IoT-Enabled Distributed Data Processing for Precision Agriculture

Grigore Stamatescu, Cristian Drăgana, Iulia Stamatescu, Loretta Ichim and Dan Popescu

Abstract— **Large scale monitoring systems, enabled by the emergence of networked embedded sensing devices, offer the opportunity of fine grained online spatio-temporal collection, communication and analysis of physical parameters. Various applications have been proposed and validated so far for environmental monitoring, security and industrial control systems. One particular application domain has been shown suitable for the requirements of precision agriculture where such systems can improve yields, increase efficiency and reduce input usage. We present a data analysis and processing approach for distributed monitoring of crops and soil where hierarchical aggregation and modelling primitives contribute to the robustness of the network by alleviating communication bottlenecks and reducing the energy required for redundant data transmissions. The focus is on leveraging the fog computing paradigm to exploit local node computing resources and generate events towards upper decision systems. Key metrics are reported which highlight the improvements achieved. A case study is carried out on real field data for crop and soil monitoring with outlook on operational and implementation constraints.**

I. INTRODUCTION

Internet of Things (IoT) systems are based on distributed sensing, computing and communication devices that collaborate in order to monitor and control physical processes. These enable the collection of real world data at an unprecedented scale and resolution which can then be used to improve the models that define the understanding and help the forecasting of the processes, be it technical, social or environmental. New data processing infrastructure are thus needed to store and retrieve the information collected in an online manner while providing mechanisms to run the analysis and control algorithms based on this data. Beyond conventional environmental monitoring as initial key driver of IoT design, current domains include (smart) cities, industry and agriculture. Finally the outcomes of the analysis are either handled in closed loops for control actions or they are supplied to hierarchical entities for decision support.

Among the applications areas mentioned above, precision agriculture represents one of the salient areas where IoT-enabled systems can improve the quality, productivity and increase automation [1]. Main challenges in this field relate to reducing input use: water, fertiliser, work, and obtaining better crop yields which is demanded by the market to keep food costs low under the strains of increasing global population. By having access to reliable, on-line information, relayed over distributed networks, domain specialists can oversee tangible improvements [2].

The conceptual and practical challenges that we approach in the design of such systems is related to efficient data reduction and management which impacts directly the congestion and energy metrics of the deployed network. This is performed by proposing a hierarchical data processing architecture in accordance to fog computing design principles. Fog computing as a concept has initially emerged as a computing organisation alternative to leverage intelligent network edge devices which make up modern IoT systems [3]. The limited computing resources available on these embedded devices are thus exploited to reduce the large quantities of collected data and transmit only higher level information pieces upstream. Given the large heterogeneity the processing primitive can run of the edge nodes range from basic threshold detection and averaging up to more advanced outlier detection and embedded learning algorithms. Wireless sensor networks (WSN) are an enabling technology to deploy fog computing systems [4], [5] where hundreds to thousands of sensing nodes self organise intro and communicate over low power radio channels. As with the case with agriculture, large areas can thus be covered with multi-hop communication networks as the networking protocols rely on cluster heads, gateways and hubs serving as intermediary data concentrators. One alternative definition presents fog systems in opposition or as complementary to conventional centralised and large scale cloud infrastructures. The complex functionality of the cloud platform is broken down at the field level over functional or spatially distributed entities which collaborate to achieve a common monitoring, event-detection and control case. In the precision agriculture use case this can help implement an optimised distributed irrigation or fertiliser dosage schemes accounting for local properties and variance of soil, micro-climate and crop particularities. The need to integrate fog computing with cloud computing in this particular scenario lays with the fact that joint observations can be derived when federating high-level information across multiple farms.

The main novelty of the paper is justified by the application of fog computing data aggregation and modelling primitives in the context of IoT-enabled smart agriculture, a highly active area of research currently. The subsequent contributions of the paper can be argued:

- system architecture for hierarchical data processing and analysis based on field level IoT devices;
- data aggregation methodology based on the fog computing paradigm under precision agriculture constraints.

*This work has been funded by the Romanian Space Agency (ROSA), through the project ”Integrated Multi-Agent Aerial Robotic System for Exploring Terrestrial Regions of Interest” (MAARS), contract no. 185/2017.

The authors are with the Department of Automatic Control and Industrial Informatics, University Politehnica of Bucharest, 060042 Bucharest, Romania. Grigore Stamatescu is also with the Institute of Technical Informatics, Technical University of Graz, 8010 Graz, Austria gstamatescu@tugraz.at

II. RELATED WORK

In [6] a fog computing framework for precision agriculture is introduced. The two tiered system is able to reduce significantly the data transmitted in the network. Reducing the computational loads, and most important, the cloud computing costs associated with centralised processing is highlighted as an essential benefit of the fog approach. The authors of [7] propose a hybrid IoT for smart farming in rural areas. The communication network uses 6LoWPAN local radio for the field interfaces while long range connections are implemented over WiFi. A 6LoWPAN border router and dedicated gateway are used to assure cross-domain integration of the networks from field level, intermediate long range relays and cloud. Network requirements for smart agriculture applications are also discussed in terms of throughput, latency and mobility support. These offer a good reference to quantify the data aggregation potential in conjunction with the sensing and control requirements. A distributed computing architecture is presented in [8] which the agricultural system basic components such as: crop, soil, climate, water and nutrients, energy. The messaging system is standardised around the Message Queuing Telemetry Transport (MQTT) to interlink sensors, actuators, communication nodes, devices and subsystems [9]. A decision tree is designed for irrigation control and integrated on the edge devices for in situ decision making. At the top level cloud services supply data through an end-user dashboard for high level decision support.

[10] introduce an intelligent irrigation system based on distributed sensor using the LoRA long range, low rate, nodes and gateways. The FIWARE infrastructure is leveraged as data management middleware platform which provides the support services. Several operation scenarios are discussed based on the scalability requirements, in terms of tens of thousands of nodes. Reference computational resource assessment for cpu, memory and network is also reported. Large scale IoT monitoring is discussed in [11]. The focus is on the ground level clustering mechanisms that support the timely collection of data and generating of the field level monitoring events. Aerial robotic platform support is provided through suitable high level control of trajectories for data collection and backhaul. Data reduction is achieved by thresholding over locally computing moving averages in conjunction with expert knowledge adapted to the monitored processes. Several radio access technologies are available to achieve reliable transmissions [12].

III. SYSTEM ARCHITECTURE AND METHODOLOGY

A. SYSTEM ARCHITECTURE FOR DATA COLLECTION AND PROCESSING

The proposed system architecture that we have designed for the purpose of efficient data collection and processing in precision agriculture is illustrated in Figure 1. It consists of the following information and physical layers: field layer, fog computing layer, cloud computing layer, data presentation layer, which are linked by cross-layer upstream and downstream data and control information flows. The layer functionality is detailed next:

- Field layer: includes the actual sensors deployed in the precision agriculture application to measure the physical parameters of interest; these include air temperature, air humidity, solar radiation, soil temperature at various depths, windspeed and rainfall; the field layer can also be expanded to accommodate intelligent actuators e.g. for irrigation or fine grained nutrient dosage, to execute commands incoming from higher level systems;
- Fog Computing layer: the fog nodes collect data from the sensors and run the data processing primitives for intelligent aggregation in order to reduce network traffic and energy expenditure; the main idea is to locally derive basic model characteristics of the particular process which are sent to the cloud in compact form; correlations between the sensed variables can also be exploited at this level for local decisions thus avoiding completely the increased cost and latency of the upper layers;
- Cloud computing layer: data is streamed towards a common cloud platform; regarding the particular implementation we use the ThingSpeak [13] platform in conjunction with Matlab algorithm development for higher level processing routines; at the cloud layer the model parameters allow the reconstruction of the time series characteristics if needed, while accounting for the inherent modelling errors;
- Data presentation layer: is concerned with the front-end software systems that present the outcomes of the data analysis to end-users or decision makers with the ability to provide mobile access and timely alerts in the case of event detection; parametrisation of the process by domain experts is also achieved at this layer.

A more detailed algorithm flowchart is provided in Figure 2. It includes the steps for algorithm description which runs on the fog computing node.

In-field measurements are uploaded to the IoT application in two ways depending on the type of information: events and measurements. Note that, a primary batching procedure is usually available for most of the monitoring systems, basically consisting of performing minimum, maximum and mean value during a specific period of time. We consider this as the starting point for further local data processing.

Primary batch aggregation Note that, a primary batching procedure is usually available for most of the monitoring systems, basically consisting of performing minimum, maximum and mean value during a specific period of time. We consider this as the starting point for further local data processing.

For instance, batches are defined within 30 minutes. Once a new batch is available, min, max and $mean$ values are computed (step A).

Check for outliers procedure For each batch of measurements, an outliers' check procedure is performed, considering an acceptance bandwidth of data variance for the measured value around the mean (step B). The procedure

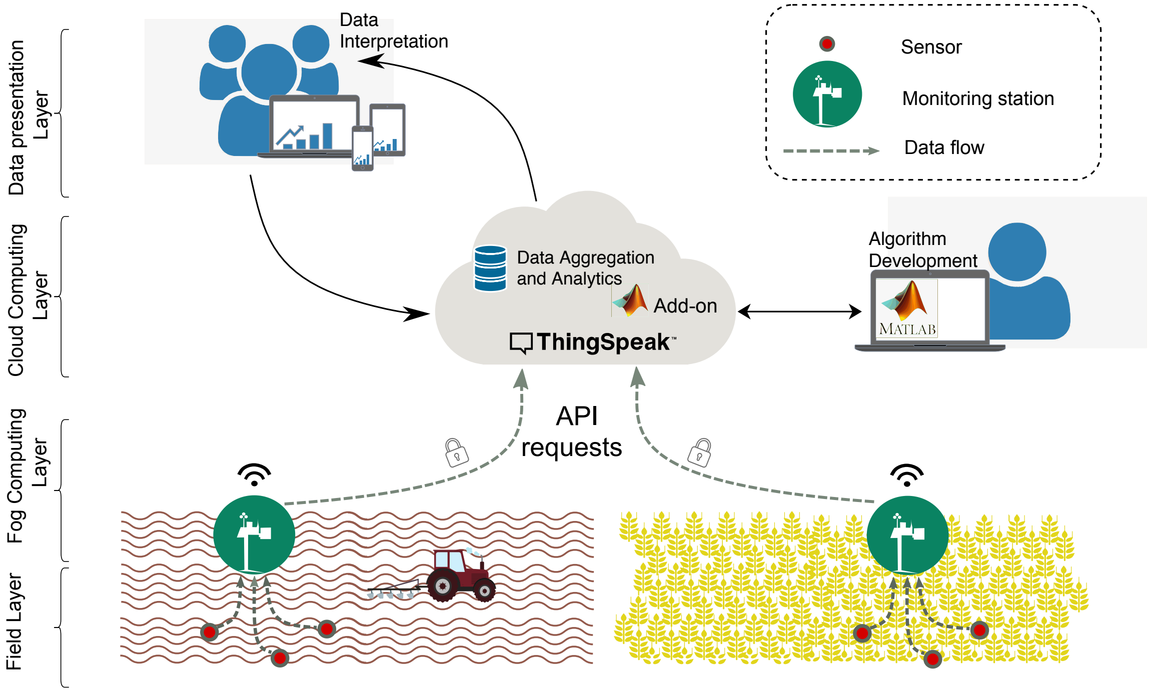


Fig. 1: Distributed data processing based on fog computing for precision agriculture

outputs an event if the minimum or maximum values exceeds the thresholds. The event E is defined as:

$$E = \{e(x_i) \in Q, T_{min} < x_i < T_{max}\} \tag{1}$$

where:

- x_i is the measured value at iteration i
- T_{min} and T_{max} are thresholds computed as:

$$T_{min} = mean(1 - w) \tag{2}$$

$$T_{max} = mean(1 + w) \tag{3}$$

where w is a weight for acceptance bandwidth size define.

Relevant data extraction Aggregated data sets are achieved based on different methods. All seek for relevant data point, aiming to a reduced size set providing at the same time a satisfying reconstruction of the initial data.

One effective method, in terms of data volume, is based on using the min and max values extraction, computed for 24 hours. It is obvious that this method is suitable only for measurements that follow a regular shape during time, with insignificant variations during a day. A measurement for which this method is suitable is the soil temperature.

Instead, change detection is a common method applicable for irregular shaped data sets. This method follows extraction of data points where trend changes occur.

Given a set of data point $(x_i, y_i), i = 1, ..., n$, trend t_i is followed for each pair x_i, x_{i+1}, such that for

$$\begin{aligned} x_{i+1} - x_i > \delta &\implies t(i) = 1 \\ x_{i+1} - x_i < \delta &\implies t(i) = -1 \\ x_{i+1} = x_i &\implies t(i) = 0 \end{aligned} \tag{4}$$

Then, if $t(i) \neq t(i+1)$ means that a trend change is detected. The coresponding data point $x_(i+1)$ is added to the relevant data set.

Relevant data extraction (step C) is performed when a set of primary aggregated batches is available.

B. DATA AGGREGATION

One reference method of extracting high level information from sensor data is Symbolic Aggregate Approximation (SAX) [14]. It operates by assigning label symbols to segments of the time series thus porting it in a unified lower dimension representation. It belongs to the family of time series data mining techniques leading to non-parametric modelling. Ranges are identified through the data histogram or in a uniform manner. The method provides linear complexity and opens up the use and application of multiple statistical learning tools. Parametrisation of SAX is highly important by defining the number of segments and the alphabet size which can influence the quality and robustness of the result.

The background on which SAX has been defined is established by PAA [15] where symbols are attributed to the aggregated numerical values listed by PAA. Several discrete event models can incorporate the resulting aggregated segments e.g. Markov models in order to compute the probability of the observed patterns for future observations. According to the PAA method description, starting with a time series X of length n, this is approximated into a vector $\bar{X} = (\bar{x}_1, ..., \bar{x}_M)$ of any length $M \leq n$, with n divisible by M. Each element of the vector $\bar{x}_i$ is calculated by:

$$\bar{x}_i = \frac{M}{n} \sum_{j=n/M(i-1)+1}^{(n/M)i} x_j \tag{5}$$

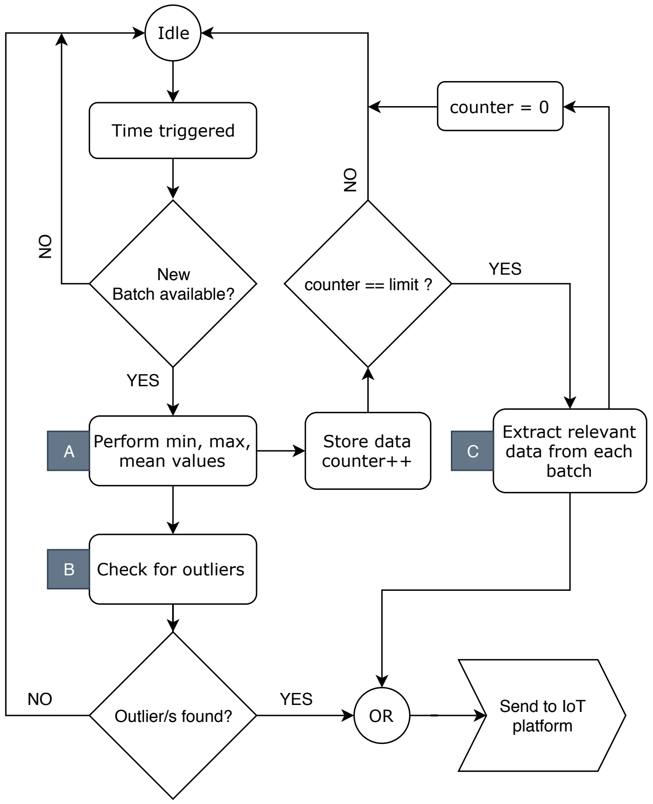


Fig. 2: Fog Computing algorithm

The dimensionality of the time series is thus reduced from n to M samples by initially dividing the original data into M equally sized frame and then compute the mean values for each frame. A new sequence is achieved by putting the mean values together which is considered to be the PAA transform (approximation) of the original data. With regard to computational considerations, the PAA transform complexity can be reduced from $O(NM)$ to $O(Mm)$ with m being the number of frames as tuning parameter of the method. The distance measure between two time series vector approximations $\bar{X}$ and $\bar{Y}$ is defined as:

$$D_{PAA}(\bar{X}, \bar{Y}) = \sqrt{\frac{n}{M}}\sqrt{\sum_{i=1}^{M}(\bar{x}_i - \bar{y}_i)} \tag{6}$$

It has been shown by the proposers of the method that PAA satisfies the lower bounding condition and guarantees no false dismissals such that:

$$D_{PAA}(\bar{X}, \bar{Y}) \leq D(X, Y) \tag{7}$$

C. INTERPOLANT METHODS

The Cloud-based application rebuilds data sets by estimates based on interpolation mechanisms. For performance evaluation we showcase three methods: the common linear interpolant (also referred as piecewise linear interpolant) and two closely related interpolants, cubic *spline* and shape preserving *Piecewise Cubic Hermite Interpolating Polynomial (pchip).*

Given a set of data points $(x_i, y_i), (x_{i+1}, y_{i+1}), ...,$ (x_n, y_n), the linear interpolation is defined as the concatenation of linear interpolants between each pair of data points, thus a set of straight lines between each data points. Any pair of data points with $x_i \neq x_{i+1}$ determines a unique polynomial p of degree less than two whose graph passes through the two points with the property:

$$p(x_i) = y_i \tag{8}$$

with the form:

$$p(x) = a_1 x + a_0 \tag{9}$$

a 1-D linear interpolation.

In general, given n points $(x_i, y_i), i = 1, ..., n$, with disting x_i, a polynomial of degree less than n whose graph passes through the n points denoted $P_n(x)$, is expressed in the *Lagrange* form as:

$$P_n(x) = \sum_{i=1}^{n}\left(\prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}\right) y_i \tag{10}$$

The Lagrange form in (10) can be written out in power form of an interpolating polynomial as,

$$P_n(x) = a_1 x^{n-1} + a_2 x^{n-2} + ... + a_{n-1} x + a_n \tag{11}$$

where the coefficients a_k are computed through a system of linear equations:

$$\begin{bmatrix} x_1^{n-1} & x_1^{n-2} & ... & x_1 & 1 \\ x_2^{n-1} & x_2^{n-2} & ... & x_2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^{n-1} & x_n^{n-2} & ... & x_n & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{12}$$

Considering this, a piecewise linear interpolant is produced by first computing the divided difference:

$$\delta_i := \frac{y_{i+1} - yi}{x_{i+1} - xi} \tag{13}$$

Then the interpolant is constructed as:

$$P(x) = y_i + \delta_i (x - x_i) \tag{14}$$

Further, for piecewise cubic polynomials, considering an interval $x_i \leq x \leq x_{i+1}$ let $h_i := x_{i+1} - xi$ be the length of an i^{th} interval and $d_k := P'(x_i)$. Therefore, using this derivative it is possible to adjust the interpolant in order to enforce smoothness, by forcing the pair of derivatives from consecutive piecewice cubics to agree.

All piecewise cubic hermite interpolating polynomials are continuous and have a continuous first derivative. In particular, *spline* is oddly smooth, meaning that it's second derative also varies continously.

Instead, *pchip* is not as smooth as *spline*, it is actually designed so that it never overshoots the data. The slopes are chosen so that $P(x)$ preserves the shape of data and also respects monotonicity.

IV. EXPERIMENTAL RESULTS

We collect experimental data from a network of field devices installed on site at an experimental research farm. Form the long term monitoring dataset we select a sample for analysis that covers one month of data. The data is preprocessed for missing values, noise removal and averaged over 30 minute intervals.

We first illustrate the application of the SAX method on the measured values for soil temperature and solar radiation in Figure 3 and Figure 5. Segment levels codify the evolution of the respective time series and provide a compact representation with considerable impact on the data storage and transmission requirements at the fog node. Finer grained patterns can be observed by zooming in at the daily level as is illustrated in Figure 5. Based on the selected segment labels, if the expected value deviates significantly by entering a different label segment, an event detection primitive can trigger a communication message from the node upstream.

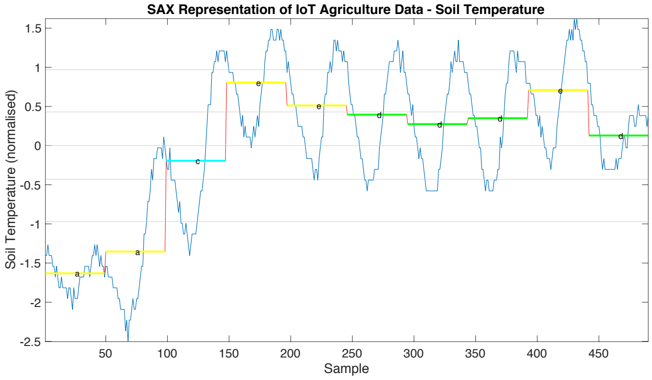


Fig. 3: Symbolic Aggregation Approximation - Soil Temperature

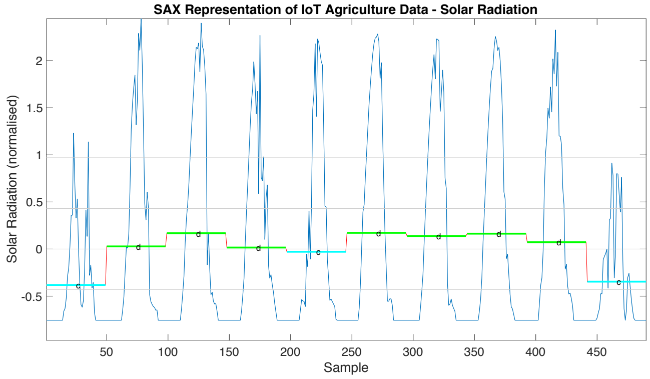


Fig. 4: Symbolic Aggregation Approximation - Solar Radiation

In order to evaluate reconstructed data consistency, achieved through different estimating algorithms, more precisely the proposed interpolants, some well known goodness-on-fit statistics are performed:

- Sum of squares of errors (SSE) - measures the total deviation of the response values from the fit to the response values and is defined as:

$$SSE = \sum_{i=1}^{n} w_i (x_i - P(x_i))^2 \tag{15}$$

where w_i is the weight for the i^{th} error between estimated i^{th} value and the empiric data

- R-square - measures how successful the fit is in explaining the variation of the data and is expressed as:

$$R - square = 1 - \frac{SSE}{SST} \tag{16}$$

where

$$SST = \sum_{i=1}^{n} w_i (x_i - \bar{x_i})^2 \tag{17}$$

where $\bar{x_i}$ is the mean value of x_i dataset.

- Root mean square error (RMSE) - is an estimate of the standard deviation of the random component in the data and is expressed as:

$$RMSE = \sqrt{\frac{SSE}{n}} \tag{18}$$

Results are summarised in Table I.

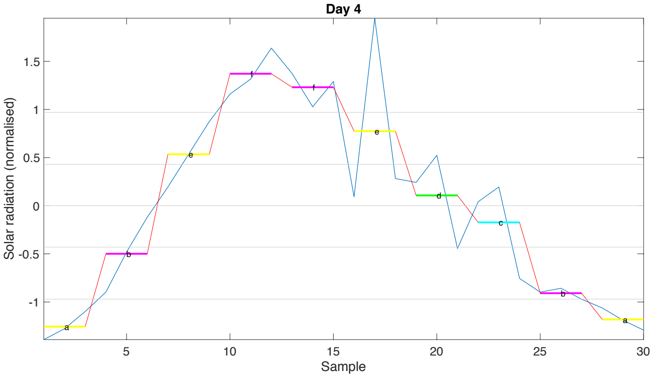
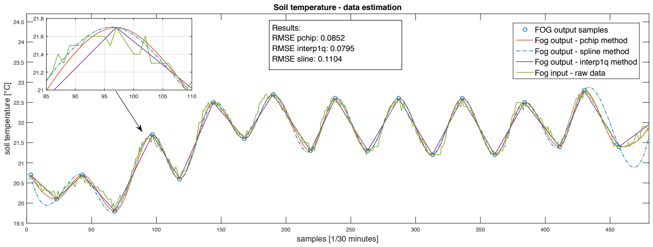


Fig. 5: Solar Radiation - Day level aggregation

TABLE I: Goodness-on-fit statistics results

	pchip RMSE	sline RMSE	interp1q RMSE
soil temperature	0.0852	0.1104	0.0795
solar radiation	0.2627	0.3691	0.2551



Results:
RMSE pchip: 0.0852
RMSE interp1q: 0.0795
RMSE sline: 0.1104

Fig. 6: Soil Temperature

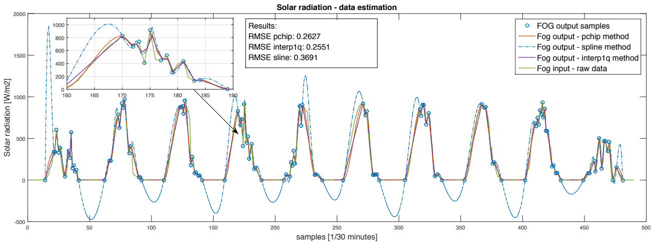


Fig. 7: Solar Radiation

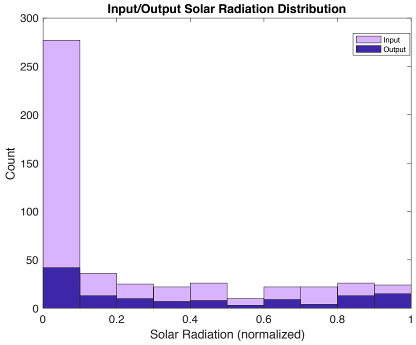


Fig. 8: Solar Radiation - Histogram

Figure 6 and Figure 7 graphically depict the results of applying the alternative methods of interpolation on the two time series. In Figures 8 and 9 the histograms quantify the associated data reduction between the raw input data and the interpolant methods presented.

For this case, the monotonicity property of *pchip* is more desirable than the smoothness property of *spline*, which in some places overshoots the data, thus one may prefer the good behavior of the shape preserving *pchip* method. Note that, as with the linear interpolation, when there are two consecutive points with the same value, the interpolant is constant over that interval. This behaviour was expected and it is appropriate in this context.

Even if the metrics indicate better fitting for linear interpolation through the studied cases, one can choose the *pchip* method, given that the results are quite close and it does a much more visual pleasing representation, in particular better modelling the peeks and following the expected behaviour around the baseline.

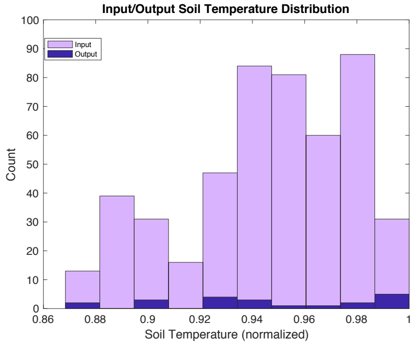


Fig. 9: Soil Temperature - Histogram

V. CONCLUSIONS

The paper presented a system architecture and distributed data processing application based on IoT in precision agriculture. By exploiting the dense spatial and temporal distributions of the sensing nodes, intelligent data reduction through aggregation and model reconstruction is illustrated for significant benefits for network congestion and energy efficiency. As the results achieved show promise, future work is focused on extensive evaluation for online decision making by domain experts in order to improve the reconstructed data quality.

REFERENCES

[1] S. Heble, A. Kumar, K. V. V. D. Prasad, S. Samirana, P. Rajalakshmi, and U. B. Desai, "A low power iot network for smart agriculture," in *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, Feb 2018, pp. 609–614.

[2] O. Elijah, T. A. Rahman, I. Orikumhi, C. Y. Leow, and M. N. Hindia, "An overview of internet of things (iot) and data analytics in agriculture: Benefits and challenges," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3758–3773, Oct 2018.

[3] M. R. Anawar, S. Wang, M. Azam Zia, A. K. Jadoon, U. Akram, and S. Raza, "Fog computing: An overview of big iot data analytics," *Wireless Communications and Mobile Computing*, vol. 2018, 2018.

[4] V. Mihai, C. Dragana, G. Stamatescu, D. Popescu, and L. Ichim, "Wireless sensor network architecture based on fog computing," in *2018 5th International Conference on Control, Decision and Information Technologies (CoDIT)*, April 2018, pp. 743–747.

[5] V. Mihai, C. E. Hanganu, G. Stamatescu, and D. Popescu, "Wsn and fog computing integration for intelligent data processing," in *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, June 2018, pp. 1–4.

[6] E. Guardo, A. Di Stefano, A. La Corte, M. Sapienza, and M. Scatà, "A fog computing-based iot framework for precision agriculture," *Journal of Internet Technology*, vol. 19, no. 5, pp. 1401–1411, 2018.

[7] N. Ahmed, D. De, and I. Hussain, "Internet of things (iot) for smart precision agriculture and farming in rural areas," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4890–4899, Dec 2018.

[8] F. J. Ferrández-Pastor, J. M. García-Chamizo, M. Nieto-Hidalgo, and J. Mora-Martínez, "Precision agriculture design method using a distributed computing architecture on internet of things context," *Sensors*, vol. 18, no. 6, p. 1731, 2018.

[9] T. Yokotani and Y. Sasaki, "Transfer protocols of tiny data blocks in iot and their performance evaluation," in *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, Dec 2016, pp. 54–57.

[10] C. Kamienski, J.-P. Soininen, M. Taumberger, R. Dantas, A. Toscano, T. Salmon Cinotti, R. Filev Maia, and A. Torre Neto, "Smart water management platform: Iot-based precision irrigation for agriculture," *Sensors*, vol. 19, no. 2, p. 276, 2019.

[11] D. Popescu, C. Dragana, F. Stoican, L. Ichim, and G. Stamatescu, "A collaborative uav-wsn network for monitoring large areas," *Sensors*, vol. 18, no. 12, p. 4202, 2018.

[12] G. Stamatescu, D. Popescu, and R. Dobrescu, "Cognitive radio as solution for ground-aerial surveillance through wsn and uav infrastructure," in *Proceedings of the 2014 6th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, Oct 2014, pp. 51–56.

[13] M. Maureira, D. Oldenhof, and L. Teernstra, "Thingspeak—an api and web service for the internet of things." 2011, available online.

[14] E. Keogh, J. Lin, and A. Fu, "Hot sax: Efficiently finding the most unusual time series subsequence," in *Proceedings of the Fifth IEEE International Conference on Data Mining*, ser. ICDM '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 226–233. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2005.79

[15] K. Chakrabarti, E. Keogh, S. Mehrotra, and M. Pazzani, "Locally adaptive dimensionality reduction for indexing large time series databases," *ACM Trans. Database Syst.*, vol. 27, no. 2, pp. 188–228, Jun. 2002. [Online]. Available: http://doi.acm.org/10.1145/568518.568520