



CENTRE FOR RENEWABLE &  
SUSTAINABLE ENERGY STUDIES

# Quality Checking of Weather Data

Information Document

Don Fitzgerald

15 October 2021

# CONTENTS

1: Quality checking and formatting tool for SAURAN data .....	3
1.1: Opening the tool online.....	3
1.2: Using the tool.....	4
1.3: Interpreting the output file .....	5
1.3.1: Missing time stamps .....	5
1.3.2: DNI error .....	5
2: Manual quality checking .....	7
2.1: Data types.....	7
2.2: Performing quality checking.....	7
2.2.1: Missing time stamps .....	8
2.2.2: Outlying data .....	9
2.3: Fixing date issues .....	10
2.4: Illustrating data allocation.....	12
Contact Details.....	13



## INTRODUCTION

Quality checking of data used in research is imperative - data may be incorrect, missing or skewed due to faults occurring in the data capturing or storage process. It is the responsibility of the researcher to perform quality checking on their data before implementing it.

In data science, it is good to keep in mind that data is a test of some physical phenomenon, but the data can never perfectly represent reality. It is for this reason that the data must be tested for quality, to gain some insight into how close to reality the data is. In the case of solar data, the true irradiance hits the instrument, where it may be partially blocked from entering the device due to dirt build up on the outside of the instrument; the solar tracking device may also be slightly out of alignment, not pointing directly at the sun or not be perfectly level; the device itself may be slightly miscalibrated and not be measuring irradiance accurately. All these factors contribute towards a measured datapoint that is different from reality. It is important for the researcher to have a rough estimation of how different the data may be (quality) so that they can apply a similar variability or probability to the conclusions drawn from this data.

In this document it will be shown how to perform quality checking on SAURAN data using an online tool developed for such a purpose; and manually using Microsoft Excel (MS Excel).

-----X-----



## 1: Quality checking and formatting tool for SAURAN data

This section will detail how to make use of a tool that automatically performs quality checking and formatting of data downloaded from the SAURAN website (<https://sauran.ac.za/>). This tool should have been downloaded alongside this document as a \*.ipynb file. If you do not have this file, then you can contact Don Fitzgerald at [don@sun.ac.za](mailto:don@sun.ac.za). Alternatively, you may consult the SAURAN website, where you can find this file by navigating to any of the weather stations and clicking on this link:

### Download data

From date

Year Month Day

2021 9 17

Date of earliest data: 2010/5/25

To date

Year Month Day

2021 9 17

Date of latest data: 2021/9/17

☒ Day-averaged ☐ Hour-averaged ☐ Minute-averaged

Download

[How to check data quality](#)

### 1.1: Opening the tool online

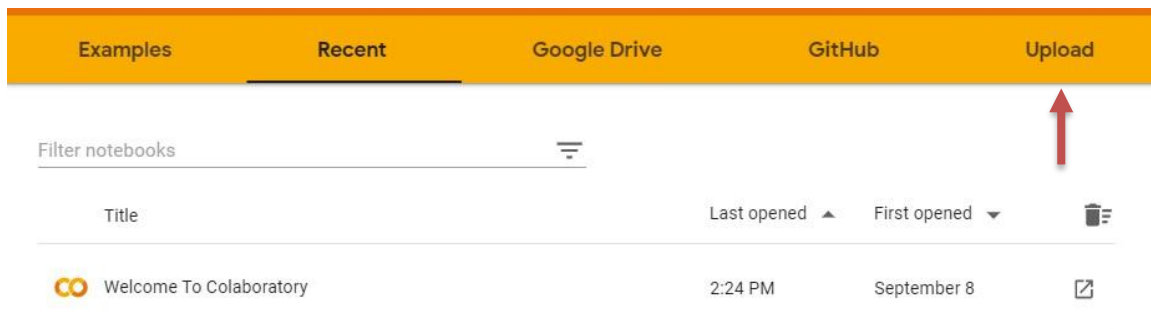
The tool is written in Python and may be executed in your web browser using Google Colaboratory. It is recommended that Google Chrome is used as the web browser when opening this tool (as opposed to Edge, Internet Explorer, Firefox or Safari). Using Google Chrome, navigate to this link: <https://research.google.com/colaboratory/>. Alternatively, you can just do a Google search for “Google Colab” and follow the appropriate link.

Note that you require a Google account to access Google Colaboratory. You may need to sign in at this point, but you do not need to register to use this tool. Every Google account holder has access to Google Colaboratory.

*Side note: if you are proficient in Python and would like to run or view this code locally, you can open it using Jupyter Notebooks. Alternatively, you can use Google Colaboratory to convert the \*.ipynb file into a standard \*.py Python file and then open it in any IDE. Python together with the Pandas library are excellent tools for data analysis. Here is a link to a video that can help you get started if you are interested: <https://youtu.be/sZDgJKI8DAM>*




You should be met with the following welcome screen, and you can hit the “Upload” button:



Alternatively, you can click “File” in the top left corner and then “Open notebook”. Once “Upload” is selected, click “Choose File” and navigate to the \*.ipynb file downloaded with this document. The tool should open, and you may follow the instructions as seen in the tool itself.

## 1.2: Using the tool

You may run the tool by clicking the play button in each cell from top to bottom:

- ▶ -- click the play button, wait for it to finish and then move onto the next cell.  
 [Show code](#)
- ▶ Hit play and upload a weather file from the SAURAN website (\*.csv).  
[Show code](#)
- ▶ Hit play to process the data and download the output file.  
[Show code](#)

**NOTE:** If you would like to run this program again, click the play button of the second cell and upload another file.

Wait for each cell to finish loading before moving to the next one. The second cell will require you to upload raw data from the SAURAN website. Please note that the program is only designed to handle hourly- and minute-averaged data and **does not process daily averaged data**. It is recommended that the quality checking be done on hourly-averaged data for the best results.

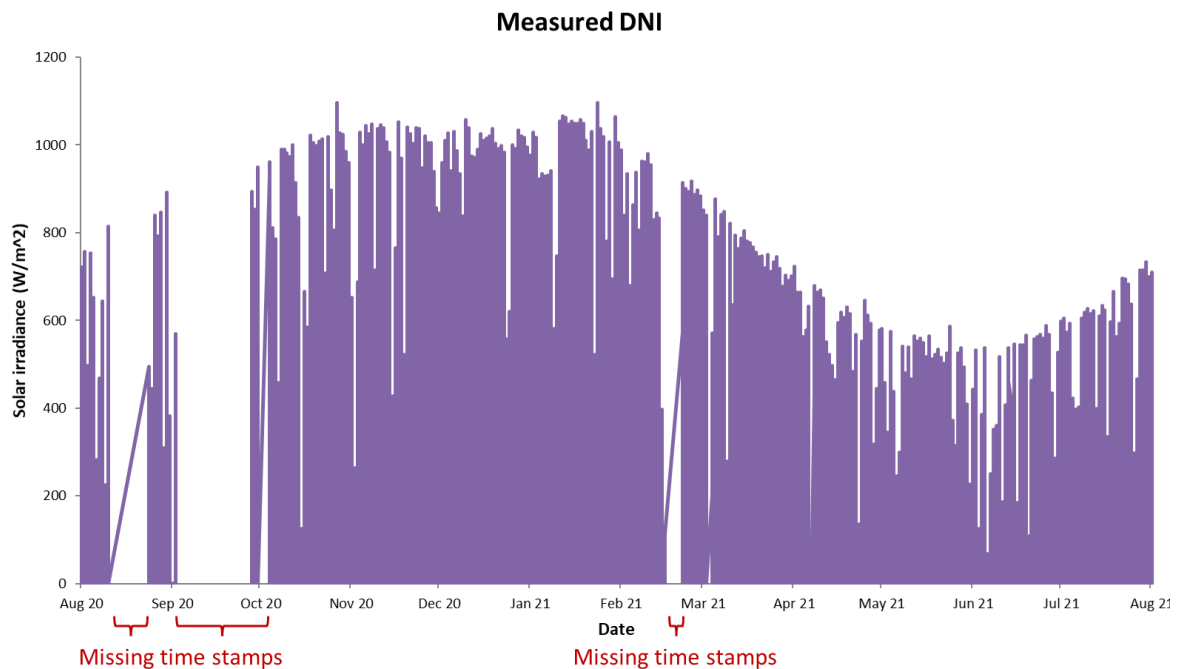
You will need to upload the raw data as obtained from the SAURAN website in a \*.csv file format. Before uploading the document, please make sure there are no missing time stamps in the first 3 rows of data. Once the file has been processed it will download the output file to your computer. If you run this tool more than once you may see an “allow multiple files to be downloaded” pop-up message appear in the top left corner of the web browser for your confirmation.

### 1.3: Interpreting the output file

The tool outputs a Microsoft Excel workbook (\*.xlsx). This workbook contains two worksheets, namely; 'Data\_Quality\_Check'; and 'Original\_Data'. The first sheet shows the results of a basic data quality assessment, while the second sheet provides the original raw data.

#### 1.3.1: Missing time stamps

The first step in data quality assessment is to check for missing time stamps. In cell G2, the total number of time stamps in the dataset is given, while cell H2 shows the number of missing time stamps. The missing timestamps may then be inserted using secondary data as is explained in the next section of this document. If you look at the 'Measured DNI' or 'Measured GHI and DHI' chart, you can determine the exact location of the missing time stamps:



#### 1.3.2: DNI error

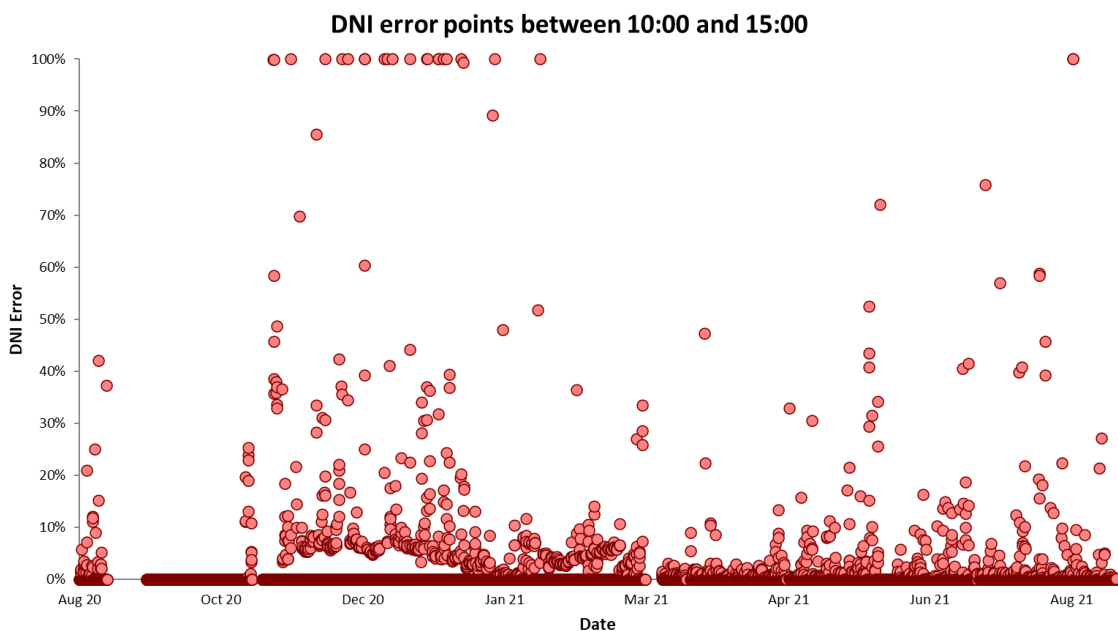
In datasets downloaded from the SAURAN website, the majority include a column called 'DNI\_calc'. This stands for the calculated DNI, which is calculated using the measured GHI, measured DHI and the zenith angle( $\theta$ ) as follows:

$$DNI_{calc} = \frac{GHI - DHI}{\cos(\theta)}$$

The calculated DNI gives us an indication of what the DNI should be, and if the instruments are reading accurately enough for that timestamp. The  $DNI_{calc}$  is compared to the measured DNI by calculating the  $DNI_{error}$  like so:

$$DNI_{error} = \left| \frac{measured\ DNI - DNI_{calc}}{DNI_{calc}} \right|$$

The  $DNI_{error}$  is not accurate in low-light conditions, therefore low-light conditions are filtered out of the results given in the output file. Only data recorded between 10am and 3pm with irradiance levels above  $5\text{ W/m}^2$  are considered when calculating the  $DNI_{error}$ . The rule of thumb is that if the  $DNI_{error}$  rises above 4%, this should be flagged as poor-quality data. This could be due to misalignment of the solar tracker, or due to dirt build-up on the instruments themselves. The number of datapoints which have a  $DNI_{error}$  are given in cell I2 of the output file and the ratio of  $DNI_{error}$  datapoints to the total number of datapoints is given in cell J2. These are also illustrated in the chart:



This now gives the researcher an indication of data quality. It is at the prerogative of the researcher to decide whether this data quality is acceptable for its application or if they would like to replace the poor-quality data with secondary data, as described in the next section of this document.

## 2: Manual quality checking

In this section, it will be demonstrated how to perform quality checking of SAURAN data manually in MS Excel. In this demonstration we will only consider Irradiance data, but the same procedure is done when checking other datatypes such as wind, temperature, relative humidity, etc.

### 2.1: Data types

Weather data is typically recorded through ground stations, satellite data or a mixture of both. Ground stations provide the most accurate data, assuming the instruments are well maintained and calibrated. Satellite data averages weather data over a larger area (e.g. 50km x 50km), and therefore has a large resolution. Some online applications provide a mixture of ground and satellite data, averaged using algorithms over large areas.

Common practice in research is to make use of primary and secondary data. The researcher uses their more accurate data as their primary source and the purpose of the secondary data is to replace faulty or missing primary data during quality checking. For example, a researcher may use ground data (SAURAN) as primary data and satellite data (Meteonorm) as secondary.

### 2.2: Performing quality checking

This involves analysing the primary data and identifying faulty or missing data, which will be replaced by secondary data. An easy way is to set up equations in Microsoft Excel (MS Excel) that flag faulty data points. As an example, the following data will be considered:


**Primary data:** Ground station  
**Source:** SAURAN  
**Resolution:** Hour  
**Site:** Stellenbosch University  
**Date:** 1 January 2018 – 31 December 2018

**Secondary data:** Satellite data  
**Source:** HelioClim-3 (SoDa-Pro)  
**Resolution:** Hour  
**Site:** Stellenbosch University  
**Date:** 1 January 2018 – 31 December 2018



### 2.2.1: Missing time stamps

The data may contain missing time stamps due to faults with the weather station. These faults may include power loss, battery failure, lightning strikes, etc. Here is an example of a dataset with missing hourly time stamps between 04:00 and 07:00:



	A	B	C	D	E	F	G	H
1	TOA5	SUN - Stel	Latitude: -	Longitude	Elevation: 119 m			SAURAN
2	Date	RecNum	GHI_CMP1	DNI_CHP1	DHI_CMP1	DHI_CMP1	UVA_Avg	UVB_Avg
3	TS	RN	W/m^2	W/m^2	W/m^2	W/m^2	W/m	W/m
4			Avg	Avg	Avg	Avg	Avg	Avg
5	2018/01/21 00:00	11816	0	0	0	0	0	0.011612
6	2018/01/21 01:00	11817	0	0	0	0	0	0.011615
7	2018/01/21 02:00	11818	0	0	0	0	0	0.011604
8	2018/01/21 03:00	11819	0	0	0	0	0	0.011612
9	2018/01/21 04:00	11820	0	0	0	0	0	0.011606
10	2018/01/21 07:00	11823	64.8533	143.033	41.543	39.6425	0.574733	0.304363
11	2018/01/21 08:00	11824	263.181	325.005	162.152	143.045	3.65767	1.1237
12	2018/01/21 09:00	11825	481.682	666.721	118.517	106.965	9.91196	2.14137

An easy way to check for missing data is to look at the hour value of each datapoint. The hour value may be retrieved using the following equation (for cell A5 from the figure above as an example):

**=HOUR(A5)**


**IMPORTANT NOTE:** If the equation above returns a “#VALUE!” statement in the cell, it means that MS Excel doesn’t recognise the information in the first column as a date and therefore cannot perform the date-related function. To fix this you will have to manually separate the date column out into individual components and then recombine it. This is shown in further detail in the next sub-section called “fixing date issues”.

Missing time stamps can then be flagged by comparing the hour values to the previous cell’s value, which should be either 1 hour or 23 hours (when the data passes into a new day). The following equation may be used (for cell A6 from the figure above as an example):

**=IF(OR((B6-B5)=1,(B5-B6)=23),"",1)**

These flags can then be used to fix the missing time stamps by inserting a line of blank data. This line of data can then be replaced with secondary data. The entire flag column can be searched for “1” via “values” to find the missing time stamps:

	A	B	C
1	TOA5		
2	Date		
3	TS		
4		HOUR	Flag
5	2018/01/21 00:00	0	
6	2018/01/21 01:00	1	
7	2018/01/21 02:00	2	
8	2018/01/21 03:00	3	
9	2018/01/21 04:00	4	
10	2018/01/21 07:00	7	1
11	2018/01/21 08:00	8	
12	2018/01/21 09:00	9	



	A	B	C
1	TOA5		
2	Date		
3	TS		
4		HOUR	Flag
5	2018/01/21 00:00	0	
6	2018/01/21 01:00	1	
7	2018/01/21 02:00	2	
8	2018/01/21 03:00	3	
9	2018/01/21 04:00	4	
10	2018/01/21 05:00	5	
11	2018/01/21 06:00	6	
12	2018/01/21 07:00	7	
13	2018/01/21 08:00	8	
14	2018/01/21 09:00	9	

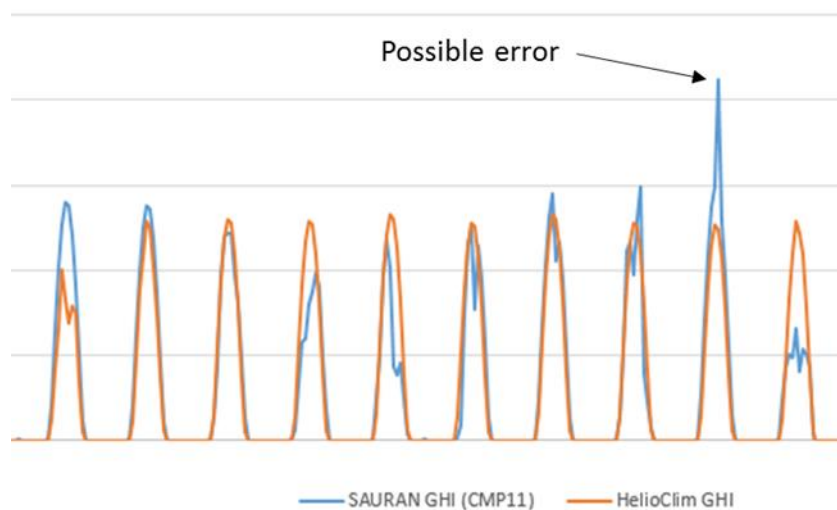
### 2.2.2: Outlying data

The data itself may be incorrect due to faulty readings by the instruments themselves. This may be checked in two ways. Firstly, if valid and invalid ranges are known for the data, an equation may be used to flag suspicious data. As an example, in Stellenbosch it is unlikely for the temperature to drop below -5°C and to increase higher than 55°C. Therefore, the following equation may be used:

**=IF(OR(D4<(-5),D4>55),1,"")**

The suspicious data should be considered, and if the researcher decides the data is incorrect, they should replace it with secondary data. This is to be done at the discretion of the researcher.

The second way to check the validity of the primary data is to plot it against the secondary data and look for large deviations. An example of this is shown here:



It must be noted that there will be natural deviation in values such as GHI since the satellite data is averaging cloud cover over a large area while the ground station is experiencing cloud cover as a single point. The figure above shows a large deviation from the trend. This data point should be considered as a *possible* error and the researcher should decide if the datapoint is plausible, otherwise, it should be replaced with secondary data.

The DNI and DNI calculated can also be checked by determining the DNI error as explained in the first section of this document titled “Quality checking and formatting tool for SAURAN data”. This may be implemented manually using MS Excel.

## 2.3: Fixing date issues

If date related function is used in MS Excel and it returns “#VALUE!”, it typically means that MS Excel doesn’t recognise the data as a date. This is shown here:

B5		=HOUR(A5)	
	A		B
4	TmStamp		
5	23/08/2020 00:00:00		#VALUE!
6	23/08/2020 01:00:00		
7	23/08/2020 02:00:00		

This is fixed by separating out the date into individual values and then recombining them as a date. Firstly, insert a blank column to the right of the date column. Next, highlight the entire date column and, in the “Data” tab at the top of the page, click on “Text to Columns”, which is in the

“Data Tools” group. In the window that pops up, make sure “Delimited” is selected and hit next. Under “Delimiters” make sure only “Space” is selected and then click “Finish”. This will separate the date and time as follows:

	A	B
4	<b>TmStamp</b>	
5	23/08/2020	00:00:00
6	23/08/2020	01:00:00
7	23/08/2020	02:00:00
8	23/08/2020	03:00:00
9	23/08/2020	04:00:00

We need to perform this again with the date. So next, add 2 blank columns to the right of the date column. Highlight the entire date column and hit Text to Columns > Delimited > Next and then this time instead of selecting “Space” as the Delimiter, select “Other” and enter a forward slash “/” to the right. Finally hit Finish to complete the action and the data should look like this:

	A	B	C	D
4	<b>TmStamp</b>			
5	23	8	2020	00:00:00
6	23	8	2020	01:00:00
7	23	8	2020	02:00:00
8	23	8	2020	03:00:00
9	23	8	2020	04:00:00

Finally, we can recombine the data back into a date by inserting a blank column and adding the following equation as per the example of the figure above:

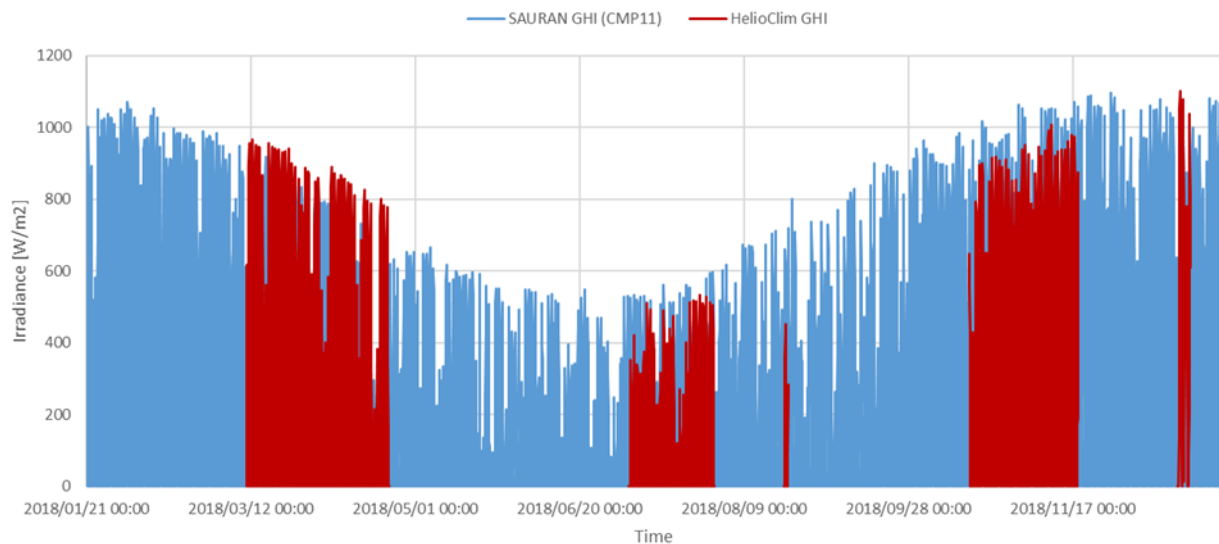
**=DATE(C5,B5,A5) + TIME(HOUR(D5),0,0)**

You can copy this new column and paste it back as “values” to remove its dependency on the other cells. You can then delete the original date data columns and work with the new date column. If the date isn’t displaying correctly, highlight the entire column of the new dates, right click on them, and select “Format Cells...”. In the “Number” tab, under “Category”, choose “Custom”. Then under “Type” to the right, scroll down and select “yyyy/mm/dd hh:mm” and then click OK. The data should now look like this:

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1		Latitude:	Longitude	Elevation: 119 m										
2		RecNum	BattV_Mir	TrackerWl	Tracker2W	ShadowW	SunWM_A	Shadowb	DNICalc_A	AirTC_Avg	RH	WS_ms_S	WindDir_I	Win
3			Volts	kW/m^2		kW/m^2	W/m^2	W/m^2		Deg C	%	meters/s	°Deg	Deg
4	TmStamp	RecNum	Min	Avg	Avg	Avg	Avg	Avg	Avg	Avg	Smp	WVc	WVc	WVc
5	2020/08/23 00:00	34496	13.71	0	0	0	0	0	0	7.984	89.9	0	0	0
6	2020/08/23 01:00	34497	13.7	0	0	0	0	0	0	7.742	93.1	0	0	0
7	2020/08/23 02:00	34498	13.71	0	0	0	0	0	0	7.943	89.9	0	0	0
8	2020/08/23 03:00	34499	13.7	0	0	0	0	0	0	7.426	91.8	0	0	0
9	2020/08/23 04:00	34500	13.7	0	0	0	0	0	0	7.272	95	0	0	0

## 2.4: Illustrating data allocation

The final step a researcher may take is to illustrate their dataset in a way that shows how much of their data is primary versus secondary data. This could be done as in a chart as follows:



## Contact Details

**Author of this report and Python tool:**

Don Fitzgerald

Email: [don@sun.ac.za](mailto:don@sun.ac.za)

**The Centre for Renewable and Sustainability Studies:**

4th Floor Knowledge Centre,  
Corner of Banghoek and Joubert Street,  
Stellenbosch, 7600

Email: [crses@sun.ac.za](mailto:crses@sun.ac.za)

Phone: +27 21 808 4069

