Outlier detection methods and sensor data fusion for precision agriculture

Andrei B. B. Torres[1], **José Adriano Filho**[1], **Atslands R. da Rocha**[1],
Rubens Sonsol Gondim[2], **José Neuman de Souza**[1]

[1]Grupo de Redes de Computadores, Engenharia de Software e Sistemas (GREat)
Universidade Federal do Ceará (UFC) – Fortaleza, CE

[2]Embrapa Agroindústria Tropical – Fortaleza, CE

{andreitorres,atslands}@great.ufc.br, adriano.liukin@gmail.com
rubens.gondim@embrapa.br, neuman@ieee.org

Abstract. *Precision agriculture is a concept regarding the use of technology to increase production yield while preserving and optimizing resources. One of the means to achieve that goal is to use sensors to monitor crops and adjust the cultivation according to its needs. This paper compares different techniques for sensor data fusion and detection and removal of outliers from gathered data to improve sensors accuracy and to identify possible sensor malfunction. As a case study, we monitored an experimental crop of precocious dwarf cashew using soil moisture sensors. Combining generalized ESD method and a weighted outlier-robust Kalman filter generated the best result, leading to more accurate data.*

1. Introduction

Precision Agriculture is “a management system that is information and technology based, is site specific and uses one or more of the following sources of data: soils, crops, nutrients, pests, moisture, or yield, for optimum profitability, sustainability, and protection of the environment” [Natural Resources Conservation Service, 2007].

In Brazil, The Brazilian Agricultural Research Corporation (in Portuguese, Empresa Brasileira de Pesquisa Agropecuária - Embrapa) recommends the use of technology as the way to make agricultural growth economically viable and sustainable, particularly due to the challenges faced by the country, such as the rational use of water, need for increased production, climate change and production efficiency [Bernardi et al., 2014].

In this context, Internet of Things (IoT) technology provides support to managing and monitoring crops, apart from optimizing the usage of resources such as water, pesticide, and fertilizers. One valuable resource that allows for a more efficient usage of water is the soil conditioner (e.g.: hydrogel), that can improve the soil moisture retention. However, it requires constant soil moisture monitoring to guarantee its efficiency, which can be provided by sensors on an IoT network.

IoT technologies have made viable the usage of a large number low-cost sensors for a fraction of the price of a high-cost specialized sensor [Sanchez et al., 2013]. Although specialized sensors have high accuracy, they also tend to have high initial and maintenance costs, which may make a large scale implementation unfeasible. Low-cost sensors do not have the same accuracy, but by applying data fusion techniques, its usage becomes feasible, by using many sensors to gather more precise data.

The goal of this research is to compare outlier detection methods as fusion techniques, and filtering algorithms to compensate the low accuracy and inherent noise in sensors readings to output a more accurate data for precision agriculture. As a case study we monitored Embrapa's experimental crop of precocious dwarf cashew using soil moisture sensors.

2. Related Work

An systematic review was performed to locate relevant work about data fusion that focused on outlier detection and improving the accuracy of sensors. Ravichandran & Arulappan [2013] propose an algorithm for validating data from Wireless Sensor Network (WSN), and the authors used real data samples from a prototype experiment of environmental monitoring. They suggested an algorithm using a combination of techniques to improve the detection of different data faults. Their proposed algorithm does not consider data fusion or signal filtering and smoothing.

Callegaro et al. [2014] propose an architecture for information fusion focused on low-cost sensors in a WSN. It targets increasing data reliability, performing a case study with atmospheric pressure sensors, focusing on outlier detection and comparing various techniques, but it does not deal with filtering and smoothing.

Lastly, Andrade et al. [2016] propose a clustering technique to be used with an outlier detection method. Similarly to the work above, the authors analyzed statistical-based methods for outlier detection with low-cost humidity and temperature sensors. Their work is the most similar to the presented in this paper, but as with the other related projects, data filtering and smoothing is not handled.

All of the cited work conclude that the usage of outlier detection techniques improved the reliability of data from WSN and that it is crucial for these techniques to be lightweight to be viable for usage in sensor nodes.

3. Data Fusion

Despite being widely used, the term "Data Fusion" does not have a definitive meaning, varying according to the context in which it is used (data fusion, sensor fusion, information fusion) [Nakamura et al., 2007]. Boström et al. [2007] proposed the following definition: "Information fusion is the study of efficient methods for automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human or automated decision making." From this point on, this work will adopt this definition for data fusion.

Data fusion can be classified according to various aspects: relationship among the sources, abstraction level of the data, among others. On fusion based on the relationship among the sources, there is no direct dependency betwixt the sensors. The relationship can be classified as complementary (sensors provide different information of the same scenario), redundant (sensors provide the same information of the same scenario) and cooperative (information from different sensors can be fused to generate a new data).

Data fusion based on the abstraction level of the data can be divided into four layers [Iyengar et al., 2001; Nakamura et al., 2007]: low-level fusion (measurement) deals with raw data that can be combined to generated more accurate information; medium-level fusion (feature) represents the features/attributes of an entity that may be fused to create other features; high-level fusion (decision) combines decisions or features to obtain a more confident decision; and multilevel fusion represents the possibility of merging data from any level, such as fusing a measurement with a feature to generate a decision.

This work will focus in data fusion at the low-level layer, by fusing the raw data provided by redundant soil moisture sensors and improving its reliability by removing possible outliers and removing inherent noise of the measurements.

4. Outlier and Filtering Methods

Outliers are values that deviate from other readings in a sample, and they can be caused by the variability in the readings or errors in the experiment. If the outlier is identified as an error it should be removed, otherwise, it can provide valuable information [SEMATECH, 2003]. So, correctly identifying outliers is vital when analyzing data from sensors.

Moreover, after the outliers have been identified and removed (if necessary), the data might still contain noise from the environment, sensors, or by other electronic components, and the usage of a filter will help in generating a clearer and more useful data.

Various techniques were developed to detect outliers by using robust statistics, but there is no way to blindly use them on any kind of data. We have selected some methods to identify which would be more suitable for the needs of our application domain:

- **Chauvenet's Method:** This method works by taking the difference of the suspect data from the mean and divide it by the standard deviation. A normal distribution function is then used to determine the probability of the result of being outside 2 standard deviations [Taylor, 1997];
- **Peirce's Criterion:** considered more rigorous than Chauvenet's [Ross, 2003], it is an statistical method derived from a Gaussian distribution, and it can be used to detect two or more outliers.
- **Z-Score and Modified Z-Score:** Also called standard score, it is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured. Z-score has a limit in identifying possible outliers, expressed by $(n-1)/\sqrt{n}$, where n is the number of samples, and the result is the maximum Z-score possible [SEMATECH, 2003]. It is recommended to use the Modified Z-Score because it uses median absolute deviation (MAD) and median instead of mean and standard deviation;
- **Generalized Extreme Studentized Deviate (g-ESD):** ESD is a method similar to Grubbs test, but it does not requite an exact number of outliers to be detected, only an upper bound. It is recommended to specify a higher number of outliers than expected [SEMATECH, 2003].;
- **Adjusted boxplot:** Hubert & Vandervieren [2008] introduced a boxplot robust for skewed distribution, since regular boxplot has the limitation that the more skewed the data, too many observations might be detected as outliers.

Signal smoothing is the practice to reduce signal noise by changing values so that individual points that are higher than their neighbors are reduced, and that low values are increased [O'Haver, 1997]. Numerous filters have been developed with different techniques and goals, and we have explored some of them: Kalman filter, Weighted outlier-robust Kalman (WRKF), Savitzky-Golay, Robust locally weighted scatterplot smoothing (RLOESS and RLOWESS) and Scale-space.

5. Case Study

The goal of the experimental crop of cashew is to assess techniques to increase soil water retention and allow plants to grow more and to survive during drought seasons. To help with the experiment, we monitored the crop using soil moisture sensors in a uniform soil and evaluated various methods for data fusion, outlier detection, and signal filtering.

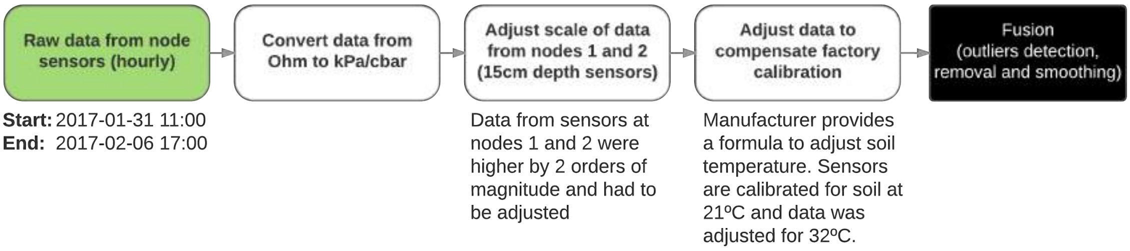
The experimental crop is being carried out by Embrapa in a field at Paraipaba - CE, and it presents particularities and problems that can be mitigated through this work. Currently soil moisture information is collected manually by visiting analog meters (tensiometers) once a day, and once a month the data is relayed to the research office.

5.1. Data Collection

We have built four sensor nodes inspired on the vinduino project [Van Der Lee, 2017], using Arduino Pro Mini and Irrometer Watermark 200SS soil moisture sensors. Each node includes three Watermark sensors at different depths (15cm, 45cm, 75cm). This is part of an ongoing project of an IoT data fusion architecture for Precision Agriculture, to improve sensors accuracy, detect events and anomalies and automated decision taking.

5.2. Data Analysis

Sensor nodes collected data for one week, and they were processed on a computer using Matlab R2016b. The main goal was to find a data fusion method adequate for precision agriculture sensors, and that can be run in nodes without local storage, meaning that it could not be a method that relied on full data in a time series or resource-intensive due to the nodes limited hardware. We have followed the structure presented in Figure 1: first we collected data from the nodes and applied the necessary adjustments (e.g.: converting from Ohms to KPa). Afterwards, we applied the selected outlier detection methods to the adjusted data. Finally, the selected filters were used to smooth out any noise. For this analysis, we use a fixed temperature value to adjust sensor calibration but a new experiment is underway with a soil temperature sensor to do a more precise calibration.


Raw data from node sensors (hourly)
Convert data from Ohm to kPa/cbar
Adjust scale of data from nodes 1 and 2 (15cm depth sensors)
Adjust data to compensate factory calibration
Fusion (outliers detection, removal and smoothing)

Start: 2017-01-31 11:00
End: 2017-02-06 17:00

Data from sensors at nodes 1 and 2 were higher by 2 orders of magnitude and had to be adjusted

Manufacturer provides a formula to adjust soil temperature. Sensors are calibrated for soil at 21ºC and data was adjusted for 32ºC.

Figure 1. Data analysis flowchart

6. Results

We had some interesting scenarios to be assessed (Figure 2):

- **15cm depth:** node 2 started acting increasingly erratic;
- **45cm depth:** node 3 have clear outliers, with values 100x higher than the average (Y axis in Figure 2 had to be in logarithmic scale to make the data readable);
- **75cm depth:** node 4 had almost flat values when there should have been variation.

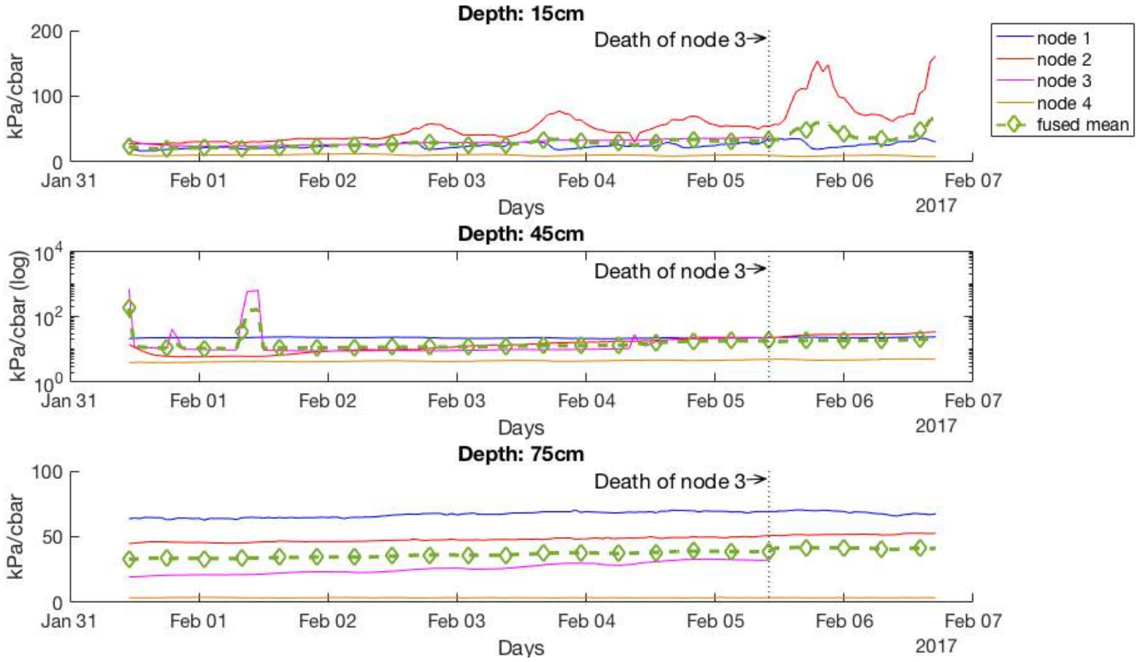


Figure 2. Raw data provided by sensor nodes

With that in mind, we had the following expectations: (i) the outlier detection method had to correctly identify the outliers from the sensors at 45cm on node 3; (ii) it also had to perform better than a regular mean of raw data; and it also had to avoid marking and removing false outliers.

6.1. Outlier Detection

6.1.1. Sensors at 15 cm depth

In Figure 3 we present outliers detected by Peirce's Criterion and modified Z-Score. The other methods are not presented because they did not detect any outlier (full listing at Table 1 at Section 6.1.4). Peirce's Criterion detected some peaks from node 2 but failed to detect the highest values, probably because node 3 died halfway through 2017/02/05 and it had only 3 values to compare. Modified Z-Score failed to detect the first (smaller) peaks but continued to work after the death of node 3 and successfully detected outliers.

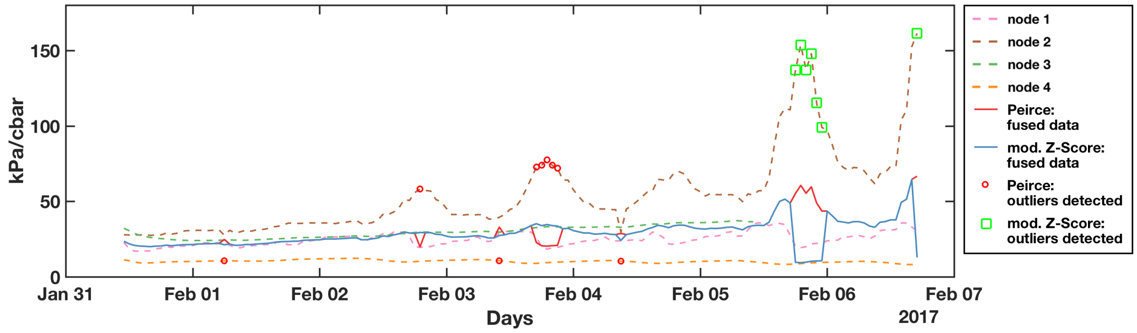


Figure 3. Outliers and fused data (15cm depth): Peirce's Criterion and Modified Z-Score

6.1.2. Sensors at 45 cm depth

We had more interesting results due to the clear presence of outliers from node 3. Figures 4 and 5 show outliers detected by the methods, and once again some were unable to find any outlier. Also, note that due to the extreme values of the outliers on Jan 31st and Feb 01st the Y-axis had to represented in a logarithmic scale to make the data readable.

- Adjusted boxplot: failed to detect outliers on Jan 31st and Feb 01st, probably due to the lack of more data and the spacing from the data points. On Feb 05 it started detecting outliers from node 4 when data from the other sensors started equalizing.
- Generalized ESD: it successfully detected outliers on Jan 31st and Feb 01st and also detected the same outliers of the adjusted boxplot.
- Peirce's criterion and modified Z-score: these methods were the most sensitive, detecting outliers in five of the seven days monitored. The issue is that they ended up removing valid data and would require manual fine tuning to each data set.

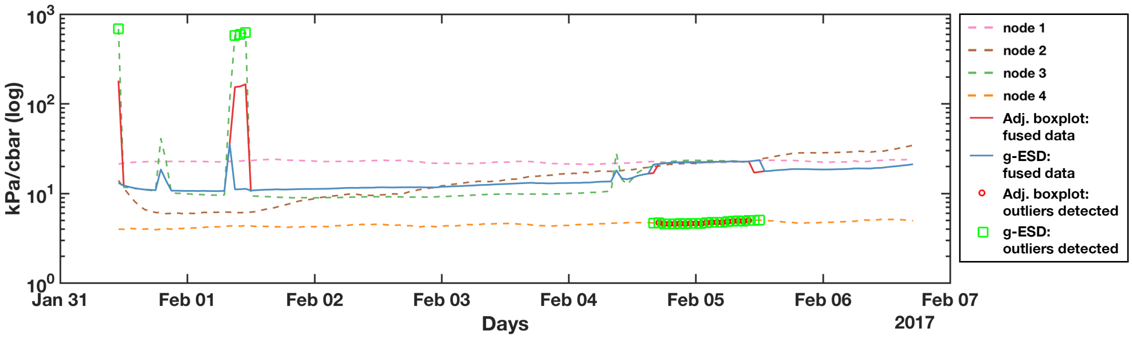


Figure 4. Outliers and fused data (45cm depth): Adjusted Boxplot and g-ESD

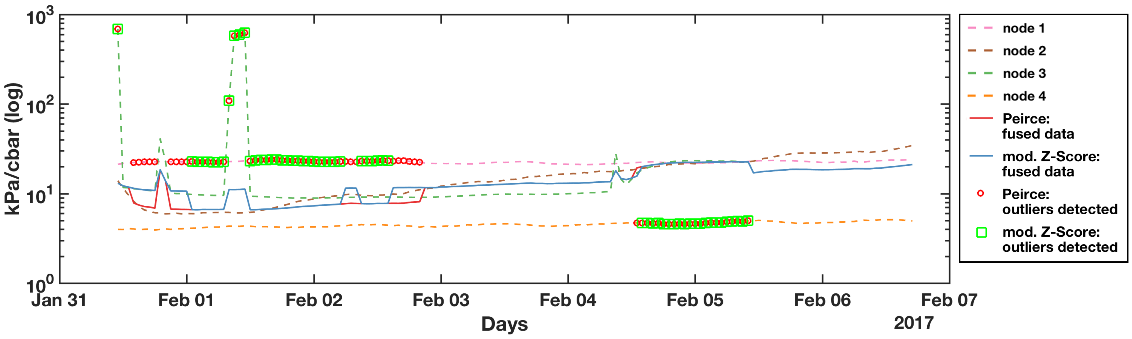


Figure 5. Outliers and fused data (45cm depth): Peirce's Criterion and Modified Z-Score

6.1.3. Sensors at 75 cm depth

All of the tested methods functioned as expected and did not detect any outliers since the values reported by the sensors did not vary.

6.1.4. Analysis

We have decided to adopt the generalized ESD method in our Precision Agriculture project because it managed to detect obvious outliers and not remove uncertain data that might still contribute valuable information. Table 1 resumes the performance of each method. Now it falls to the filters to smooth out noise and abrupt changes in the data.

Table 1. Total outliers detected and removed

	15cm	45cm	75cm	Comments
Adjusted boxplot	0	18	0	did not remove obvious outliers
Chauvenet	0	0	0	no outlier detected
Generalized ESD	0	25	0	select method
Peirce's criterion	9	76	0	too sensitive
Z-score	0	0	0	no outlier detected
Modified z-score	7	63	0	too sensitive

6.2. Filtering

In this section, we took the data fused by the ESD method and fed it to the selected filters mentioned at Section 4. All filters managed to reduce the existing noise in the output of the ESD method, but results varied regarding the anomalous readings (spikes).

6.2.1. Sensors at 15 cm depth

In Figure 6 it is noticeable that RLOESS and RLOWESS failed to follow the original data curve after February 5th and their output stopped being reliable. Kalman filter created a more reliable and smoother output, but it was still heavily influenced by the surge at the end of Feb 5th. Moreover, the WRKF filter had the best output of them all. Figure 7 presents the output of Savitzky-Golay and scale-space, where both managed to remove noise but followed closely the format of the original data from ESD.

6.2.2. Sensors at 45 cm depth

Although ESD removed obvious outliers, some anomalous data remained and they need to be treated by the filters. Figures 8 and 9 shows that Kalman, Savitzky-Golay, and

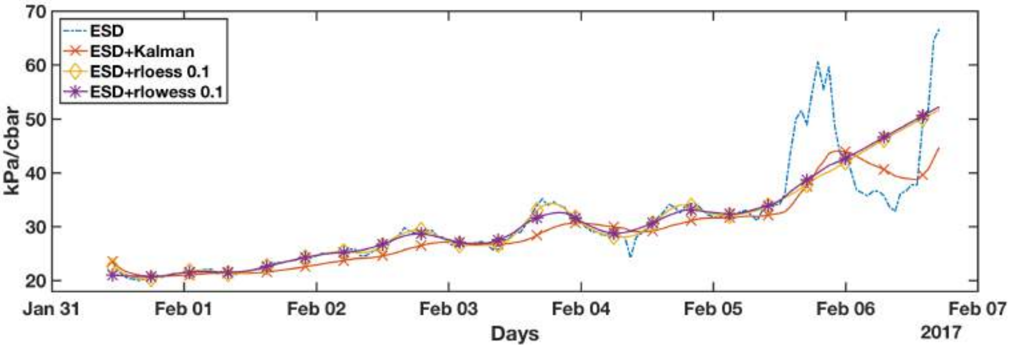


Figure 6. Filtered data for sensors at 15cm depth (a)

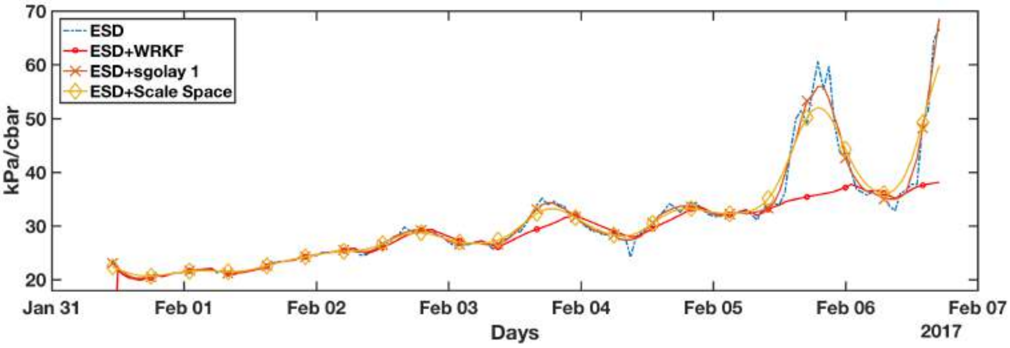


Figure 7. Filtered data for sensors at 15cm depth (b)

Scale-space were affected by spikes on Jan 31st and Feb 01st, and Kalman did better with the anomalous reading on Feb 5th. RLOESS and RLOWESS were not affected by the first remaining outliers, but only RLOWESS was resilient to spikes on Feb 5th. Once again, the best filter was WRKF, although on Feb 1st and 2nd the resulting signal is not as smooth as the other filters.

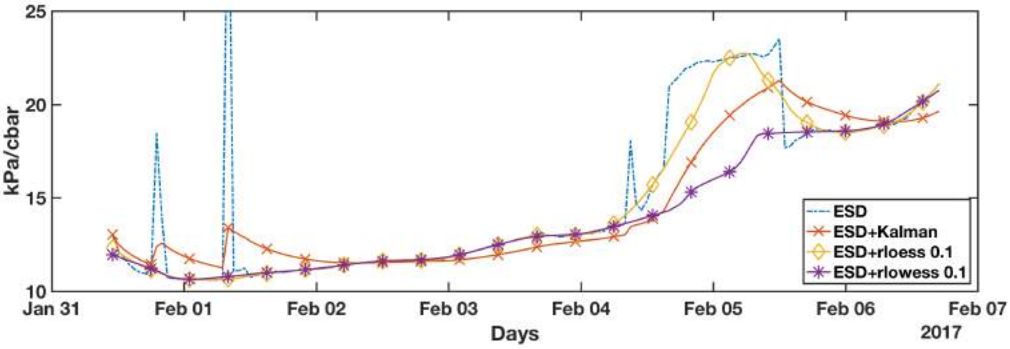


Figure 8. Filtered data for sensors at 45cm depth (a)

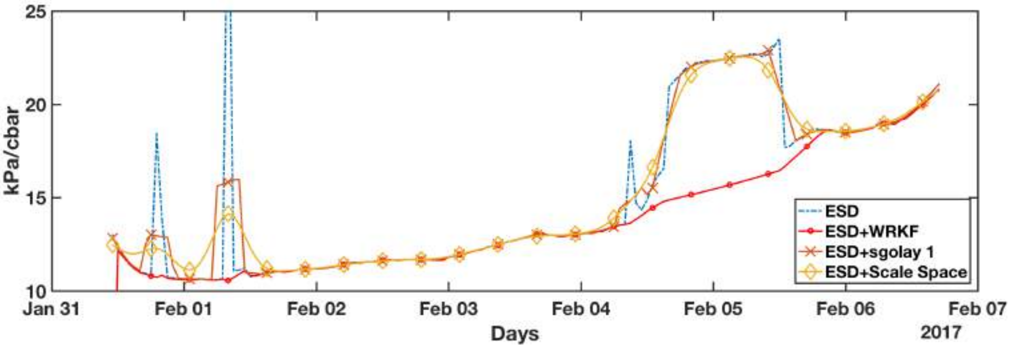


Figure 9. Filtered data for sensors at 45cm depth (b)

6.2.3. Sensors at 75 cm depth

The performance of almost every filter was similar with data from sensors at 75cm depth (Figures 10 and 11), except of Kalman, that presented a negative bias on the output, and WRKF, that once again resulted in a somewhat noisy output.

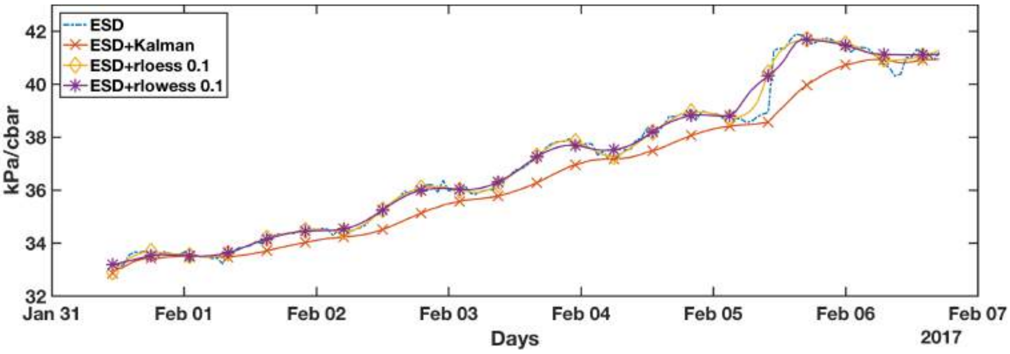


Figure 10. Filtered data for sensors at 75cm depth (a)

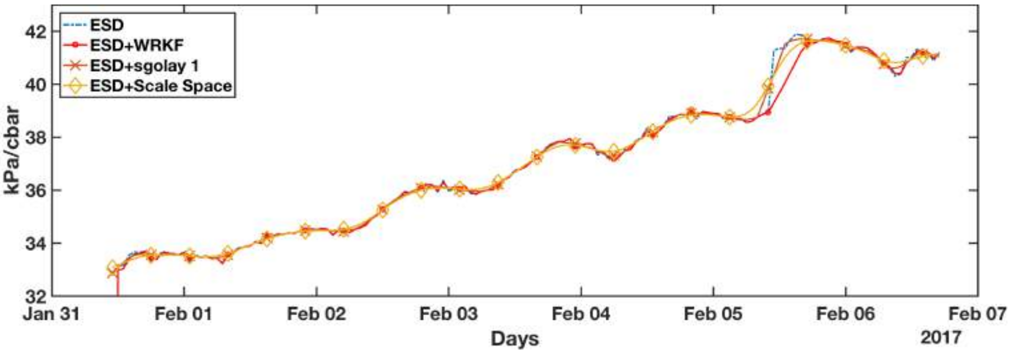


Figure 11. Filtered data for sensors at 75cm depth (b)

6.2.4. Analysis

Despite presenting some noise, most notably on data from sensors at 75cm depth, WRKF had the best performance in being resilient to residual outliers and anomalous data (possibly introduced by malfunctioning sensors).

7. Conclusion

We have monitored an Embrapa's experimental crop of precocious dwarf cashew using soil moisture sensors. Next, we performed information fusion on the gathered data provided by applying several methods of outliers detection and removal. Afterward, we tested filters to smooth out existing noise and any residual outlier or anomalous data. By analyzing each method, we have come to the conclusion that the best solution is combining generalized ESD (to fuse the data and remove outliers) and weighted outlier-robust Kalman filter (WRKF). ESD managed to fuse data from four sensors and removed outliers there were certain to be caused by errors while maintaining relevant data, and WRKF managed to filter out residual outliers and smooth peaks in the data. Also, by the test results, we can determine that fours equal sensors is the lowest number necessary to identify outliers and get more reliable data, since at least one sensor will malfunction.

For future works, we are currently working on having raw data converted into feature information and getting everything to be executed locally in the sensors nodes. Besides, we intend to increase the number of sensors in the crop.

References

Andrade, A.; Montez, C.; Moraes, R.; Pinto, A. R.; Vasques, F.; & Siva, G. L. "Outlier Detection Using k-means Clustering and Lightweight Methods for Wireless Sensor Networks." In *42nd Annual Conference of IEEE Industrial Electronics Society (IECON)*, S. 1–6 [2016]. doi:10.1109/IECON.2016.7794093.

Bernardi, A. C. d. C.; Naime, J. d. M.; de Resende, A. V.; Bassoi, L. H.; & Inamasu, R. Y. *Agricultura de precisão: resultados de um novo olhar.* Embrapa, Brasília, DF [2014]. ISBN 9788570353528.

Boström, H.; Andler, S. F.; Brohede, M.; Johansson, R.; Karlsson, A.; Laere, J. V.; Niklasson, L.; Nilsson, M.; Persson, A.; & Ziemke, T. "On the Definition of Information Fusion as a Field of Research." In *IKI Technical Reports*, (October):S. 1–8 [2007]. doi:HS-IKI-TR-07-006.

Callegaro, R.; Montez, C.; Pinto, A. R.; & Moraes, R. "Uma Arquitetura para Fusão de Dados e Detecção de Outliers em Sensores de Baixo Custo de Redes de Sensores sem Fio." In *Anais do II Workshop de Comunicação em Sistemas Embarcados Críticos - WoCCES*, S. 3–16 [2014]. doi:10.5753/cbie.wcbie.2015.1007.

Hubert, M. & Vandervieren, E. "An adjusted boxplot for skewed distributions." In *Computational statistics & data analysis*, Band 52 (12):S. 5186–5201 [2008].

Iyengar, S. S.; Chakrabaraty, K.; & Qi, H. "Introduction to special issue on 'distributed sensor networks for real-tie systems with adaptive configuration'." In *Journal of the Franklin Institute*, Band 338:S. 651–653 [2001].

Nakamura, E. F.; Loureiro, A. a. F.; & Frery, A. C. "Information fusion for wireless sensor networks." In *ACM Computing Surveys*, Band 39 (3) [2007]. ISSN 03600300. doi:10.1145/1267070.1267073.

Natural Resources Conservation Service. "Precision Agriculture: NRCS Support for Emerging Technologies." In *Agronomy Technical Note* [2007].

O'Haver, T. "A pragmatic introduction to signal processing." [1997].

Ravichandran, J. & Arulappan, a. I. "Data validation algorithm for wireless sensor networks." In *International Journal of Distributed Sensor Networks*, Band 2013 (iv) [2013]. doi:10.1155/2013/634278.

Ross, S. M. "Peirce's criterion for the elimination of suspect experimental data." In *Journal of Engineering Technology*, Band 20 (2):S. 38–41 [2003].

Sanchez, L.; Muñoz, L.; Galache, J. A.; Sotres, P.; Santana, J. R.; Gutierrez, V.; Ramdhany, R.; Gluhak, A.; Krco, S.; Theodoridis, E.; & Pfisterer, D. "SmartSantander: IoT experimentation over a smart city testbed." In *Computer Networks*, Band 61:S. 217–238 [2013]. ISSN 13891286. doi:10.1016/j.bjp.2013.12.020.

SEMATECH, N. "e-Handbook of Statistical Methods." [2003]. URL `http://www.itl.nist.gov/div898/handbook/`. Accessed: 2017-03-16.

Taylor, J. R. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements.* University Science Books, 2nd Auflage [1997]. ISBN 0935702423,9780935702422.

Van Der Lee, R. "Vinduino: Open license project for agricultural irrigation management." [2017]. URL `https://github.com/ReiniervdL/Vinduino`. Accessed: 2017-03-17.