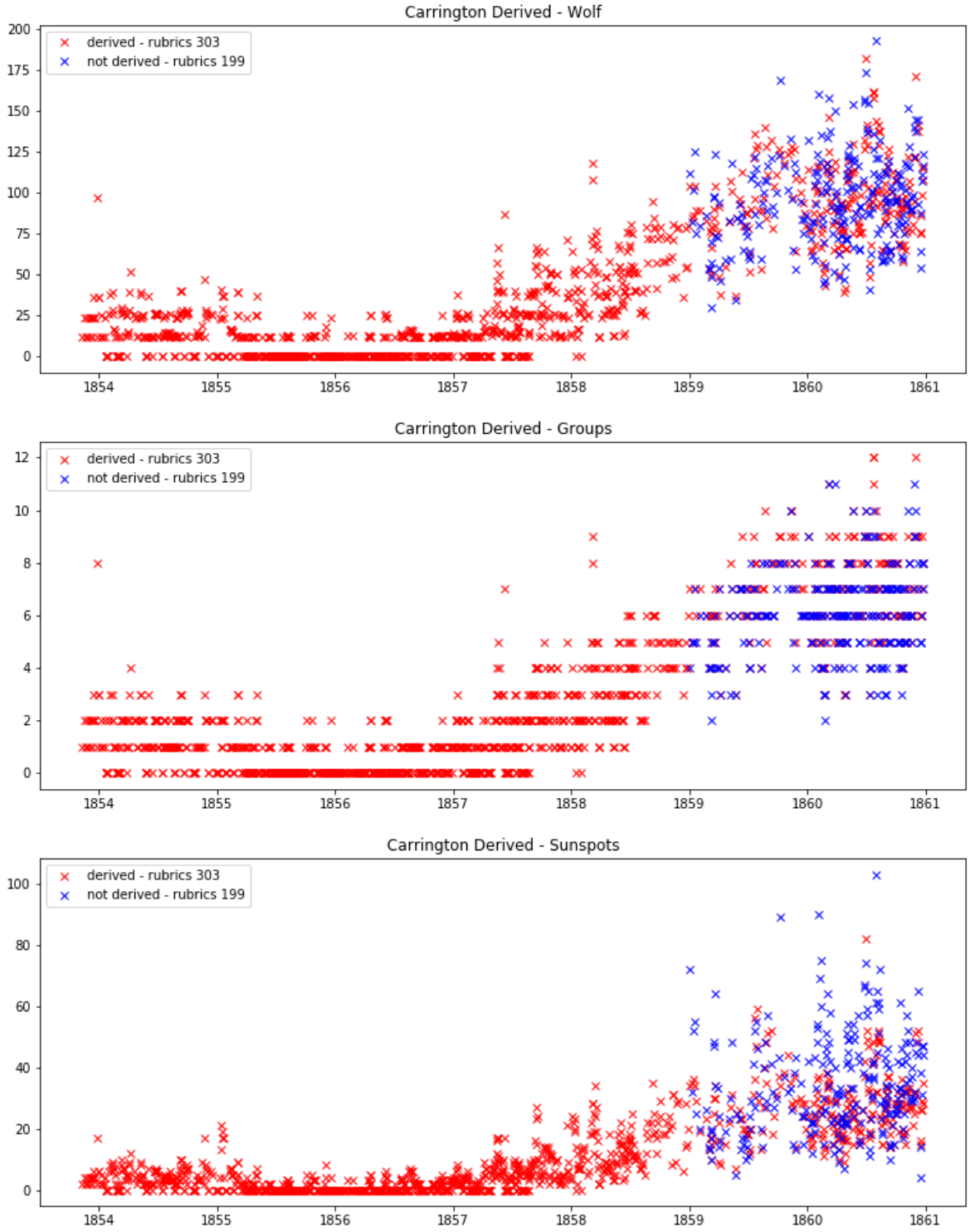

Figure 3: Carrington derived

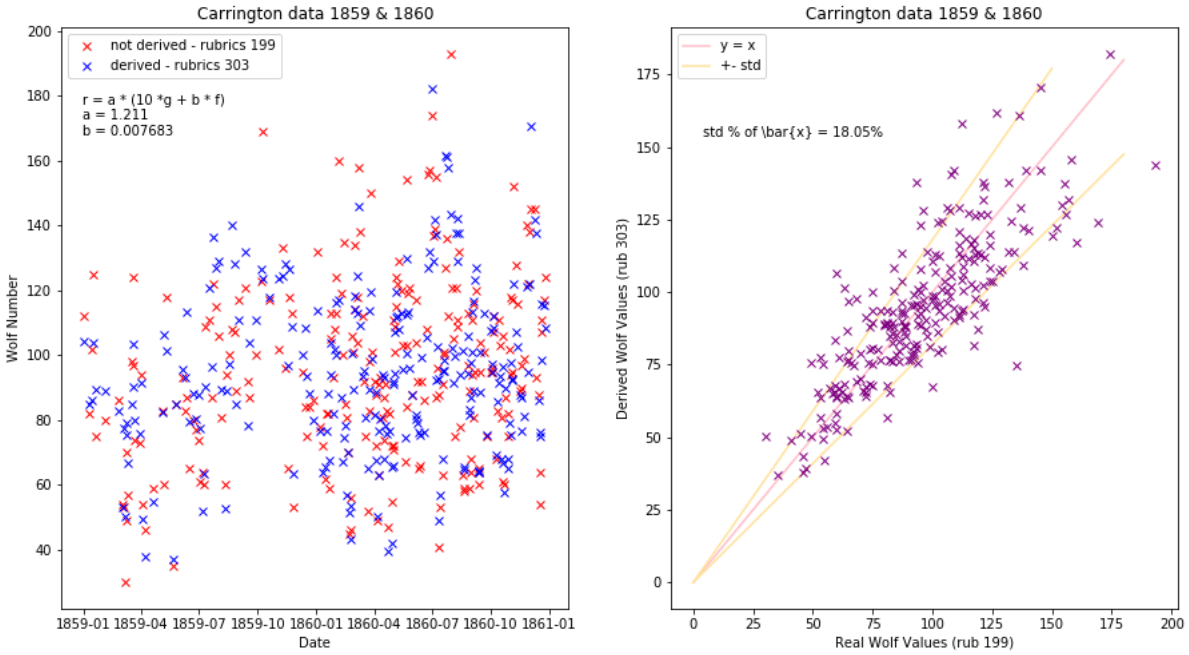

Figure 4: Carrington wolf fit

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

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

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

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

6.1 Converting the f ('aire')

This section has been moved to the log