Facultad de Ingeniería - Universidad de la República

Calidad de Datos desde “small data” hacia “big data”: desafíos y técnicas
2017 Course Assignment

Author: Pablo Cerveñansky (C.I. 4.747.574-1)

Index

1. Introduction

Our Master's Thesis project consists in studying data compression techniques in Wireless Sensor Networks (WSNs). We are particularly interested in signals with certain temporal and/or spatial correlation among each other, since this feature allows us to achieve better compression ratios. In general, datasets gathered by WSNs which measure physical magnitudes (i.e. temperature, moisture, pressure) exhibit some degree of correlation, so we searched for datasets with that characteristic to analyse on the Data Quality course assignment.

The datasets we chose [1] include soil moisture and meteorological measurements from stations installed in the area surrounding Davos, Switzerland in the period between October 2010 and October 2013. This information was used by researchers interested in studying the role of soil moisture in relation with the snowpack runoff and catchment discharge in high alpine terrain [2]. There are 7 stations in total, labeled 1202, 1203, 1204, 1205, 222, 333 and SLF2, all of which can be located in the map of Figure 1.1.

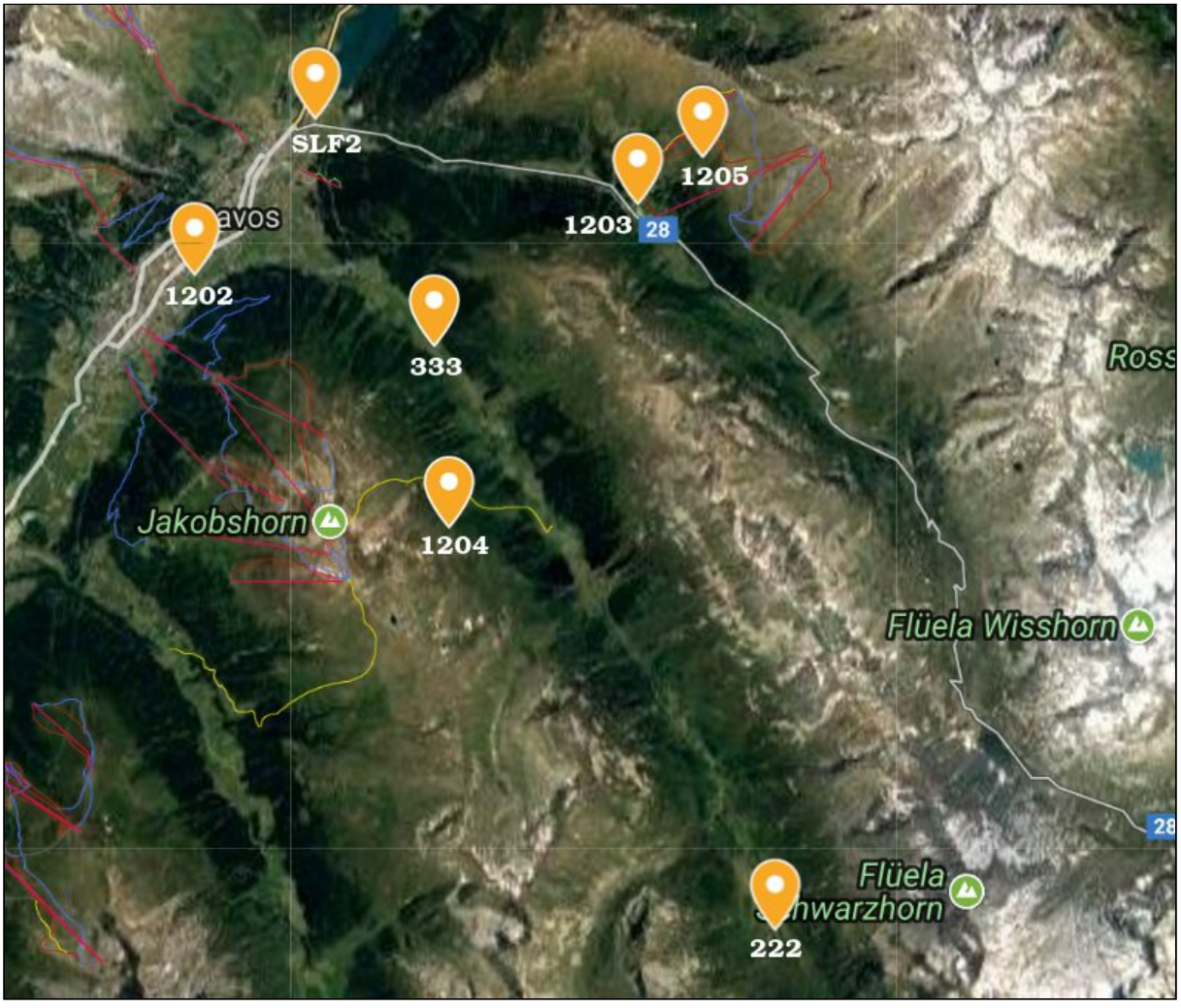

Figure 1.1: Map with the 7 stations in which the data was gathered. [3]

For each of the stations the following three datasets are provided:

1) The soil moisture measurements dataset is stored in files *vwc_[station-id].dat/.smet*. The data was recorded using Decagon 10HS sensors [4] installed at 10, 30, 50, 80 and 120 cm depth. There were two sensors installed per depth, labelled A and B in the files.

2) *station_[station-id].smet* data files contain in-situ measured meteorological parameters, such as air temperature, relative humidity, wind velocity and snow height.

3) *interpolatedmeteo_[station-id].smet* data files contain meteorological parameters data derived by interpolating data from several stations in the Davos area to the stations' location. The dataset was generated from the output of the *Alpine3D* model [6].

In [1] the researcher states that files with *.smet* extension must follow the *SMET* format [7]. In addition to the actual data, these files contain a header section which provides metadata regarding the station location and the magnitude unit of measurement for each column. *SMET* format also allows to specify an offset and a multiplier to be applied on the values of each column. For example, the offset can be used to handle Celsius to Kelvin conversion when measuring temperature.

Our original plan was to conduct a data quality assessment of each of the three datasets. We begun by analysing the first dataset, working in the data profiling, studying its data quality dimensions and then cleaning the data (Section 2). This work took us longer than expected so for the two remaining datasets the work is not as thorough. In both cases (Sections 3 and 4) we tried to focus on the characteristics that weren't already observed in the first dataset.

We are attaching the code with this document so all of our experiments can be easily reproduced. In every section we describe the command used to run each script. We used *python* version 2.7.13 but any 2.7.x version should work. The three datasets can be downloaded in a single file which is available here [1] under the title "Data and Resources".

2. Soil moisture dataset

2.1 Dataset overview

The first thing we noticed is that not every data file has the same file extension: some of the files have *.dat* extension (stations 1202, 1203, 1204 and 1205) and others have *.smet* extension (stations 222, 333 and SLF2). However, all of the files have the same structure, which is not consistent with the *SMET* format specifications in [7]. As an example, Figure 2.1 shows an overview of the *vwc_SLF2.smet* data file.

```
timestamp -10cm_A -30cm_A -50cm_A -80cm_A -120cm_A -10cm_B -30cm_B -50cm_B -80cm_B
-120cm_B
2010-10-01T00:00 -999.000000 -999.000000 -999.000000 ... -999.000000 -999.000000
2010-10-01T01:00 -999.000000 -999.000000 -999.000000 ... -999.000000 -999.000000
. . .
2012-03-09T22:00 0.383000 0.345000 0.316000 ... 0.466000 0.374000
2012-03-09T23:00 0.382000 0.344000 0.316000 ... 0.466000 0.374000
. . .
2013-09-30T23:00 -999.000000 -999.000000 -999.000000 ... -999.000000 -999.000000
2013-10-01T00:00 -999.000000 -999.000000 -999.000000 ... -999.000000 -999.000000
```

Figure 2.1: Overview of the *vwc_SLF2.smet* data file.

By analysing Figure 2.1 we can make the following observations:

- Only the first line has metadata (i.e. the name of the dataset columns) and the rest of the file consists of the data rows.
- There is no metadata describing which station the data file belongs too. This information is only available in the name of the file.
- By reading the dataset description in [1] it's straightforward knowing which column represents which data in the real world. For example, in data file *vwc_SLF2.smet* the column labeled "-50cm_A" represents soil moisture measurements for sensor labeled A at 50 cm depth in station SLF2.
- There is no metadata defining the unit of measurement for the soil moisture values. We had to read the Decagon 10HS sensor specification [4] to learn that these sensors measure the Soil Volumetric Water Content (VWC), which is equal to the volume of water divided by the volume of the sample. Higher VWC values mean more soil moisture.
- The string "-999.000000" is heavily repeated in the data rows. By reading the metadata of an actual *.smet* file (e.g. *interpolatedmeteo_1202.smet* from the third dataset) we realized that this string is used when there is no data available for the column in a particular timestamp (i.e. the string represents an unknown/null value).

2.2 Data profiling and data quality dimensions

In order to have a high order view of the dataset contained in the files we've decided to apply the data profiling techniques introduced in the course. We decided to use *pandas* [8], which is a highly customizable data analysis tool for Python.

We began by processing and gathering the following statistics for each data file:

- Number of columns and rows
- Number of null columns (i.e. columns with "-999.000000" in every row)
- Number of invalid rows (i.e. rows with missing columns, invalid values or duplicate datetime value)
- Number of null rows (i.e. rows with "-999.000000" in every column minus datetime)
- Timestamp of the first and last rows

The results obtained are presented in Table 2.1. The experiments can be reproduced by running the script with the following command: "python script.py vwc *folder_path*", where *folder_path* is the path of the folder in which the data files are stored.

Data file	# cols.	# null cols.	# rows	# invalid rows (%)	# null row (% of valid)	First timestamp	Last timestamp
vwc_1202.dat	10	0	26,305	0	10,438 (39.68%)	2010-10-01 00:00:00	2013-10-01 00:00:00
vwc_1203.dat	10	3	26,305	0	10,657 (40.51%)	2010-10-01 00:00:00	2013-10-01 00:00:00
vwc_1204.dat	10	1	26,305	0	8,899 (33.83%)	2010-10-01 00:00:00	2013-10-01 00:00:00
vwc_1205.dat	10	5	26,305	0	9,396 (35.72%)	2010-10-01 00:00:00	2013-10-01 00:00:00
vwc_222.smet	10	0	26,310	6 (0.02%)	809 (3.07%)	2010-10-01 00:00:00	2013-09-30 23:00:00
vwc_333.smet	10	0	26,094	6 (0.02%)	1,321 (5.06%)	2010-10-10 00:00:00	2013-09-30 23:00:00
vwc_SLF2.smet	10	0	26,311	6 (0.02%)	1,826 (6.94%)	2010-10-01 00:00:00	2013-10-01 00:00:00

Table 2.1: Statistics for each of the files in the soil moisture dataset.

Many observations can be made by looking at the data in Table 2.1. To begin with, we noticed that all of the data files have the same number of columns but there are three files with at least one null column.

Another observation is that the number of valid rows is almost the same for every data file. There are 5 files with exactly 26,305 valid rows in which the first and last timestamps match. The other two files are *vwc_222.smet* and *vwc_333.smet*. File *vwc_222.smet* has 26,304 valid rows because it seems to be missing the last row (i.e. the last timestamp is off by an hour, being "2013-09-30 23:00:00" instead of "2013-10-01 00:00:00"). If we compare file *vwc_333.smet* with the other files we notice that there is a 9 day offset for the first timestamp

and this data file also seems to be missing the last row. If we had these missing 217 (i.e. 9*24 + 1) rows, then the number of valid rows would also be 26,305.

In addition, we noticed that the number of null rows varies substantially among the data files. In the first four files the percentage is above 30% while in the last three it's under 10%. If we fill the missing rows with null rows in *vwc_222.smet* and *vwc_333.smet*, then the number of null rows becomes 810 (3.08%) and 1,538 (5.85%) respectively.

Also, it's important to notice that all of the invalid rows were due to a duplicated datetime value. For example, in Figure 2.2 we can see an example for file *vwc_222.smet*, in which there are four rows with duplicate timestamp "2011-10-01T00:00". We consider only three rows to be invalid because the values for every column match among the four rows. If the value for any column was different among any of the row pairs then we would have consider all four rows as invalid.

```
. . .
2011-09-30T23:00 0.515000 0.445000 0.323000 0.196000 0.403000 0.519000 0.422000 0.274000 0.185000 0.428000
2011-10-01T00:00 0.515000 0.445000 0.323000 0.196000 0.403000 0.519000 0.422000 0.274000 0.184000 0.427000
2011-10-01T00:00 0.515000 0.445000 0.323000 0.196000 0.403000 0.519000 0.422000 0.274000 0.184000 0.427000
2011-10-01T00:00 0.515000 0.445000 0.323000 0.196000 0.403000 0.519000 0.422000 0.274000 0.184000 0.427000
2011-10-01T00:00 0.515000 0.445000 0.323000 0.196000 0.403000 0.519000 0.422000 0.274000 0.184000 0.427000
2011-10-01T01:00 0.515000 0.445000 0.323000 0.196000 0.403000 0.519000 0.422000 0.274000 0.184000 0.427000
. . .
```

Figure 2.2: Section of the *vwc_SLF2.smet* data file with duplicate rows marked in red.

Next, we continued applying the data profiling technique and gathered stats (such as mean and standard deviation) for each of the sensor values in each data file. We used the *describe* method [9] from *pandas* library to get the results which we are presented in Tables P1-P7 in Section A.2 in the Appendix. Again, the experiments can be reproduced by running the script using command "python script.py vwc *folder_path*". This script also creates plots for each data file, which can be seen in Figures C1-C7 in Section A.1 in the Appendix.

We used the information from Tables 2.1 and P1-P7 and Figures C1-C7 to analyse the different data quality dimensions of the dataset:

(I) Accuracy

The accuracy of the dataset is highly dependant on the reliability of the sensors used to register the measurements, so we decided to study their characteristics. In Figure 2.3 we present some of the specs of the 10HS sensors. We can observe that the sensor measurement range is 0-57% VWC with an accuracy of ± 3% VWC. Having already calculated the minimum and maximum values for each sensor in Tables P1-P7 allowed us to do a quick sanity check. We found that no sensor has registered a value below 0% VWC, since the minimum value is 3.8% VWC for sensor "-10cm_A" in file *vwc_333.smet* (Table P6). However, there are two sensors which have maximum values above 57% VWC: for sensor

"-10cm_B" in file *vwc_1205.dat* (Table P4) the maximum is 59.9% VWC and for sensor "-30cm_B" in file *vwc_SLF2.dat* (Table P7) the maximum is 58.0% VWC. Since these values are only slightly above the measurement range we believe that there could have been a minor calibration issue with these two sensors.

ACCURACY	Apparent Dielectric Permittivity (ε_a): ± 0.5 from ε_a of 2 to 10, ± 2.5 from ε_a of 10 to 50 Soil Volumetric Water Content (VWC): Using standard calibration equation: ± 0.03 m^3/m^3(± 3% VWC) typical in mineral soils that have solution electrical conductivity < 10 dS/m; using soil specific calibration, ± 0.02 m^3/m^3 (± 2% VWC) in any soil
RESOLUTION	ε_a: 0.1 from ε_a of 1 to 30, 0.2 from ε_a of 30 to 50 VWC: 0.0008 m^3/m^3 (0.08% VWC) in mineral soils from 0 to 0.50 m^3/m^3 (0-50% VWC)
RANGE	ε_a: 1 (air) to 50 VWC: Calibration dependant; up to 0 - 57% VWC with polynomial equation

Figure 2.3: 10HS Large Soil Moisture Sensor specifications. More specs in [5].

Another point that we think it's worth mentioning is that sensors in different stations have different measurement precision. The sensors in stations 1202, 1203, 1204 and 1205 have a precision of 6 digits after the decimal (e.g. 0.347450 or 34.7450% VWC), while sensors in stations 222, 333 and SLF2 have a precision of 3 digits after the decimal (e.g. 0.313 or 31.3% VWC). The fact that all of the values recorded by sensors in a particular station have the same precision inclines us to think that there wasn't a fault in the sensors but in the way the data was gathered by the researchers. According to the specs in Figure 2.3 the sensor precision under certain conditions is 0.0008 or 0.08% VWC, so in every case using more than 3 digits after the decimal to represent a measured value seems an overkill.

Since in every station there are two sensors at each depth, we thought it would be interesting to compare the VWC measurements of sensors at the same depth to see if we can notice any accuracy issues. The way we plotted the data for each sensor separately in Figures C1-C7 is useful for making the comparison. Also, in Table 2.2 we compare the VWC mean for each sensor.

If we consider station 222, for example, both plots (see Figure C5) are similar and the (absolute) difference in the mean for sensors at the same depth is fairly low, i.e. 0.002 for -10cm, 0.027 for -30cm, 0.044 for -50cm, 0.027 for -80cm and 0.03 for -120cm. An analogous observation can be made for stations 333 and SLF2. However, in this case for the sensors at 120cm depth the differences are slightly higher, i.e. 0.11 in station 333 and 0.12 in station SLF2. This difference can be easily spotted in the respective plots (Figures C6 and C7). Station 1205 has a reduced number of sensors and the differences are not significant (see Figure C4).

In the remaining stations (1202, 1203 and 1204) the plots (Figures C1, C2 and C3) show more discrepancies, especially in station 1202. Considering the data in Table 2.2 we observe that not only are the differences higher but in some cases the tendencies change as the depth increases. The cell colours used in Table 2.2 allow us to easily spot this phenomena. For example, in station 1202 for sensors labeled A the mean always increases as we get deeper, but this is not the case for sensors labeled B, i.e. there is a decrease in the mean from -50cm to -80cm and from -80cm to -120cm. We cannot be certain if this is an accuracy issue since we are not quite familiar with the physical magnitude that is being measured (i.e. VWC) and whether it can have fluctuating values for sensors near each other. There is also the possibility that distances between sensors at the same depth changes among the different stations. We already mentioned that the precision for sensors in each station is not homogeneous, which leds us to believe that different researchers were in charge of different stations. This could also help to explain a possible difference in the distance between sensors at the same depth.

Sensor	Mean for vwc_1202.dat	Mean for vwc_1203.dat	Mean for vwc_1204.dat	Mean for vwc_1205.dat	Mean for vwc_222.smet	Mean for vwc_333.smet	Mean for vwc_SLF2.smet
-10cm_A	0.389	0.288	0.383	0.439	0.525	0.309	0.356
-30cm_A	0.413	0.249	0.314	0.312	0.452	0.342	0.341
-50cm_A	0.414	0.225	0.295	0.315	0.315	0.333	0.314
-80cm_A	0.426	0.217	0.267	-	0.198	0.204	0.453
-120cm_A	0.436	-	0.250	-	0.400	0.319	0.256
-10cm_B	0.250	0.290	0.380	0.465	0.523	0.363	0.399
-30cm_B	0.345	0.214	0.250	0.334	0.425	0.340	0.352
-50cm_B	0.476	0.224	0.243	-	0.271	0.326	0.311
-80cm_B	0.449	-	0.266	-	0.171	0.210	0.466
-120cm_B	0.437	-	-	-	0.430	0.209	0.376

Table 2.2: Changes in the mean of the VWC measured by each sensor as depth increases. A cell is green (red) if the mean of the VWC measured by the sensor is higher (lower) than the mean of the VWC measured by the sensor immediately above it. Data was gathered from Tables P1-P7.

We've already seen that there are duplicate invalid rows in three of the data files (see Table 2.1). The way we see it, this error was probably introduced by a researcher while processing the data files and was not caused by a sensor failure when recording the measurements, and this doesn't have a negative impact in the accuracy of the dataset since there is no missing or corrupt data.

All in all, we believe that the sensors used to gather the values are reliable and in spite of the minor accuracy issues we've mentioned we consider the dataset to be a fairly accurate representation of events in the real world.

(II) Completeness

In Tables 2.1 and P1-P7 we've seen that many entries in the dataset have null values. In the course we learnt that there are three cases in which to use null values: (1) when a piece of data doesn't exist, (2) when it exists but it is unknown and (3) when we ignore whether the value exists. Clearly all of the null values in our dataset belong to case (2), i.e. the VWC value exists in every moment in every place in the soil, we just ignore the value for some timestamps in our dataset.

In Table 2.3 we present the number of null values and its percentage over the total valid rows for each sensor. The sensors in station SLF2 are the ones with less percentage of null values, with every sensor having no more than 25% of null values. The stats for sensors in stations 222 and 333 are less homogeneous but in each case at least half of the sensors have less than 10% of null values. In the remaining stations (1202, 1203 and 1204) the percentage of null values for each sensor is generally high, always above 30%.

Sensor	# null values (% of valid) vwc_1202.dat	# null values (% of valid) vwc_1203.dat	# null values (% of valid) vwc_1204.dat	# null values (% of valid) vwc_1205.dat	# null values (% of valid) vwc_222.smet	# null values (% of valid) vwc_333.smet	# null values (% of valid) vwc_SLF2.smet
-10cm_A	19,717 (75%)	10,657 (40.5%)	8,899 (33.8%)	22,708 (86.3%)	3,969 (15.1%)	1,738 (6.7%)	2,069 (7.9%)
-30cm_A	10,438 (39.7%)	20,140 (76.6%)	13,325 (50.6%)	15,293 (58.1%)	3,798 (14.4%)	3,389 (13.0%)	3,146 (12.0%)
-50cm_A	10,438 (39.7%)	10,687 (40.6%)	12,821 (48.7%)	9,450 (35.9%)	15,440 (58.7%)	7,967 (30.5%)	1,828 (6.9%)
-80cm_A	10,648 (40.5%)	26,303 (99.9%)	26,292 (99.9%)	-	3,425 (13.0%)	1,820 (7.0%)	1,827 (6.9%)
-120cm_A	10,438 (39.7%)	-	26,303 (99.9%)	-	12,513 (47.6%)	7,617 (29.2%)	2,069 (7.9%)
-10cm_B	21,759 (82.7%)	10,658 (40.5%)	13,323 (50.6%)	22,590 (85.8%)	811 (3.1%)	3,215 (12.3%)	6,512 (24.8%)
-30cm_B	10,438 (39.7%)	19,745 (75.1%)	9,877 (37.5%)	22,721 (86.3%)	813 (3.1%)	7,806 (29.9%)	6,542 (24.9%)
-50cm_B	10,438 (39.7%)	22,946 (87.2%)	15,214 (57.8%)	-	811 (3.1%)	1,322 (5.1%)	6,270 (25.0%)
-80cm_B	10,438 (39.7%)	-	26,298 (99.9%)	-	13,616 (51.8%)	1,322 (5.1%)	6,275 (23.9%)
-120cm_B	10,438 (39.7%)	-	-	-	811 (3.1%)	1,324 (5.1%)	6,511 (24.8%)

Table 2.3: Number of null values (and percentage over the valid rows) for each sensor. Green cells have percentages between 0-10%, light yellow cells between 10-25%, orange cells 25-50% and red cells 50-100%. Data was gathered from Tables P1-P7.

(III) Consistency

We haven't encounter any critical consistency issues while studying the dataset. The file extension differs among the files (some are *.dat* and some are *.smet*) but the content is consistent and in every file the timestamp and the soil moisture measurements have the same format. Also, the order of the columns and the string used to represent a null value (i.e. "-999.000000") is the same in every file.

The closer we've come to finding a consistency problem was in the case of the invalid rows. As we've already explained, we found many rows with a duplicate timestamp, but every other value in the row was also duplicate so we don't think that introduced any inconsistencies to the dataset.

(IV) Timeliness

The way we see it, this concept is heavily dependant on the task someone wishes to achieve with the information in the dataset. According to [2] the dataset was used to study the degree in which soil moisture is related to flood risks in the Davos alpine area. Since the other two remaining datasets cover the same time period, we think that the dataset fulfills its mission. We must also mention that there is no volatility risk since the dataset is concerned with a time frame that is entirely in the past.

(V) Interpretability

From our perspective, the dataset could use some improvement in this dimension. All of the data files are missing metadata which could help the end user to better interpret the data rows. We think that the unit of measurement for the soil moisture values and the location of the station (i.e. latitude, longitude and altitude) should be stated explicitly somewhere in the data files, as is the case with datafiles in the other two datasets. Also, the fact that the string "-999.000000" represents unknown values should be explained somewhere. We think that using this "number" string could lead to confusion and we would prefer using a more self-explanatory string such as "nodata" or "null".

2.3 Data cleaning

After learning about the quality dimensions of the dataset we decided to create a data cleaning script to improve its quality. The script can be run using the following command: "python script.py vwc *input_path* clean *output_path*", where *input_path* is the path of the folder in which the data files are stored and *output_path* is the folder in which the clean data files are to be saved.

This is a list of the improvements made in each of the data files:

- We followed the *SMET* format specification and added a header with station metadata (e.g. station identifier, latitude, longitude). We took this information from the headers of the files from the second dataset.

- We removed the duplicate rows and filled the missing timestamps so that there are no gaps between the first ('2010-10-01 00:00:00') and last ('2013-10-01 00:00:00') rows. This means that every clean data file has exactly 26,305 data entries, one per hour. The correctness can be checked by doing the math:
  #days * #hours_per_day + 1 =(365*2 + 366)*24 + 1 = 26,305.
- Instead of representing null values with the string "-999.000000" we used "NULL". This information is represented by the "nodata" key in the file header.
- Changed the file extension to *.clean.smet* so that all the clean files have the same extension.

As an example, Figure 2.4 shows an overview of the *vwc_SLF2.clean.smet* data file, which is the file created when applying the data cleaning script on the original *vwc_SLF2.smet* file.

```
SMET 1.1 ASCII
[HEADER]
station_id   = SLF2
station_name = SLF2
latitude     = 46.812365
longitude    = 9.847212
altitude     = 1560
easting      = 783800
northing     = 187400
epsg         = 21781
nodata       = NULL
tz           = 0
timestamp -10cm_A -30cm_A -50cm_A -80cm_A -120cm_A -10cm_B -30cm_B -50cm_B -80cm_B
-120cm_B
[DATA]
2010-10-01T00:00 NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL
2010-10-01T01:00 NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL
. . .
2011-09-30T23:00 0.346 0.336 0.305 0.455 0.259 0.394 0.36 0.314 0.469 0.38
2011-10-01T00:00 0.346 0.336 0.305 0.455 0.259 0.394 0.363 0.314 0.469 0.38
2011-10-01T01:00 0.346 0.336 0.305 0.455 0.259 0.394 0.365 0.314 0.469 0.38
. . .
2013-09-30T23:00 NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL
2013-10-01T00:00 NULL NULL NULL NULL NULL NULL NULL NULL NULL NULL
```

Figure 2.4: Overview of the *vwc_SLF2.clean.smet* data file.

If we compare it to Figures 2.1 and 2.2, which show different sections of the original file, we can notice the informative header and the lack of duplicate data rows.

3. Meteorological parameters dataset

The files from this dataset follow the *SMET* specification and so the code for processing them is slightly more complex than the one from Section 2. We had to add the ability to parse information from the metadata and apply the offset and multiplier to each column. The command to run the processing script is "python script.py station *folder_path*", where *folder_path* is the path of the folder in which the *station_[station-id].smet* data files are stored.
The script gathered the same statistics than the one described in Section 2.2. We present the data next in Table 3.1.

Data file	# cols.	# null cols.	# rows	# invalid rows (%)	# null row (% of valid)	First timestamp	Last timestamp
station_1202.smet	9	0	958,488	1 (0.0001%)	0	2010-11-19 15:04:00	2013-05-21 12:53:00
station_1203.smet	9	0	1,024,089	0	0	2011-03-18 10:31:00	2013-12-27 09:11:00
station_1204.smet	9	0	1,635,534	0	0	2010-10-29 14:13:00	2014-05-06 12:22:00
station_1205.smet	9	0	1,297,671	0	0	2011-05-11 00:00:00	2014-05-06 12:23:00
station_222.smet	16	0	52,608	0	0	2010-10-01 00:00:00	2013-09-30 23:30:00
station_333.smet	16	0	52,608	0	0	2010-10-01 00:00:00	2013-09-30 23:30:00
station_SLF2.smet	17	0	84,865	0	58 (0.07%)	2009-09-29 00:00:00	2014-08-02 00:00:00

Table 3.1: Statistics for each of the files in the meteorological parameters dataset.

The first thing we noticed is that the number of columns is not the same for each file, and the numbers seem to indicate than the first four files (9 columns with the same magnitudes) might have a different source than the last three (16 matching columns with the same magnitudes, with an extra one in *station_SLF2.smet*). This is one of the differences between Table 3.1 and Table 2.1. Another difference is that in this case there are no null columns.

Data files *station_222.smet* and *station_333.smet* are the only ones with the same number of rows. In addition, they have equal first and last timestamps. All of the other files have a highly fluctuating number of rows and there isn't a single pair of files with the same first or last timestamp. Again, this is a big difference in regard to Table 2.1, in which these stats were more homogeneous among the different data files.

In addition, data files *station_222.smet* and *station_333.smet* do not have any null value in any column, while the rest of the files have several null values in most columns. This statistics were also gathered by the script but we didn't present them in a table because we didn't have time to analyse them further.

There is a single invalid row, in file *station_1202.smet*, which we present in Figure 3.1. The issue is caused by a number with two dots. In the dataset analysed in Section 2 we noticed a number of duplicate data rows, but we hadn't encounter this kind of problem. In Figure 3.1 we can also observe the *SMET* file header, which has useful metadata. This header was missing in the first dataset but we added it in the data cleaning process in Section 2.3.

```
SMET 1.1 ASCII
[HEADER]
station_id   = station_1202
station_name = Golf_course
latitude     =      46.7968286797
longitude    =      9.8297824141
altitude     =    1537
easting      =  782523
northing     =  185633
epsg         =   21781
nodata       =    -999
tz           =       0
fields = timestamp TA RH VW DW ISWR RSWR OLWR PSUM TSS
[DATA]
2010-11-19T15:04 272.55 0.793854 0.5079 207.174 5.1892 18.0097 -999 0.0 271.063
2010-11-19T15:05 272.55 0.797107 0.0335 207.781 1.8315 10.0732 -999 0.0 270.875
. . .
2011-04-14T13:52 276.92 0.425036 1.7031 30.9755 12.8205 348.596 -999 0.0 287.562
2011-04-14T13:53 277.07 0.41136 1.465.4335 0.4661 0.4335 0.2477 0.423 0.4335 273.485
2011-04-14T21:27 272.05 0.816356 0.5448 27.4183 0.0 0.3052 -999 0.0 273.75
. . .
2013-05-21T12:52 -725.85 -9.99 -999 -999 -999 -999 -999 -999 -725.85
2013-05-21T12:53 -725.85 -9.99 -999 -999 -999 -999 -999 -999 -725.85
```

Figure 3.1: Overview of the *station_1202.smet* data file, including the invalid data row.

We didn't have enough time to analyse the data quality dimensions of this dataset and we don't think any conclusions can be made from the observations made so far.

4. Interpolated meteorological parameters dataset

The files in this dataset had the most homogeneous statistics among the three datasets. Since the data files also follow the *SMET* format the script for processing them is almost the same as the one in Section 3, the only change being the name of the files. It can be run with "python script.py interpolatedmeteo *folder_path*", where *folder_path* is the path of the folder in which the *interpolatedmeteo_[station-id].smet* data files are stored.

All of the 7 data files have the following statistics:

- Number of columns: 12
- Number of null columns: 0
- Number of rows: 43,826
- Number of invalid rows: 0
- Number of null rows: 0
- First timestamp: 2009-10-01 01:00:00
- Last timestamp: 2014-10-01 02:00:00

In every file there's a row for each hour and there are no row gaps between the first and last timestamps since: #days * #hours_per_day + 2 =(365*4 + 366)*24 + 2 = 43,826.
Also, we noticed that this is the only dataset in which there isn't a single null value in any column in any data file. The way we see it, this is due to the fact that this dataset was not generated from sensors in the field, but instead the researchers used the *Alpine3D* model [6] to interpolate data from nearby stations.

We didn't have enough time to analyse the data quality dimensions of the dataset. However, we think that we would need an understanding of the *Alpine3D* model and access to the field data from the nearby stations, which is not provided, to assess the data quality more thoroughly.

A. Appendix

A.1. Soil moisture dataset: charts

Displayed below are the Charts C1-C7 with all of the data gathered by each of the stations of the soil moisture dataset (Section 2). To improve the visualization we divided the data from sensors with labels A and B into two separate plots for each station. We used the same scale and range for both axis in every chart so they can be easily compared.

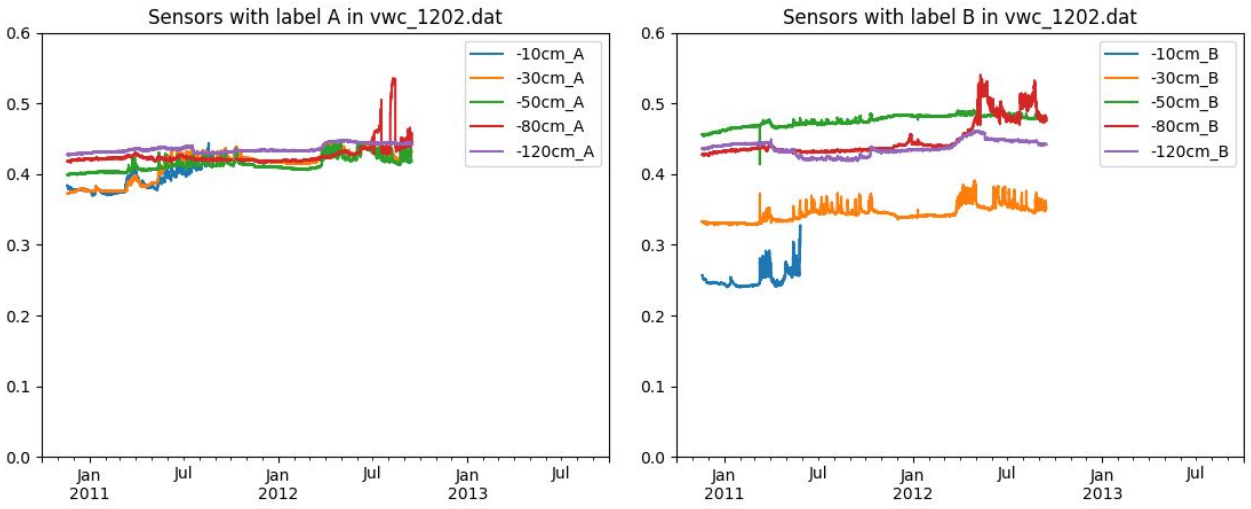

Figure C1: Chart for data file *vwc_1202.dat*.

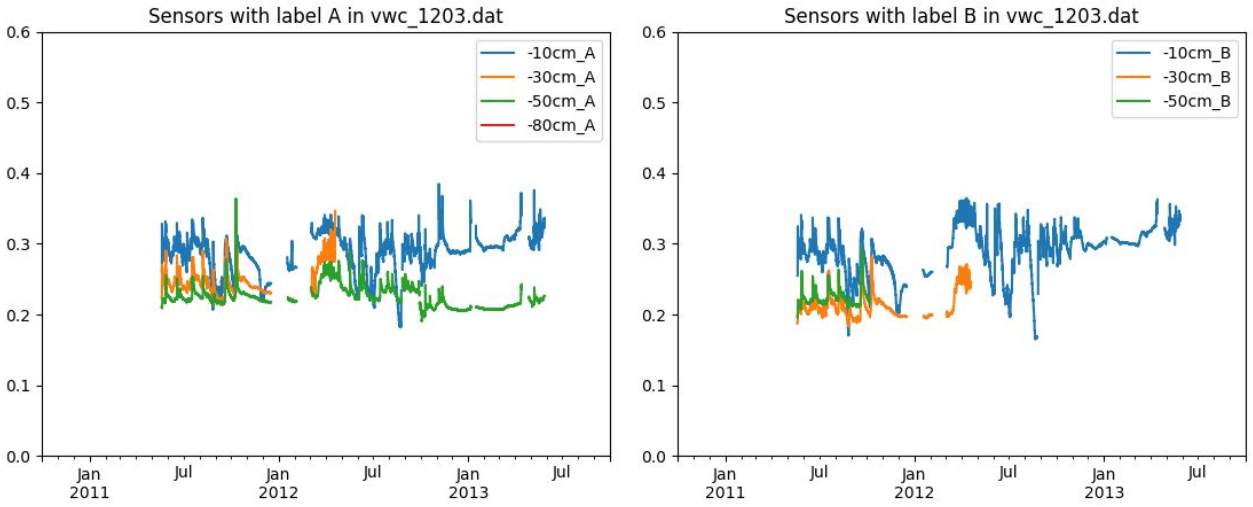

Figure C2: Chart for data file *vwc_1203.dat*.

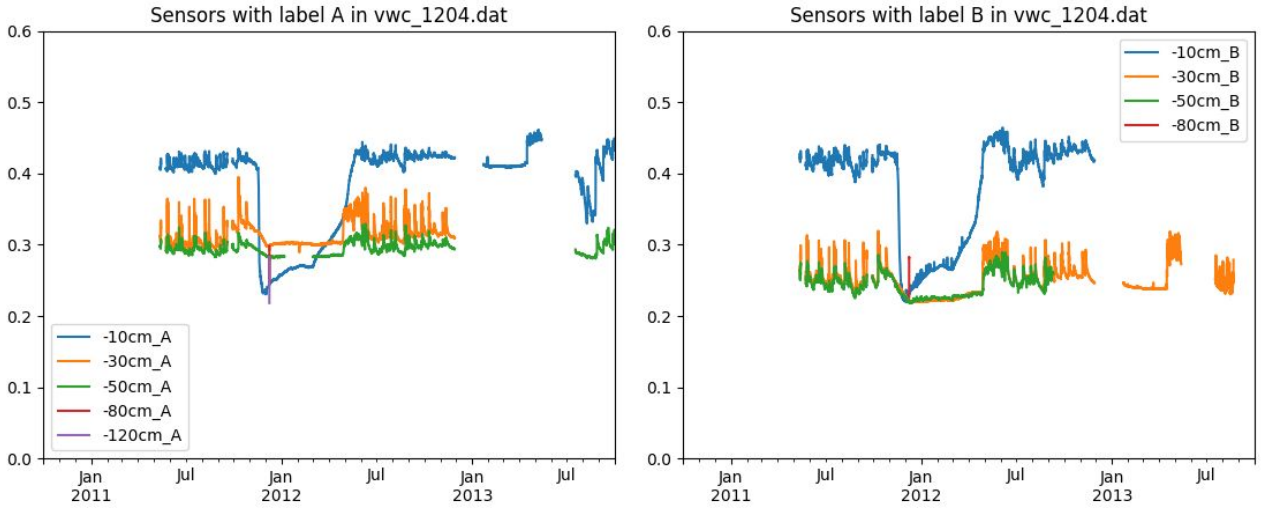

Figure C3: Chart for data file *vwc_1204.dat*.

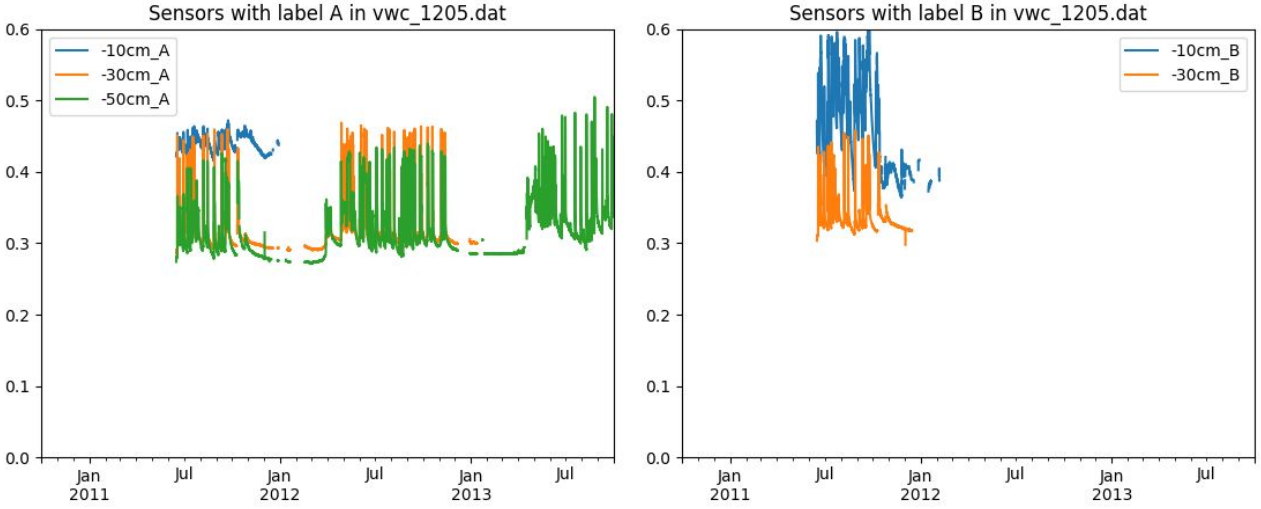

Figure C4: Chart for data file *vwc_1205.dat*.

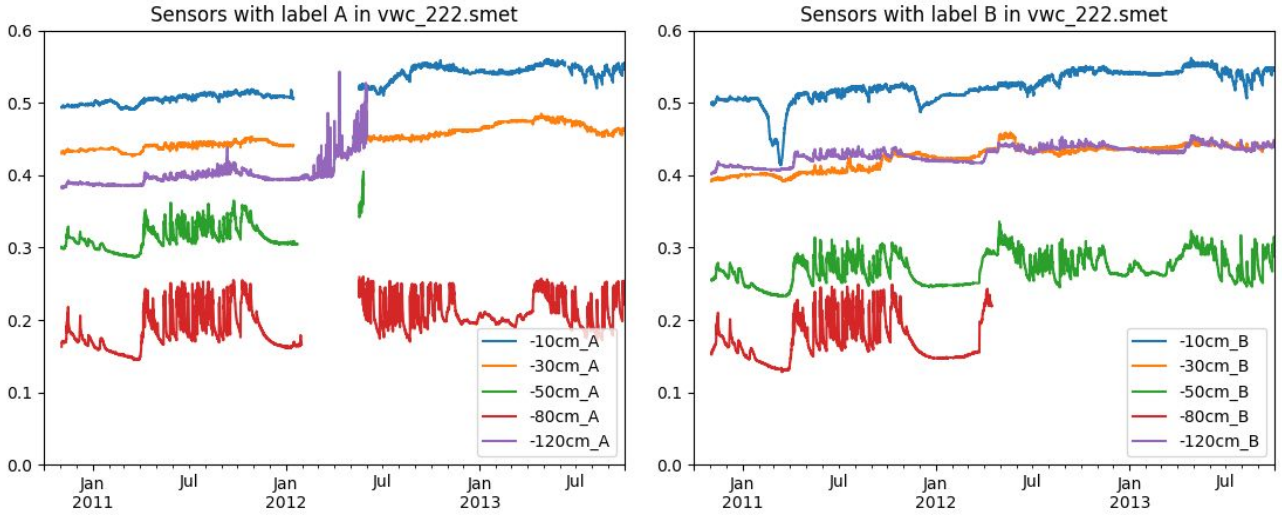

Figure C5: Chart for data file *vwc_222.smet*.

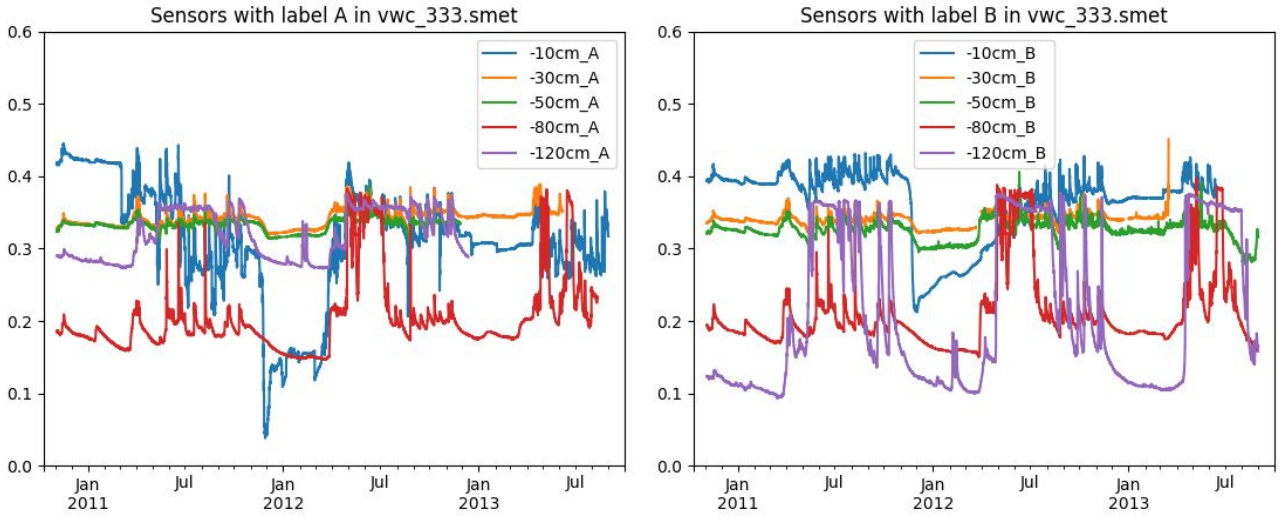

Figure C6: Chart for data file *vwc_333.smet*.

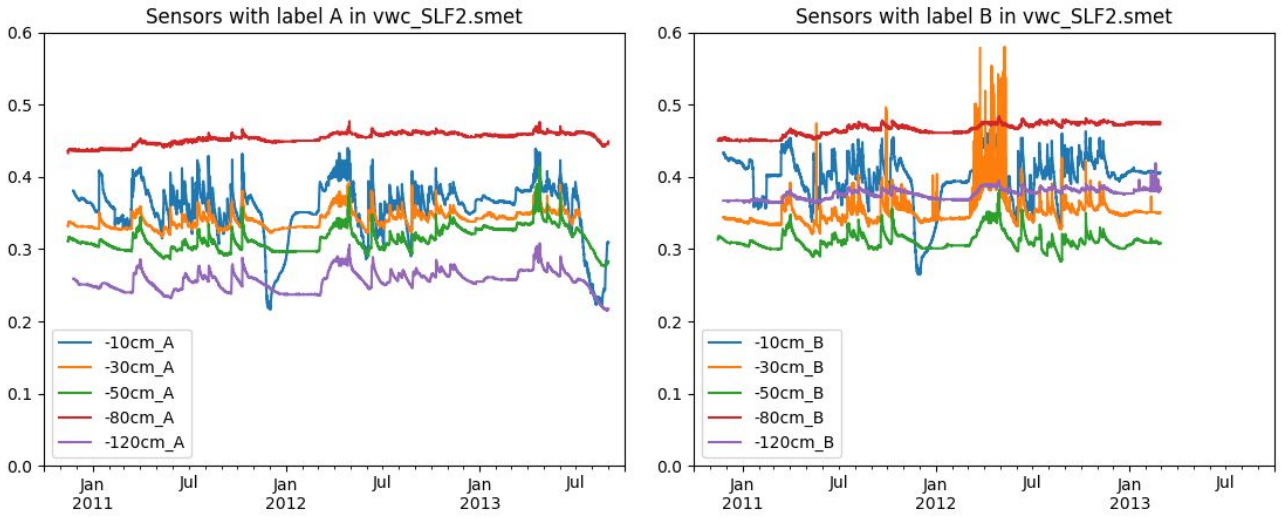

Figure C7: Chart for data file *vwc_SLF2.smet*.

A.2 Soil moisture dataset: data profiling tables

Below we present Tables P1-P7 with the data profiling information for each of the stations in the soil moisture dataset (Section 2). For each sensor we display the total number of null values and its percentage among the quantity of valid rows, the minimum and maximum values, the mean, the median and the standard deviation.

Sensor	# null values (% of valid rows)	Min.	Mean	Median	Max.	Standard deviation
-10cm_A	19,717 (75%)	0.369	0.389	0.384	0.443	0.015
-30cm_A	10,438 (39.7%)	0.372	0.413	0.419	0.446	0.022
-50cm_A	10,438 (39.7%)	0.398	0.414	0.411	0.444	0.010
-80cm_A	10,648 (40.5%)	0.416	0.426	0.422	0.535	0.015
-120cm_A	10,438 (39.7%)	0.426	0.436	0.434	0.449	0.006
-10cm_B	21,759 (82.7%)	0.240	0.250	0.246	0.328	0.011
-30cm_B	10,438 (39.7%)	0.328	0.345	0.344	0.391	0.011
-50cm_B	10,438 (39.7%)	0.414	0.476	0.478	0.490	0.008
-80cm_B	10,438 (39.7%)	0.426	0.449	0.436	0.541	0.025
-120cm_B	10,438 (39.7%)	0.418	0.437	0.436	0.461	0.010

Table P1: Data profiling for file *vwc_1202.dat* (# valid rows = 26,305).

Sensor	# null values (% of valid rows)	Min.	Mean	Median	Max.	Standard deviation
-10cm_A	10,657 (40.5%)	0.182	0.288	0.293	0.385	0.028
-30cm_A	20,140 (76.6%)	0.217	0.249	0.245	0.346	0.019
-50cm_A	10,687 (40.6%)	0.191	0.225	0.222	0.364	0.014
-80cm_A	26,303 (99.9%)	0.204	0.217	0.217	0.229	0.018
-10cm_B	10,658 (40.5%)	0.165	0.290	0.297	0.365	0.033
-30cm_B	19,745 (75.1%)	0.183	0.214	0.209	0.282	0.017
-50cm_B	22,946 (87.2%)	0.195	0.224	0.221	0.295	0.011

Table P2: Data profiling for file *vwc_1203.dat* (# valid rows = 26,305).

Sensor	# null values (% of valid rows)	Min.	Mean	Median	Max.	Standard deviation
-10cm_A	8,899 (33.8%)	0.230	0.383	0.411	0.462	0.061
-30cm_A	13,325 (50.6%)	0.289	0.314	0.309	0.395	0.015
-50cm_A	12,821 (48.7%)	0.278	0.295	0.294	0.329	0.009
-80cm_A	26,292 (99.9%)	0.222	0.267	0.259	0.331	0.033
-120cm_A	26,303 (99.9%)	0.218	0.250	0.250	0.283	0.046
-10cm_B	13,323 (50.6%)	0.219	0.380	0.416	0.464	0.071
-30cm_B	9,877 (37.5%)	0.220	0.250	0.247	0.320	0.021
-50cm_B	15,214 (57.8%)	0.218	0.243	0.243	0.289	0.016
-80cm_B	26,298 (99.9%)	0.227	0.266	0.262	0.325	0.035

Table P3: Data profiling for file *vwc_1204.dat* (# valid rows = 26,305).

Sensor	# null values (% of valid rows)	Min.	Mean	Median	Max.	Standard deviation
-10cm_A	22,708 (86.3%)	0.416	0.439	0.439	0.472	0.010
-30cm_A	15,293 (58.1%)	0.285	0.312	0.307	0.469	0.026
-50cm_A	9,450 (35.9%)	0.271	0.315	0.305	0.505	0.037
-10cm_B	22,590 (85.8%)	0.364	0.465	0.466	0.599	0.061
-30cm_B	22,721 (86.3%)	0.298	0.334	0.328	0.458	0.021

Table P4: Data profiling for file *vwc_1205.dat* (# valid rows = 26,305).

Sensor	# null values (% of valid rows)	Min.	Mean	Median	Max.	Standard deviation
-10cm_A	3,969 (15.1%)	0.491	0.525	0.520	0.560	0.021
-30cm_A	3,798 (14.4%)	0.426	0.452	0.450	0.485	0.014
-50cm_A	15,440 (58.7%)	0.286	0.315	0.312	0.405	0.018
-80cm_A	3,425 (13.0%)	0.145	0.198	0.198	0.260	0.027
-120cm_A	12,513 (47.6%)	0.382	0.400	0.397	0.543	0.017
-10cm_B	811 (3.1%)	0.414	0.523	0.522	0.562	0.022
-30cm_B	813 (3.1%)	0.392	0.425	0.432	0.459	0.166
-50cm_B	811 (3.1%)	0.233	0.271	0.269	0.336	0.019
-80cm_B	13,616 (51.8%)	0.128	0.171	0.165	0.249	0.027
-120cm_B	811 (3.1%)	0.401	0.430	0.432	0.455	0.011

Table P5: Data profiling for file *vwc_222.smet* (# valid rows = 26,304).

Sensor	# null values (% of valid rows)	Min.	Mean	Median	Max.	Standard deviation
-10cm_A	1,738 (6.7%)	0.038	0.309	0.310	0.445	0.787
-30cm_A	3,389 (13.0%)	0.320	0.342	0.343	0.389	0.011
-50cm_A	7,967 (30.5%)	0.310	0.333	0.335	0.389	0.009
-80cm_A	1,820 (7.0%)	0.147	0.204	0.190	0.382	0.051
-120cm_A	7,617 (29.2%)	0.272	0.319	0.317	0.376	0.034
-10cm_B	3,215 (12.3%)	0.212	0.363	0.380	0.432	0.049
-30cm_B	7,806 (29.9%)	0.321	0.340	0.341	0.451	0.009
-50cm_B	1,322 (5.1%)	0.275	0.326	0.327	0.406	0.014
-80cm_B	1,322 (5.1%)	0.150	0.210	0.194	0.397	0.052
-120cm_B	1,324 (5.1%)	0.093	0.209	0.159	0.375	0.106

Table P6: Data profiling for file *vwc_333.smet* (# valid rows = 26,088).

Sensor	# null values (% of valid rows)	Min.	Mean	Median	Max.	Standard deviation
-10cm_A	2,069 (7.9%)	0.216	0.356	0.364	0.440	0.039
-30cm_A	3,146 (12.0%)	0.317	0.341	0.341	0.393	0.012
-50cm_A	1,828 (6.9%)	0.277	0.314	0.311	0.414	0.016
-80cm_A	1,827 (6.9%)	0.433	0.453	0.455	0.477	0.008
-120cm_A	2,069 (7.9%)	0.215	0.256	0.257	0.308	0.014
-10cm_B	6,512 (24.8%)	0.265	0.399	0.406	0.463	0.03
-30cm_B	6,542 (24.9%)	0.321	0.352	0.348	0.580	0.018
-50cm_B	6,270 (25.0%)	0.283	0.311	0.309	0.382	0.012
-80cm_B	6,275 (23.9%)	0.449	0.466	0.468	0.484	0.009
-120cm_B	6,511 (24.8%)	0.364	0.376	0.377	0.419	0.007

Table P7: Data profiling for file *vwc_SLF2.smet* (# valid rows = 26,305).

B. References

[1] Nander Wever (2017): IRKIS Soil moisture measurements Davos; SLF; doi:10.16904/17. Link: http://www.envidat.ch/dataset/soil-moisture-measurements-davos

[2] Wever, N., Comola, F., Bavay, M., and Lehning, M.: Influence of snow surface processes on soil moisture dynamics and streamflow generation in alpine catchments, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2016-601, in review, 2017

[3] Google map with the location of the 7 stations https://drive.google.com/open?id=1DzZxUjD-TSxWX5w2iQmda_4g1HM&usp=sharing

[4] 10HS Large Soil Moisture Sensor: http://www.decagon.com/en/soils/volumetric-water-content-sensors/10hs-large-volume-vwc/

[5] 10HS Large Soil Moisture Sensor - Specifications: http://www.decagon.com/en/soils/volumetric-water-content-sensors/10hs-large-volume-vwc/#Specifications

[6] *Alpine3D*: https://models.slf.ch/p/alpine3d/

[7] *SMET* specifications: https://models.slf.ch/docserver/meteoio/SMET_specifications.pdf

[8] *pandas* library: http://pandas.pydata.org/

[9] *describe* method: https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.describe.html