DETERMINING AN EFFECTIVE SET OF TIME SERIES DATA PARAMETERS SUITABLE FOR IDENTIFYING SECURITY INCIDENTS

Vasyl Komarov, Dmitrii Chernov, Vladimir Krizhanovski

Vasyl' Stus Donetsk National University

Abstract

The purpose of the research is developing a method of data analysis, systematization, and preparation of time series of sensors values for further analysis by the method of structural functions (Allan variance), which can be used to reveal characteristics of processes and their changes with high validity. The data collection was performed in a real cyber-physical system on large range time intervals using examples of 11 different temperature sensors. The analysis of the processed digital data was performed and influences factors on the way from the measurement point to the point of receiving information were determined. It is shown that the appropriate preparation of data arrays makes it possible to obtain distinctive generalized samples of time sequences parameters, the comparison to which allows to detect security incidents and abnormal processes in IoT sensor networks. The considering of these factors makes it possible to create a tool for sensors values pre-sorting and sample data sets obtaining. The developed hardware and software tools for such preliminary processing are described.

Keywords: time series data, data preparation, Identification of differences in date, Allan variance.

INTRODUCTION

There is an issue with the analysis of data sequences that are accumulated (generated) over time. That issue has acquired special value in connection with the development of the Internet of Things (IoT) and the autonomous operation of large networks of sensors. In such a network the need for the detection of abnormal operations or external (unauthorized) interference has of important. In other words, tasks of control of the proper operation of sensor networks (safety) and control of attacks on these networks (security) are actual.

There are many methods of solving similar problems, which include the analysis of time series [1], the search for "disorder" in the process [2], the use of event tracking systems [3], statistical methods [4], etc. Many of these methods require large amounts of data to be processed, which is justified in terms of obtaining detailed information, but IoT systems create excessive communication traffic and lead to an increase in power consumption. The challenge arises to reduce the cost of computing power and energy for data transmission and processing while maintaining the accuracy and detail of time series analysis. One of the possible options is the use of the method of structural functions, namely the Allan variance and related functions, which make it possible to obtain generalized characteristics of the stochastic process, based on which it is possible to determine the key parameters, by the change of which it is possible to determine the change in the nature of the process [5]. The result of such processing is a set of data, the number of which is proportional to $\log_2 n$, where n is the number of numerical values in the sequence. Thus, it is possible to achieve a significant reduction in the amount of data that must be transmitted for processing. Of course, it is important to understand what exactly is needed and possible to get using this method.

The most practically significant results of the use of system functions and Allan dispersion were obtained in the analysis of highly stable electromagnetic oscillators, gyroscopes, and other stable systems, where the deviations are many orders of magnitude smaller compared to the averaged values in the time series (when it can be determined) [5]. At the same time, there are examples of using this method for processes in which a different relationship between the deviation of parameters and their average value is observed [6]. For example, the analysis of the results obtained from application of Allan dispersion to the results of experiments with sensors of the Internet of Things shows that the calculated indicators of the processes are affected by the functioning properties of specific types of sensors: their dynamic characteristics, single outliers in the general data flow, etc.

The purpose of the work is to develop a method of analysis and preparation of data of time series of sensor for further processing by the method of structural functions, so that further processing allows to detect the characteristics of processes and their change with high reliability.

EXPOSITION

In research for purposes of automation of data collection an improvised IoT sensor node [7] consisting of an Arduino Uno board, an additional module with a flash drive and DS3231 real-time clock module was considered.

For an experimental study of the properties was used 4 types of sensors: DS18B20 temperature sensors (2 devices), two types of combined temperature and humidity sensors DHT11 (4 devices) and DHT22 (also 4 devices), the DS3231 includes a temperature sensor, which also was considered with other sensors. In total 11 temperature sensors located in relatively close physical conditions were used simultaneously in the study. All used sensor types are very common, so certain properties of their metrological characteristics are available in the corresponding datasheets.

The collection of dataflows from temperature sensors was provided entirely on the side of the Arduino controller by a specially developed algorithm using the PlatformIO toolkit within the VSCode IDE. The controller, performing tasks in autonomous mode, stored a measurements series of different intervals and periods for all connected sensors on the flash card. Some software tricks were made in the controller algorithm of the improvised sensor node to minimize the possibility of periodic jitter effect appearing in the data due to the moments of recording the measurement data on the flash card, creation of folders and files, which are needed quite often and could not be taken out of the total measurement time due to the limit of the controller's RAM and the lack of multitasking.

Further statistical analysis of data samples with automated conditional evaluation and systematization of the nature of changes during the measurement was performed on the PC using a specially written Python script, which generates of an aggregate brief report of data distribution by sensor type, measurement interval, value range.

The following naming conventions are used in the algorithms and below in the article: sample – data of a single measured temperature value; probe – an unitary continuous limited fragment of dataflow with given fixed sample rate and number of samples. Several probe samples are presented in the Fig. 1-3 on the left. Probes can be overlapped and nonoverlapped.

In total, 4075 nonoverlapped probes suitable for further analysis were obtained from 11 sensors. These probes include only measurements during which there was changes in the measured value, to which the Allan dispersion method can be applied. For several reasons the distribution turned out to be not homogeneous in relation to the types of sensors. The main reasons are the unequal number of various sensors involved in the experiment (which was due to the existing nomenclature at the time of the start of the experiment) and the capabilities of the used improvised model of the sensor node. In addition, the DS18B20 sensors were set to two different resolution options for separate series of measurements, which should be distinguished as two sub-types of sensors (see Fig. 2), and the integrated sensor DS3231 by default has the coarsest discretization of the sensors used, which in advance resulted in the production of a larger number samples with a fixed temperature value, which were rejected as uninformative and not suitable for variance estimation. Table 1 shows overall statistical data of the obtained probes.

Some of the obtained probes from individual sensors turned out to contain unintentionally arising unnatural values (see Fig. 3). Probably it was caused by several reasons or a combination of them, including poor contacts of the Arduino's connectors during part of the

Table 1.

Sensor types		DHT11	DHT22	DS18B20 (9 bit)	DS18B20 (12 bit)	DS3121
Devices		4	4	1	2	1
Sample rate (s)	min	10.0				
	max	2.5				
Samples	min	128				
	max	1024				
Temperature (°C)	min	55.4	50.8	44.5	42.25	43.5
	max	9.7	10.5	10.0	14.43	12.0
$\Delta T_{\min}$ (°C)		0.09	0.1	0.5	0.06	0.25
Probes	Total	1517	1510	145	452	285
	Natural	912	1410	145	412	265
	Anomaly	605	100	–	40	20

measurements, as well as fluctuations in current consumption and electromagnetic activity of the flash drive board located physically close to the data buses on the improvised sensor node (finally spikes in data received instead of the expected jitter). The DHT11 sensors turned out to be the most "sensitive" to the conditions over data collection phase of the experiment (possibly just for the reason that the connector from this group of sensors was not lucky when assembling the measuring system).

The resolution of all DHT22 and DHT11 type sensors participating in the study was 0.1 of the main temperature measurement scale unit, and DS18B20 sensors – 0.06 and 0.5 for individual instances 1 and 2, respectively. This property limits the high-frequency fluctuations of the output signal of the sensors and causes differences in the statistical properties of the accumulated data, which is confirmed by their processing.

Observation of the responses of considered sensors to temperature changes confirm that the

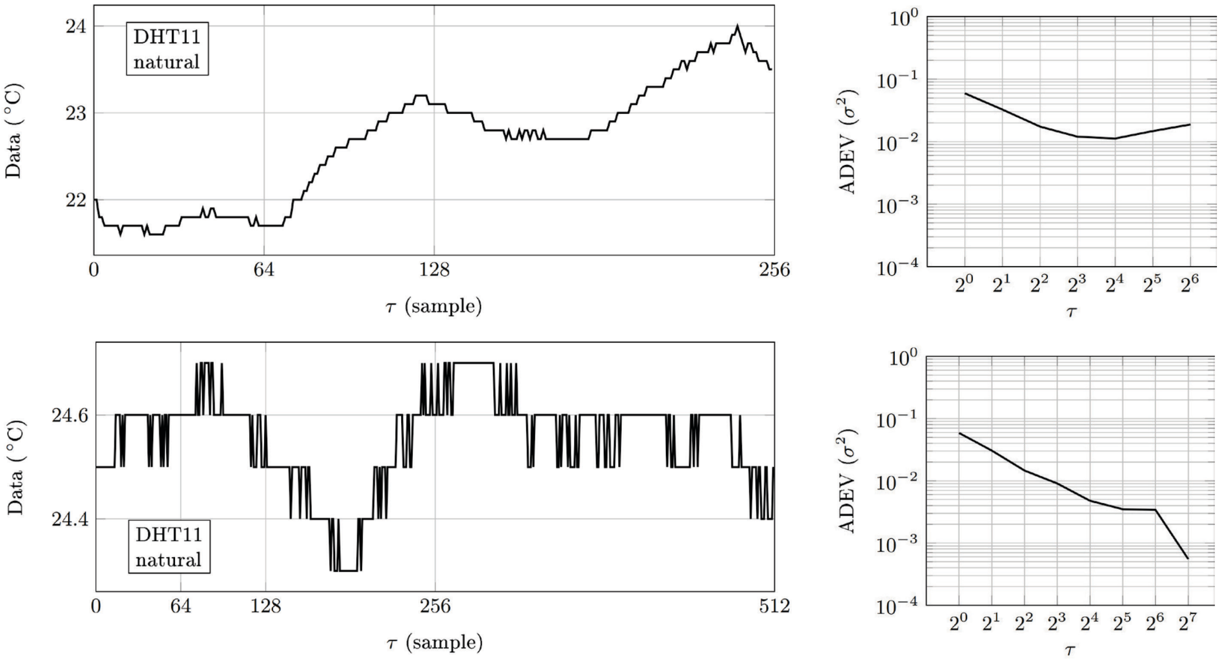

***Fig. 1.** Various samples of the DHT11 sensor data stream and the corresponding Allan variance*

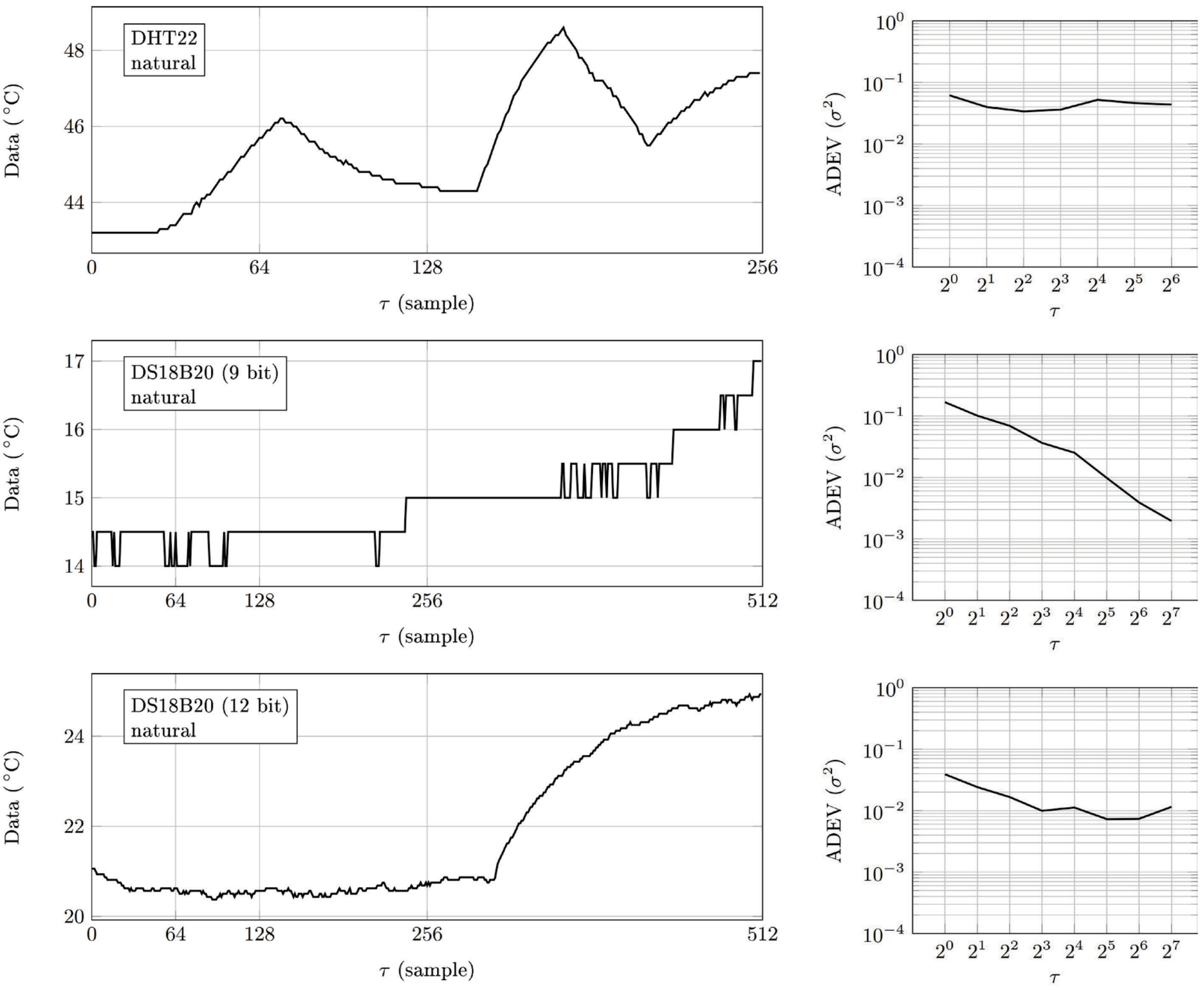

***Fig. 2.** Samples of the DHT22 and DS18B20 (with different resolution) data stream and the corresponding Allan variance*

received data have various dynamics, based on the type of sensor, the design differences, the transfer function of internal digital converters and transmission to the receiving device interface, as well as the individual specific of manufactured devices.

As it became clear after data analysis, a special limited representation of numerical values, which was used in the controller algorithm for intermediate storage, can also give an additional contribution to the transfer function for sensors with certain parameters of the digitalization of measured values due to the rising effect of multi-rate digital signal processing.

It is obvious that the time to establish the steady state of the sensor depends on its physical properties and, in the case of the studied sensors, has a systematic effect on high-frequency data changes, which can manifest itself against the background of a real change in the measured values. Differences in statistical data, caused by high-frequency fluctuations of the output signal depending on the resolution of the sensors under study, can be seen in the left section of the given graphs - in the case of DS18B20, the curves are expectedly higher for small values of τ. The graphs also clearly show the difference in the slope of the section with τ ~ 5–100 for the two types of sensors.

CONCLUSION

The dynamic characteristics of the sensors, primarily the parameters of establishing a stationary state or the relaxation time in combination with value of time interval between individual polls of the sensor state allow to rationally limit the necessary number of consecutive samples and separate the redundant data, which mainly characterizes the source of the measured data and insignificantly characterizes the measurement system.

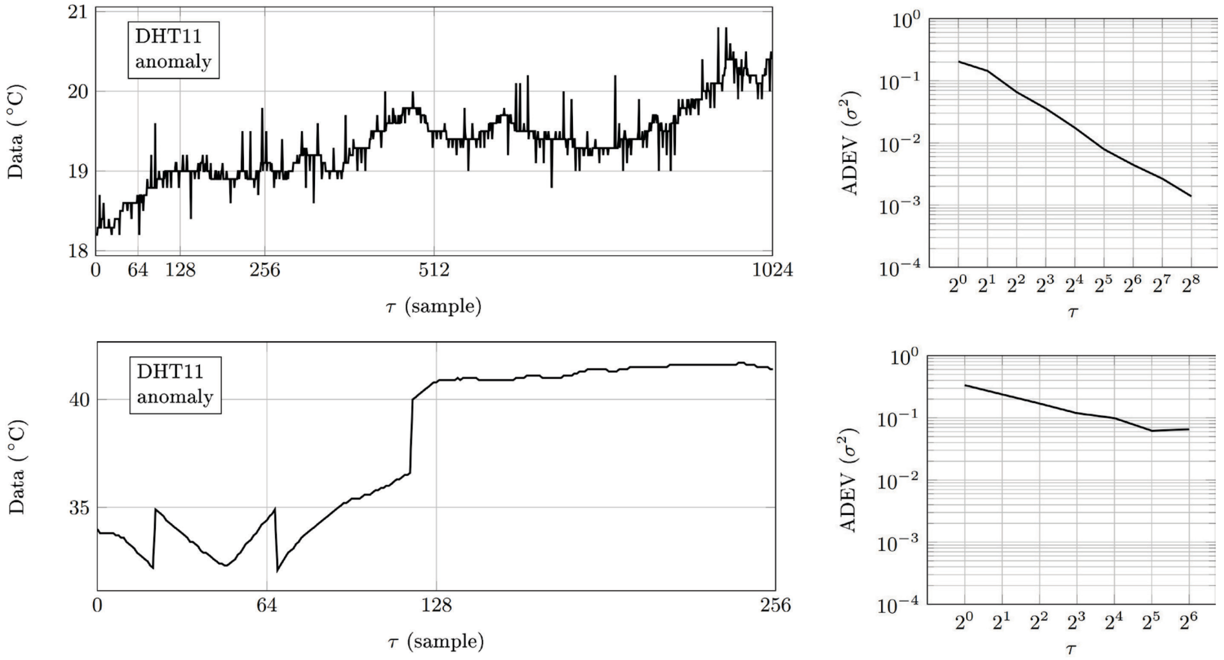

Fig. 3. *Samples of the DHT11 sensor data stream with anomalies of different kind and the corresponding Allan variances.*

Each specific manufactured temperature sensor has to a certain extent a personal signature of metrological characteristics, which can be used indirectly to identify the type of sensor or atypical data from a known instance of the sensor, which may be a sign of interference in the operation of the measuring system or, for example, an emergency change in the properties of the sensor physical environments, when controlled object cannot perform its regular function.

Processing of the data flow this way in the dynamic Allan variance mode allows to continuous monitor the state of the sensor network using relatively small computing resources and network traffic.

REFERENCE

[1] W. W. Wei. “Time series analysis.” in *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*, 2013, doi: 10.1093/oxfordhb/9780199934898.013.0022

[2] B. Yavorskyy, Yu. Leschyshyn, “Reliability of method for change-point detection of rhythm–cardiosignal,” *Visnyk TNTU*, vol 73, no 1, pp. 252-258, 2014, (In Ukrainian.)

[3] G. González-Granadillo, S. González-Zarzosa, R. Diaz, “Security Information and Event Management (SIEM): Analysis, Trends, and Usage in Critical Infrastructures,” *Sensors*, vol. 21, no. 14, p. 4759, 2021, doi: 10.3390/s21144759.

[4] Oakland, J., Oakland, R. J. *Statistical Process Control*, 7th ed. Routledge, 2018, doi: https://doi.org/10.4324/9781315160511

[5] Walls F.L., Allan, D.W. “Measurements of Frequency Stability”, *Proceedings of the IEEE*, vol. 74, no. 1, pp. 162-168, 1986.

[6] L. Galleani L., P. Tavella, “The Dynamic Allan Variance V: Recent Advances in Dynamic Stability Analysis,” in *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 63, no. 4, pp. 624-635, April 2016, doi: 10.1109/TUFFC.2015.2504124.

[7] V. G. Kryzhanovskyi, V. F. Komarov, S. P. Serhiienko and L. V. Zahoruiko, “Identification of Sensor Nodes Normal Operation Using Allan Variance,” *Visnyk VPI*, no. 3, pp. 78–83, Jun. 2021. (In Ukrainian.), doi: 10.31649/1997-9266-2021-156-3-78-83.