DATA SILSO HISTO quality control Report

Stephen Fay

July 2, 2019

1 Introduction

1.1 Github repository and project

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

1.2 Brief History et Mise en Contexte

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

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

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

2.1 Everything wrong with the data

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

2.2 Annotation keys

2.2.1 What do the flags mean?

0	1	2	3	4
same as Null	suspicious	Comment in journal = ?	move to bin	suspiciously high
5	6	7	8	9
very suspicious	misc see comment	probs ok, to be investigated	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.

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

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

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

2.3 Search and correct.

2.3.1 Outline

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

2.3.2 Log

I started this (today) on 2019.06.21 (yes, the solstice!)

- 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)

- 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 ; 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
- 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
- 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 2.2.1). 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.

- 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.

- 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.

- 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_rubr 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

- Tuesday July 2
  - exploring ways of creating visualisations that will help me catch some of the suspicious datapoints. Right not my task is hunting down those ones which I labelled 'suspicious sunspots'
  - made some pretty plots which you can find in *suspicious sunspots plots.ipynb* in the root directory
  - Using the plot I was able to check some of Weber's suspicious sunspots in the Mittheilungen. And strangely enough Weber's observations are correct! It seems he really did see 476 sunspots on the sun on the 25th of september 1870. I mean the patter fits it's just unsusually large. And you need to be dedicated to count hundreds of spots every day.
  - Made a method in the jupyter notebook mentioned above that plots an observer's stuff and color codes the flags. I investigated Tacchini's green (flag=4) datapoints and they are infact correct, they appear in the journals. There is one datapoint from Tacchini which I corrected by hand, this one had a wolf number of 61, I found it in the journals, the correct value was 6 - typo. I corrected it manually.
  - In the rubric 279 mitt 30 page 409 Tacchini observes a bunch of sunspots without observing any groups. This is annoying because it means we have no wolf! Other than that they seem to have been entered in correctly, I have not yet transfered these to the GOOD DATA SILSO database, they are still only in DATA SILSO HISTO and BAD DATA SILSO
  - I have an idea that we could do some stats and deduce a wolf number even without the groups, and give it a special flag, based on some probability we estimate a wolf number. There might be some complications with this tactic, here are 3 possible ways of doing it and their weaknesses:
    1. For each sunspots number possible {1,2,3...} find the corresponding most likely wolf number by going though each data point with that number and doing a distribution (probs normal) to find which is the most likely wolf number associated with this sunspots number. Essentially we define a function term-by term $f \stackrel{\text{def}}{=} \mathbb{N} \to \mathbb{R}$. The weakness : if we sample everything we might find that Tacchini has his own idiosyncratic way of doing sunspots and wolf, and so the data we add to his entriee would not fit well with his methods of observation.
    2. For each sunspots number do the same as above but go through the groups number! Again for each sunspots number find sift through all the data of everyone to find the best groups number $g \in \mathbb{R}$, i.e. the g s.t. it sits on top of the fitted normal distribution.
    3. For each sunspots number, methods 1) and 2) above but fitted with only Tacchini's previous and future data. The weakness of this method is that there is not much data to go on... (relative to the first two)

4. One thing that worries me in the case of Tacchini is that this guy observers a lot of sunspots. I am looking at his entries right now and in 1870 he regularly observers over 200 sunspots, on 5^{th} of April 1870 he sees 302 sunspots. That said cross-referencing his data with other observers' from the same time seems to support this hypothesis. But the more sunspots he observers, the less data we have to make a nice normal distribution for each sunspots number, there may be sunspots numbers that only appear like 3 times in the whole database, how are we to do any stats on those. The answer is the following method for tying sunspots to groups and wolf number: for each sunspots number we link it to a wolf number by finding the best fit normal distribution for the likeliest wolf number to be associated with it. Then we give this value an error bar which is bigger the less data we have. Then we do a plot x-axis = sunspots number, y-axis = wolf numbers with vertical error bars. Then we do a line of best fit through the whole lot, try several models and do a chi-squared test. It might be worth getting rid of the small ones and only look at data where $s > 20$. Again we could do this for wolf directly and for groups then wolf.

- Before embarking on this adventure I will first endeavour to plug both of Tacchini's hole's that appear in the plots I made: 3 entire years are missing From Tacchini's data 1877, 1878 and 1881
- In 1881 the rubrics number 465 contains two data sets, one of them is entirely Ricco's and the second is from Tacchini and G. Millosevich. Nowhere int the data set is it indicated who saw what, so I looked and found that there was no Alias for Millosevich. This makes me think that he must be Tacchini's assistant or observing partner. Anyway we have a big gap in Tacchini's observations, what I will do is comment all of these observations "Tacchini and Millosevich". I did that and gave them a flag, then regenerated the Tacchini graphs and there is no more gap in 1881, what's more the data looks almost identical. There is one only slightly worrying difference and maybe I'm inventing patterns where I am trying to see them but the 1881 observations are on average a tiny tiny bit higher than the surroundings. Actually I will put a picture **here (put link to picture I will include in this report)**, I think this is infact a seasonal effect, there is more atmosphere...

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

2.3.3 Python scripts - what they contain

3 Comparaison du data avant et apres + visualisations

3.0.1 The original sql data tables format

Table 2: DESCRIBE DATA

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

Table 3: DESCRIBE OBSERVERS

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

Table 4: DESCRIBE RUBRICS

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

3.0.2 My new sql data table format

Table 5: DESCRIBE DATA (the only table)

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

3.0.3 Graphs and visual representations

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

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

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

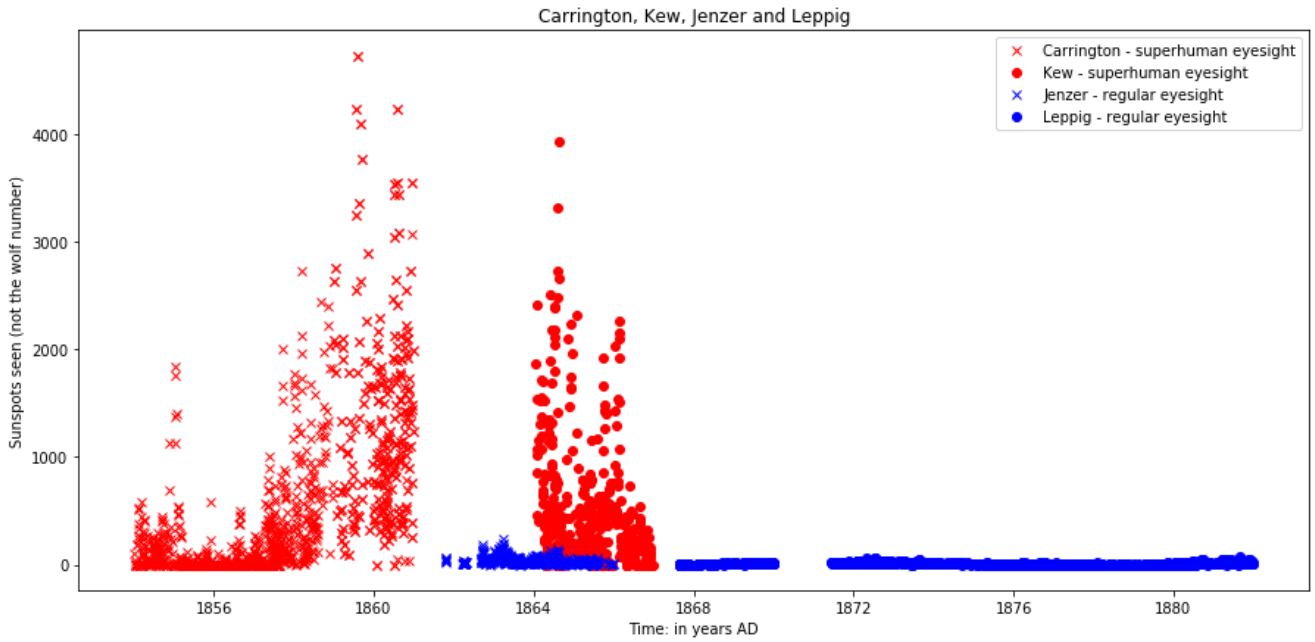

Figure 1: Carrington has great eyesight!