The quest for sense: Physical phenomena classification in the Internet of Things

João B Borges, Heitor S Ramos, Raquel A. F Mini, Aline Carneiro Viana, Antonio A. F Loureiro

The quest for sense: Physical phenomena classification in the Internet of Things*

João B. Borges*†, Heitor S. Ramos†, Raquel A. F. Mini‡, Aline C. Viana§, Antonio A. F. Loureiro†
* Department of Computing and Technology, Universidade Federal do Rio Grande do Norte, Caicó, RN, Brazil
† Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
‡ Department of Computer Science, Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, MG, Brazil
§ Inria Saclay, Palaiseau, France
Email: joaoborges@dct.ufrn.br, ramosh@dcc.ufmg.br, raquelmini@pucminas.br, aline.viana@inria.fr, loureiro@dcc.ufmg.br

***Abstract*—This paper investigates the precise identification of physical phenomena in the Internet of Things (IoT) context, which is one of the main challenges when dealing with the massive scale of IoT data. For this, we use information theory quantifiers in the characterization and classification of physical phenomena to minimize the effects of the lack of proper descriptions and the high heterogeneity of IoT sensors. Thus, by understanding the dynamics behind physical phenomena, we perform the classification of sensor data based on their expected behavior, not their data points. By using a simple classification algorithm, we show that the behavioral dynamics of some physical phenomena are more affected by different geographical regions than others. This gives a classification accuracy of 75% when all phenomena are considered and of 93% when considering only the invariant ones, with a worst case of false positives of 12%. This result indicates the high potential of our technique to correctly identify physical phenomena from sensor data, a fundamental issue for several applications, even in an unreliable IoT environment.**

***Keywords*-Internet of things, Time series characterization, Time series classification, Information theory quantifiers**

I. Introduction

The convergence between physical and informational worlds is increasingly becoming a reality. In recent years, the Internet of Things (IoT) [1], [2] has received special attention from both industry and academy, and plays an important role in this context. By their ability to interact with real world, IoT sensors are the "eyes" and "ears" for the new forthcoming systems. The number of IoT initiatives is rapidly growing around the world, with the potential to collect an unprecedented amount of data. From home automation [3] and agricultural applications [4] to large smart cities [5] and global weather forecasting [6], the range of physical phenomena being monitored, as the amount of generated data, is impressive. However, when dealing with this massive scale of IoT [7], a challenging question is *"how to precisely characterize and classify physical phenomena in data collected by a large population of sensors?"*

In this question, there are two issues to examine. First, a poor description of both data and resources of most deployed sensors in current IoT solutions, with their sensors deployed all over the world (some of those solutions have their data available on the Web, e.g., ThingSpeak IoT Platform – http://thingspeak.com [8]). Generally, they are described by a simple set of tags and textual information, freely assigned by their owners. This makes the analysis and, consequently, the understanding of the related physical phenomena prone to errors and misunderstandings. Second, the high heterogeneity of sensors brings differences in the resolution and magnitude of their collected data. This may affect distance-based algorithms when making comparisons between data collected by pairs of sensors.

Given such issues, we claim that solutions to handle data collected from IoT platforms should avoid the use and the processing of raw data. Instead, it is recommended (i) the extraction of features able to capture data attributes of the physical phenomena, and (ii) the analysis of such features using machine learning techniques (e.g., classification and clustering) [9]. This has the benefit of minimizing or circumventing the effects of inadequate descriptions and high data variability, as mentioned above. In order to perform the feature extraction and feature analysis, we will model the data collected from sensors as time series.

The starting point of the characterization and classification of physical phenomena is the identification of features to be extracted. This may lead to a problem in which features can depend on a given instance of the data and not represent the phenomena behind it. For instance, a feature vector of different sensors, even for the same phenomenon, but created at different moments and locations, may lead to completely different values, which may invalidate any further analysis [10].

To tackle all these challenges related to the precise identification of physical phenomena in the IoT context, we make the following contributions: (i) We use *information theory quantifiers* to precisely characterize physical phenomena. In this work, we will consider temperature, relative humidity, atmospheric pressure and wind speed, which are physical phenomena typically monitored in IoT solutions. These quantifiers allow us *to know the behavior behind a given physical phenomenon*; (ii) We model data collected from heterogeneous sensors, at different scales and sampling rates, as a *unified representation*. We advance the state-of-the-art of understanding and classifying IoT data by designing *strategies based on the phenomena' expected behavior, rather than on the raw data* of a sensor itself. This also helps in the scalability problem in IoT

* Invited paper

by avoiding comparisons with a large number of time series; and (iii) We perform the *classification of physical phenomena* by analyzing their placement on the Causality Complexity-Entropy Plane (CCEP). We *compare their placement with previously learned placement regions from known physical phenomena*. We also employ a robust method to minimize the lack of a proper distance measure for this plane, *helping further investigations* and *reducing the gap* between the theory from Bandt-Pompe transformations (method used in this work) and their application to problems in real-world scenarios.

The rest of this paper is organized as follows. Section II discusses the related work and motivates our main problem. Section III presents the information theory concepts and the background necessary to understand the proposed technique. Section IV presents a data characterization of the physical phenomena used in this work, and Section V shows our classification strategy to correctly identify an appropriate sensor monitoring a given data type. Section VI concludes the work and presents some open research issues.

II. Related work and problem definition

Time series analysis is a research topic that gained a new breath in last years, mainly due to the increasing attention for big data, machine learning, and IoT initiatives. A considerable number of algorithms and strategies has been proposed [9], in which classification and clustering are the most representative ones for the supervised and unsupervised cases, respectively.

These solutions are, basically, divided in those which use raw time series data and those in which some processing is applied before extracting information. The algorithms based on raw data generally compute some distance metric (e.g., Euclidean, Fréchet, dynamic time warping (DTW) [11]) between pairs of time series. Those with the lower distances are grouped together (clustering), or identified as similar (classification). However, for IoT scenarios, in which there is a vast heterogeneity of sensors, some assumptions can not be made. For instance, most of these solutions assume that time series have similar length, or they are sampled at the same constant rate. Furthermore, due to the high number of available sensors, it is not scalable to perform direct comparisons on raw data, making many strategies unsuitable for this scenario.

On the other hand, strategies based on extracting features from the time series seem to be more appropriate to the problem. These solutions are based on the analysis of a new vector of features extracted from the time series [10], [12]. One of the first advantages of this strategy is the dimensionality reduction of the time series, which are expected to have a very large length for in IoT. This reduction and equalization of sizes was also the subject of many studies in the literature, via some smoothing technique or symbolic approximation [13].

Another point to be concerned in feature extraction is related to the decision of which features to compute. Since we are interested in physical phenomena, this may lead to features that can be dependent on the current instance of time series, not representing the phenomena behind it. For instance, features from different time series, even for a same given phenomenon, may lead to completely different values, which may invalidate any further analysis. For this cases, they may change as a function of time and space, requiring another extraction for new features, that could be computationally expensive [10].

Thus, it is reasonable to consider as features those metrics that could represent the behavior of a phenomenon measured by the sensor, independently of the current sample. In this direction, metrics originated from information theory seem to be appropriate to extract this knowledge from a phenomenon. The use of information theory quantifiers has been applied for characterizing time series in several studies. These metrics proved to be effective in distinguishing different dynamic behaviors of time series [14]–[16], being useful in the characterization of real-world time series [17], [18].

In this work, we use information theory quantifiers to characterize the behavior behind physical phenomena. Our aim is better to identify time series by just comparing their behaviors, not their data points. This allows avoiding the comparisons for identification, between a large number of time series, thus, solving the scalability problem in IoT. We also base the extraction of the quantifiers in a transformation from the time series, which is valid for the dimensionality reduction of the data as well as to increase their robustness to IoT problems, discussed in the following sections.

III. The Bandt-Pompe transformation and information theory quantifiers

Before extracting the information theory quantifiers from physical phenomena, we have to first transform their data from time domain to another more appropriate one. In this section we present the Bandt-Pompe transformation and the process of calculating those quantifiers from it. We also show that this transformation meets our needs, since it better represents the time series dynamic behavior, being fundamental for the proper calculation of our metrics. We also present the CCEP, which is a method that combine those metrics in a plane, so different time series dynamics could be distinguished by different placement regions on it.

A. The Bandt-Pompe Distribution

The Bandt-Pompe method [19] consists of transforming a given time series onto a set of symbolic patterns, each of size D, according to the ordinal relation between neighboring data samples. From these patterns, it follows by assigning a probability distribution to the time series, by counting the frequency of each pattern [17], [20]. In the following we present a more formal description for this process.

For a given time series $\mathbf{X} = \{x_1, \ldots, x_T\}$ and an embedding dimension $D \in \mathbb{N}$, the Bandt-Pompe transformation performs by generating partitions $\mathbf{P}_t \subseteq \mathbf{X}$ of size D, such that, for each instant $t \in \{D, \ldots, T\}$, $\mathbf{P}_t = \{x_{t-(D-1)}, x_{t-(D-2)}, \ldots, x_{t-1}, x_t\}$. The ordinal relation for each of these instants t consists in the permutation $\pi = (r_0, r_1, \ldots, r_{D-1})$ of $(0, 1, \ldots, D-1)$ subject to $x_{t-r_{D-1}} \leq x_{t-r_{D-2}} \leq \cdots \leq x_{t-r_1} \leq x_{t-r_0}$ [14].

In other words, π represents the necessary permutation for the elements of $\mathbf{P}_t$ to be sorted in ascending order. For each permutation π of all $D!$ possible permutations of order D, let $\mathbf{P}_\pi = \{t : D \leq t \leq T\}$ be the set of all partitions that has the permutation type π, and $|\mathbf{P}_\pi| \in \{0, \ldots, T - D + 1\}$ be the number of partitions of type π, then the probability distribution $P = \{p(\pi)\}$ is defined by

$$p(\pi) = \frac{|\mathbf{P}_\pi|}{T - D + 1}, \tag{1}$$

satisfying the conditions $p(\pi) >= 0$ and $\sum_\pi p(\pi) = 1$.

The choice of D depends on the length T of the time series, and the condition $T \gg D!$ must be satisfied in order to obtain reliable statistics [20]. For practical purposes, Bandt and Pompe [19] recommended D in the interval $[3 \ldots 7]$.

A variant of this method considers inserting an embedding delay $\tau \in \mathbb{N}$ [17], such that the elements of each partition be separated by intervals of size τ, corresponding to a sample by regular spaced intervals [20]. Thus, the partitions are defined as $\mathbf{P}_t = \{x_{t-(D-1)\tau}, x_{t-(D-2)\tau}, \ldots, x_{t-\tau}, x_t\}$, and the probability distribution as

$$p(\pi) = \frac{|\mathbf{P}_\pi|}{T - (D + 1)\tau}. \tag{2}$$

A detailed study on the impact of τ in the complexity measures is given by Zunino et al. [20]. It is clear that the method originally proposed by Bandt and Pompe is equivalent to the case where $\tau = 1$.

The Bandt-Pompe transformation allows the calculation of measures from this distribution generated by the ordinal permutations, with the advantages to be simple, fast to calculate, robust in the presence of observational and dynamic noise, and invariant with respect to nonlinear monotonous transformations [14], [17]. With this method, some details of the amplitude and variability of the original time series are lost, however, it is very suitable for the analysis of experimental data since it avoids amplitude threshold dependencies that affect other methods based on range partitions [20].

B. Information theory quantifiers

Following the initial purpose of Bandt and Pompe, the information quantifiers are calculated from the distribution $p(\pi)$, for all $D!$ permutations π of order D. For our purposes these will be the metrics used to express different time series dynamics and, thus, making their distinction possible.

1) Permutation Entropy: Since the Bandt-Pompe distribution is based on the permutations of neighboring values, the authors proposed a variation from the classical entropy of Shannon, which they called *permutation entropy*, and is defined as

$$H[P] = -\sum p(\pi) \log p(\pi), \tag{3}$$

where $0 \leq H[P] \leq \log D!$. The permutation entropy is equivalent to the Shannon entropy and is a measure of uncertainty associated to the process described by P [17]. Lower values of $H[P]$ represent an increasing or decreasing sequence of values in the permutation distribution, indicating that the original time series is deterministic. On the other side, high values of $H[P]$ indicate a completely random system [19].

2) Normalized Shannon Entropy: The maximum value for $H[P]$ occurs when all $D!$ possible permutations have the same probability to occur, which is the case for the uniform distribution P_u of permutations. Thus, $H_{\max} = H[P_u] = \log D!$, where $P_u = \{1/D!, \ldots, 1/D!\}$. [20].

Rosso et al. [14] defined the normalized Shannon entropy, from the permutation entropy case, as

$$H_S[P] = \frac{H[P]}{H_{\max}}, \tag{4}$$

where $0 \leq H_S[P] \leq 1$.

3) Statistical Complexity: Another statistical measure that can be computed from the permutation distribution is the statistical complexity. Defined by Lamberti et al. [21], this measure is another point of view concerning the knowledge of some underlying process, based on the Jensen-Shannon divergence JS between the associated probability distribution P and the uniform distribution P_u, i.e., the trivial case for the minimum knowledge from the process. The statistical complexity is given by

$$C_{JS}[P] = Q_{JS}[P, P_u] H_S[P], \tag{5}$$

where $P = \{p(\pi)\}$ is the Bandt-Pompe distribution, P_u is the uniform distribution, and $H_S[P]$ is the normalized Shannon entropy, as described above.

The disequilibrium $Q_{JS}[P, P_u]$ is given by

$$Q_{JS}[P, P_u] = Q_0 JS[P, P_u] \tag{6}$$

$$= Q_0 \left\{ S\left[\frac{P + P_u}{2}\right] - \frac{S[P] + S[P_u]}{2} \right\}, \tag{7}$$

where S is the Shannon entropy measure and Q_0, given by

$$Q_0 = -2 \left\{ \left(\frac{D! + 1}{D!}\right) \ln(D! + 1) - 2\ln(2D!) + \ln(D!) \right\}^{-1}, \tag{8}$$

is a normalization constant equal to the inverse of the maximum value of $JS[P, P_u]$, so $0 \leq Q_{JS} \leq 1$ [16], [17].

C. Causality Complexity-Entropy Plane

Both normalized Shannon entropy and statistical complexity measures are are very suitable quantifiers to extract knowledge from the dynamic behavior of a given process and, thus, distinguishing them. Based on these measures, Rosso et al. [14] proposed the CCEP, which is a 2-dimensional metric space built by the statistical complexity (C_{JS}) as the y-axis and the normalized Shannon entropy (H_S) as the x-axis.

An important characteristic of the CCEP is the placement of time series with different dynamical behaviors at specific regions in the plane. For instance, the plane allow us distinguishing deterministic, stochastic, and chaotic time series, based on their placement on it [14]. Figure 4 illustrates different planes for $D = \{4, \ldots, 7\}$. For different values of H_S and order D, the statistical complexity measure ranges between a minimum and maximum limits, as illustrated in the figure (for more details, please see [14]).

To understand the intuition behind the CCEP, it is important to understand the particularities of each of their measures. With the normalized Shannon entropy it is possible to measure the amount of uncertainty one may have from the process, ranging from the certain prediction of the possible values ($H = 0$) to the maximum uncertainty (uniform distribution) ($H = 1$) [21]. On the other hand, the statistical complexity measure is able to measure the levels in which these dynamics affect the time series. A higher statistical complexity means that both regular and random behaviors are present in a given time series at the same level, i.e., in a equilibrium. Otherwise, it measures if these behaviors are present but one having more influence than the other, where a zero value means only one of them is present. Those measures, when combined, reveal some details of the dynamics of the time series, which can be used for effectively discerning among different behaviors, such as deterministic, randomness, and chaotic behaviors [14].

IV. Characterization of physical phenomena

Focusing on the provision of a better understanding of real-world data, this section presents a characterization of physical phenomena (e.g., temperature, atm. pressure, humidity, and wind speed), when applying the previous discussed metrics. As previously discussed, the Bandt-Pompe approach has some properties that make it well suited for its application to real data. Furthermore, using the normalized Shannon entropy (H_S) and the statistical complexity (C_{JS}) as information quantifiers along with the CCEP, allow us distinguishing the different time series behaviors.

a) Dataset selection: When dealing with IoT sensors, there are some uncertainty and unreliability that must be considered before dealing with their data. For instance, for sensors available in IoT platforms, such as the case of ThingSpeak, the data and resources usually are not described or, when it does, have a very poor description. Even if the descriptions were given, since there is no validation, there is no guarantee that it is correct. Generally, they are described by a simple set of tags and textual information, freely assigned by their owners and/or users, which makes the search for a keyword or expression prone to errors and misunderstandings [8].

Table I summarizes the lack of descriptions for the IoT sensors available in the ThingSpeak platform, collected at three distinct periods: October 31, 2015, and July 28, 2016, and April 14, 2017. As noted, the number of available sensors, as the time series per monitored phenomenon increased throughout the years, but also their problems.

For those sensors where some description was provided, Fig. 1 illustrates some of the most frequent words used to describe them, for April, 2017. While it can be noticed that most descriptions are, indeed, related to environmental and physical phenomena, they are still really poor, mainly if compared to more complex scenarios, such as semantic and ontology solutions [22]–[24]. The top-5 most frequent words for the three years are: *temperature*, *humidity*, *temp*, *field*, and *sensor*. These last three are generic descriptors, which add no valuable information to the monitored phenomena.

TABLE I
Summary of problems in the descriptions of IoT sensors, from the ThingSpeak platform, for the periods: October 31, 2015, July 28, 2016, and April 14, 2017. * t/s: time series per sensor.

Issue	2015	2016	2017
# of sensors	4064	7427	13013
# of time series	15170 (3.7 t/s)	28966 (3.9 t/s)	51185 (3.9 t/s)
No description	1655 (40.7%)	3681 (49.6%)	6632 (50.9%)
No location	3208 (78.9%)	6046 (81.4%)	10856 (83.4%)
No tags	3233 (79.5%)	6179 (83.2%)	11045 (84.9%)
No desc. & tags	1564 (38.5%)	3499 (47.1%)	6318 (48.5%)

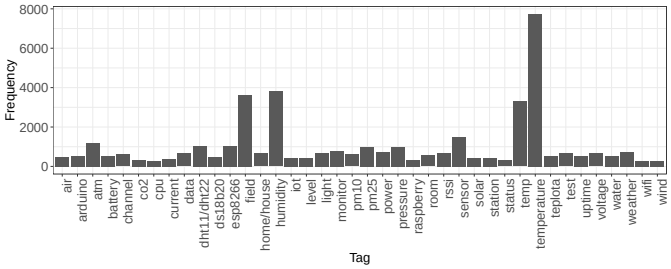

Fig. 1. Most frequent words used to describe IoT sensors, from descriptions of sensors available in the ThingSpeak platform, collected at April of 2017.

Thus, given this uncertainty and unreliability of sensors, in order to correctly evaluate our hypothesis, we decided not to apply the proposed strategy directly to the IoT data. Instead, we looked at similar data but from more reliable sources. Thus, to investigate the feasibility of using CCEP to study physical phenomena, we considered time series data measured by international airport stations from 60 different airports from all regions of the USA territory. Table II presents the list of these places. Data are measures from the historical weather conditions of temperature, relative humidity, atmospheric pressure, and wind speed, in the period from 2000 to 2015. All data were collected from the Weather Underground[1] platform.

b) Dataset characterization: The dataset was randomly split in half, where the time series from 30 places was used to the characterization step and the other 30 to validate the classification of the phenomena. To illustrate the behavior of the data we are dealing with, Fig. 2 gives an example of time series for 2015, measured at the Logan International Airport, Boston, MA. For each phenomenon, it is also showed a Lowess smoothing of the data with smoother span $f = 0.1$.

We can see that the behavior of the data for each different phenomenon is quite different. From a more seasonal behavior for temperature to a more "random" behavior for the wind speed. Fig. 3 shows the Bandt-Pompe distributions for the dataset in the period between 2000 and 2015, considering an embedding dimension $D = 4$. We can see that, for each phenomenon, the ordering patterns have different probabilities. This is the behavior that is captured by the aforementioned information quantifiers H_S and C_{JS}. To better understand these behaviors, Fig. 4 presents the CCEP for the whole time series, considering different embedding dimensions $D = \{4, \ldots, 7\}$.

[1]Weather Underground – http://wunderground.com.

TABLE II
LIST OF PLACES IN USA, ORDERED BY STATE, USED FOR THE CHARACTERIZATION AND CLASSIFICATION PHASES.

ID	Place	Type	ID	Place	Type
1	Anchorage	F	31	Detroit	C
2	Phoenix	C	32	Minneapolis	F
3	Los Angeles	C	33	Charlotte	C
4	Oakland	F	34	Raleigh	F
5	Ontario	F	35	Omaha	F
6	Sacramento	F	36	Newark	F
7	San Diego	F	37	Albuquerque	F
8	San Francisco	C	38	Las Vegas	C
9	San Jose	F	39	Buffalo	F
10	Santa Ana	F	40	New York (Central Park)	C
11	Denver	C	41	New York (JFK)	F
12	Hartford	F	42	New York (LaGuardia)	C
13	Fort Myers	C	43	Cleveland	C
14	Fort Lauderdale	C	44	Columbus	F
15	Jacksonville	F	45	Cincinnati	F
16	Miami	C	46	Portland	C
17	Orlando	F	47	Philadelphia	C
18	Tampa	C	48	Pittsburgh	F
19	West Palm Beach	F	49	Nashville	F
20	Atlanta	C	50	Austin	C
21	Honolulu	C	51	Dallas	C
22	Kahului	F	52	Dallas (Fort Worth)	C
23	Chicago (Midway)	C	53	Houston (Bush)	F
24	Chicago (O'Hare)	C	54	Houston (Hobby)	C
25	Indianapolis	F	55	San Antonio	C
26	New Orleans	F	56	Salt Lake City	F
27	Boston	C	57	Washington (Reagan)	C
28	Baltimore	C	58	Washington (Dulles)	F
29	Kansas City	F	59	Seattle	C
30	St. Louis	C	60	Milwaukee	F

Legend: C - Used for Characterization / F - Used for Classification

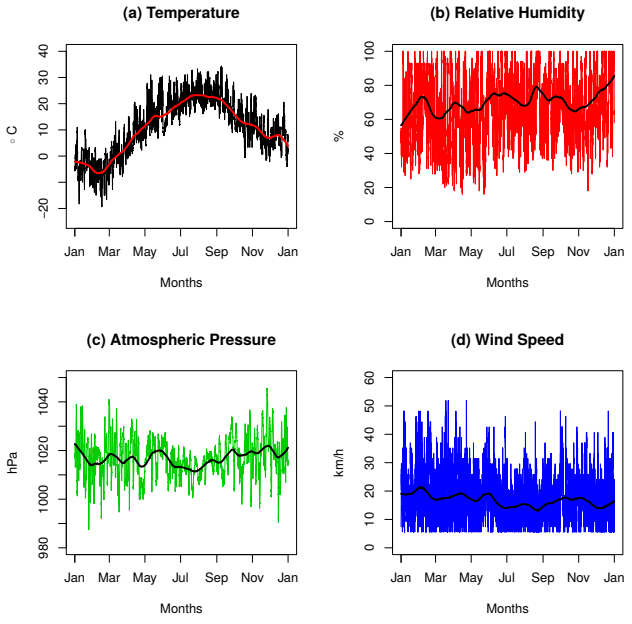

Fig. 2. Historical data of 2015 from the Logan International Airport, Boston, MA. Phenomena (a) temperature, (b) relative humidity, (c) atmospheric pressure, and (d) wind speed, along with a Lowess smoothing ($f = 0.1$).

The rightmost region of CCEP (near $H_S = 1$ and $C_{JS} = 0$) represents a totally random behavior such as a white noise. On the other hand, the time series that present strong regularity and more correlation between neighboring values tend to lie in the leftmost part of CCEP (near $H_S = 0$ and $C_{JS} = 0$). The region of high C_{JS}, i.e., the upper-center region of the plane represents time series with chaotic behavior [14].

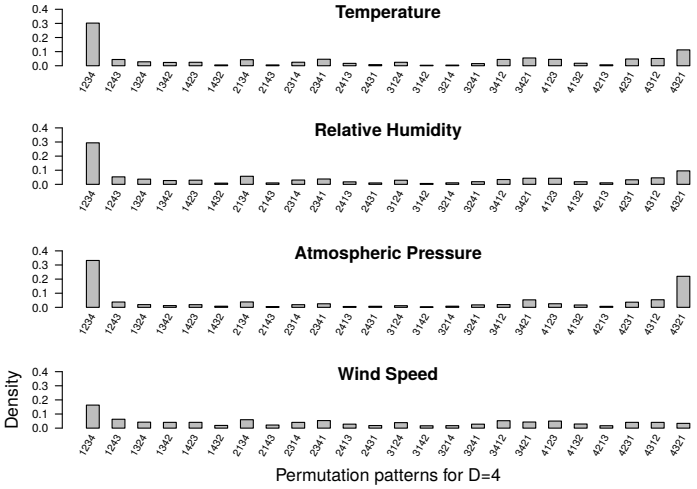

Fig. 3. Bandt-Pompe distributions of Boston's weather historical data in the period between years 2000 and 2015. All plots considered $D = 4$.

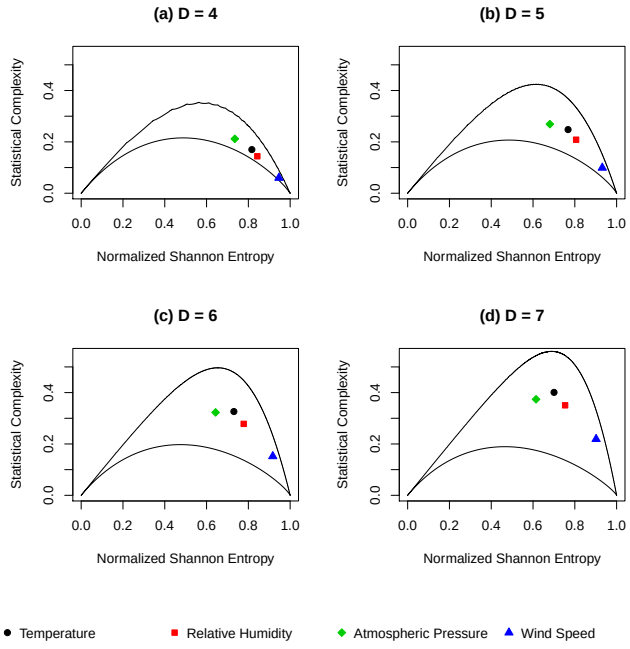

Fig. 4. CCEP of Boston's weather historical data between years 2000 and 2015, for $D = \{4, \ldots, 7\}$.

Fig. 4 shows that, for all values of D, phenomena lie in the region of medium-high entropy values ($0.6 < H_S < 1$), which are similar to the region that characterizes "colored" random noises, representing different correlation values in the time series structures [14]. Also, as D increases, the regularities of the phenomena are better captured by the information quantifiers. This can be expressed by the increasing of C_{JS}, the decreasing of H_S, and the more clear separation between the points for each phenomenon in the plane.

Following our analysis, in order to verify the feasibility of

this method in characterizing the physical phenomena *per se*, Fig. 5 presents the CCEP representation for all phenomena studied herein, from the 30 places used for the characterization phase, with an embedding dimension $D = 6$. The reason for choosing $D = 6$ is due to a trade-off between a concise capture of the regularities of the phenomena and the small value of $D!$, which impacts on the required length of the time series.

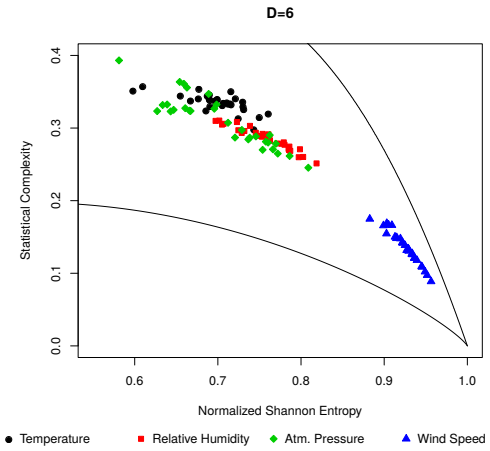

Fig. 5. CCEP of historical weather data in the period between years 2000 and 2015 of all the places used for characterization, with $D = 6$.

Fig. 5 also shows that, with exception for the atmospheric pressure, all the phenomena lie at a specific region in the plane and has an "expected behavior", described by the shape in which the points are scattered. The time series for the atmospheric pressure present different behaviors in different places, and are more spread over the plane.

Fig. 6 illustrates the two most extreme time series, i.e., the ones which are farther from their cluster regions when considering temperature and atmospheric pressure. We can see that, while there is a small variation between the temperatures of Denver and Phoenix, for the atmospheric pressure, this difference is more apparent, and the time series of Denver seems to be more noisy than that for Honolulu. This difference on behavior results in different values of the information quantifiers and, thus, different placements in CCEP. This may lead to the need of a study about the geographical influence on these measures, which we will conduct as a future work.

V. Classification of Physical Phenomena

In this section, we present our strategy for the classification of physical phenomena from the IoT sensors. The first step is to learn the expected behaviors from different phenomena and, hence, identify their placement regions in the CCEP by clustering their placements. Thus, we will be able to verify if a given time series are similar to an already known phenomenon, by comparing the distance of its placement in the plane to the previous learned placements.

A. Identifying regions on CCEP

To define the expected placement region in the CCEP for a given phenomenon, we have to consider their concentration of points in the plane and estimate a centroid to represent it.

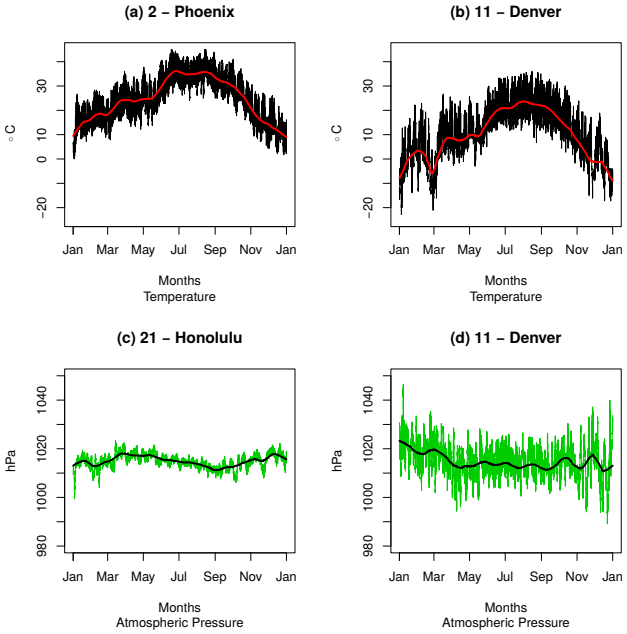

Fig. 6. Examples of historical data for 2015 of the most divergent places in the CCEP. For the temperature of (a) Phoenix and (b) Denver, and the atmospheric pressure for (c) Honolulu and (d) Denver, with a Lowess smoothing ($f = 0.1$).

Fig. 7 presents a heat map, built with a kernel density estimation from the points showed in Fig. 5, illustrating the concentration of points in CCEP for the phenomena under analysis. We can see that, for the temperature, relative humidity and wind speed, there is a regularity in the concentration of points around a particular natural centroid. Furthermore, they form three different grouped clusters. For the case of atmospheric pressure, the points are more spread and two centroids are highlighted, resembling a bimodal data.

Issues related to mixture models to fit bimodal data will not be covered in the present work and will be the subject of a future study. For our current purposes, it is sufficient to identify the centroids in which the points are surrounding and accept the fact that the atmospheric pressure can be characterized by two different regions on the plane.

Another point we must also be concerned when discovering the regions in the plane is about the effects of noise and imprecision in the information quantifiers. Thus, before obtaining the centroids, we first apply the Skinny-dip clustering algorithm [25] on the points for each phenomenon. Skinny-dip is a noise-robust clustering algorithm based on the Hartigan's dip test of modality and is able to reasonably detect the more distinguishing concentration of points in a given region, even under a rate of 80% of noise [25].

Fig. 8 depicts the resulting clusters after the Skinny-dip processing for the points of each phenomenon. Gray points are those considered noisier by the algorithm. We can see that, for the atmospheric pressure points, there are clearly two formed clusters. After discovering the most significant points, we performed a kernel density estimation (KDE) to

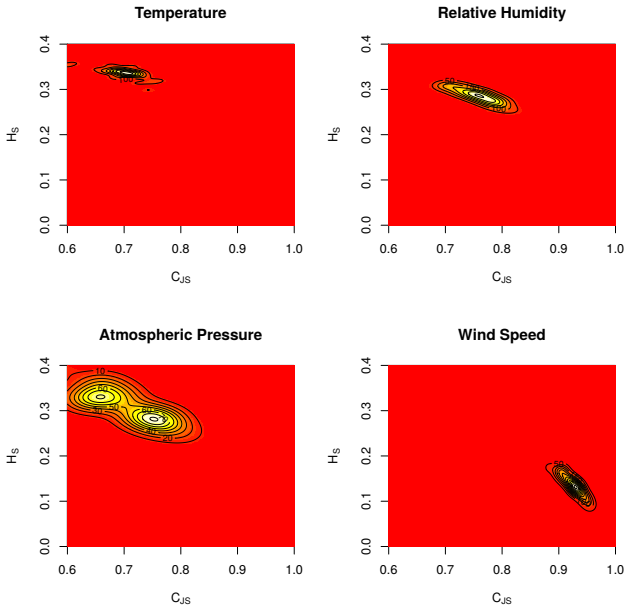

Fig. 7. Heat map illustrating the concentration of points from information quantifiers related to the characterization of the temperature, relative humidity, atmospheric pressure, and wind speed phenomena, with $D = 6$.

each cluster and interpolate between their points to compute the centroids. Table III illustrates the values of the centroids computed for each discovered cluster.

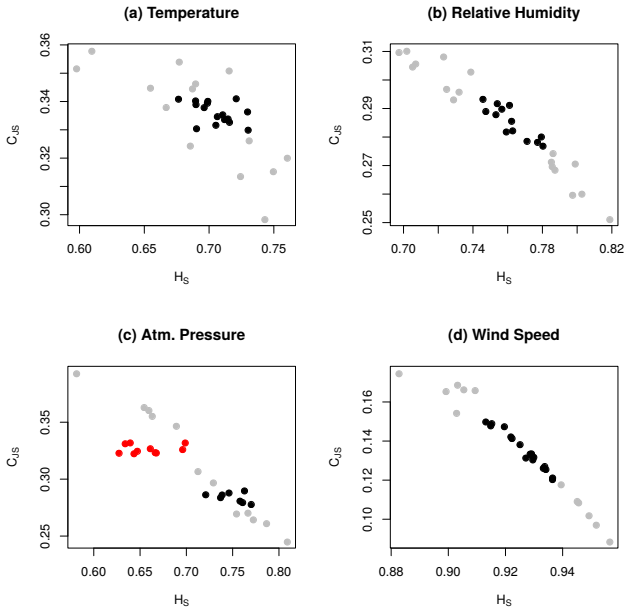

Fig. 8. Most relevant points discovered by Skinny-dip clustering algorithm for the physical phenomena, with $D = 6$.

B. Classification of time series in the CCEP

The next step towards the classification of a given time series is to place it in the CCEP and verify if it lies close to some know phenomenon placement. A point to consider here is that the notion of distance in this plane is still an open research question. For instance, although the values for H_S ranges from 0 to 1, the values for C_{JS} are bounded by their limits, impacting on direct distance metrics. This occur because C_{JS} behavior is governed by patterns in the probability distribution space that gives, for the same entropy, different levels of statistical complexity.

However, to illustrate the feasibility of the method for the classification of physical phenomena, we used the Euclidean distance, despite not being the most appropriate, to compute the nearest centroid for the time series of the places present in our dataset. We show the classification results in terms of Accuracy (A), true positive rate per class (tp) and false positive rate per class (fp), where $A = \frac{tp+tn}{tp+tn+fp+fn}$, where tn is the true negative rate and fn is the false negative rate.

Table IV summarizes the results for the classification process using the CCEP method with the Euclidean distance (CCEP_E) to discover the type of the time series. To perform this identification, we compute the Euclidean distance between the position of a given time series in the plane to the already known centroids, showed in Table III. We use a simple classification technique that assigns a given time series to the same type of the closest centroid.

TABLE III
CENTROIDS OF CCEP REGIONS FOR EACH PHENOMENON, FROM A KERNEL DENSITY ESTIMATION ON THE MOST RELEVANT POINTS.

Phenomenon	H_S	C_{JS}
Temperature	0.711	0.334
Relative Humidity	0.755	0.290
Atmospheric Pressure 1	0.739	0.286
Atmospheric Pressure 2	0.658	0.324
Wind Speed	0.930	0.131

TABLE IV
RESULTS OF THE NUMBER OF TRUE POSITIVE AND FALSE POSITIVE IDENTIFICATIONS FOR THE PHYSICAL PHENOMENA TIME SERIES WITH THE CCEP METHOD FOR THE EUCLIDEAN DISTANCE (CCEP_E).

Metric	**Value**
Accuracy (A)	0.75 (90/120)
Accuracy (without Atm. Pressure)	0.93 (84/90)
Temperature (tp)	0.83 (25/30)
Humidity (tp)	0.67 (20/30)
Pressure (tp)	0.53 (16/30)
Wind Speed (tp)	0.97 (29/30)
Temperature (fp)	0.10 (9/90)
Humidity (fp)	0.10 (9/90)
Pressure (fp)	0.12 (11/90)
Wind Speed (fp)	0.01 (1/90)

For the two general configurations we consider the cases where the atmospheric pressure time series, as their centroids, were present in the experiment and not. We can see that, since the atmospheric pressure phenomenon was the most difficult to estimate its behavior, when it is included in the experiment the number of correct identifications is about 75%, which means specifically a number of 90 out of 120 in total. The total of 120 series is from the 30 places used for classification, such that each place has four time series for their phenomena. Without the pressure being considered, this number increases to 93%,

84 out of the 90 total time series were correctly classified.

For each individual phenomenon class, we can see a reasonable total of true positives, with the wind speed being the most correctly identified, with a total of 97%. The atmospheric pressure has the lowest true positive rate, only 53%. This result is indeed expected, since the behavior of the pressure phenomenon was the hardest to be characterized. On the other hand, as the most stable phenomena in the characterization, the wind speed was the one with the best results.

Another important aspect for this classification is regarding the false positive values. In fact, according to the application that will be using an IoT sensor, maybe worse than not finding a sensor to be used is to find a wrong sensor. For this sort of problems, the method also seems to be reasonable with the highest rate of wrong identification being 12% for the atmospheric pressure phenomenon. Furthermore, as mentioned before, even with this promising results, the Euclidean distance is not the most appropriate for the current method. It is expected that, with a proper distance metric to this plane, these results will be improved.

VI. Conclusions and future directions

In this work, we proposed the application of information theory quantifiers, namely the normalized Shannon entropy and statistical complexity, to extract knowledge regarding the expected behavior of physical phenomena in the context of IoT. We showed that the Bandt-Pompe approach and the CCEP plane provide robust quantifiers to face the challenges related to this particular scenario.

In order to perform the classification of the physical phenomena, we proposed a definition of regions within the plane. Those placement regions were defined by the application of a noise-robust strategy for finding their centroids, which resulted in significant quality for the process. All these contributions clearly advance the state of the art in the characterization and classification of physical phenomena in the Internet of Things.

As future work, we open a number of questions related to the advances in the definition of distance metrics in the CCEP space, the need for studying the geographical influence on this information measures, and novel approaches to minimize the effect of IoT related problems in the correct characterization and identification of physical phenomena.

Acknowledgment

We would like to thank the research agencies, CAPES, CNPq, FAPEMIG, and grants #15/24536-2 & #15/24494-8, São Paulo Research Foundation (FAPESP).

References

[1] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[2] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, sep 2013.

[3] M. Coronado and C. A. Iglesias, "Task automation services: Automation for the masses," *IEEE Internet Computing*, vol. 20, pp. 52–58, 2016.

[4] M. Wu, Y. Wang, and Z. Liao, "A new clustering algorithm for sensor data streams in an agricultural IoT," *Proceedings - 2013 IEEE International Conference on High Performance Computing and Communications, HPCC 2013 and 2013 IEEE International Conference on Embedded and Ubiquitous Computing, EUC 2013*, pp. 2373–2378, 2014.

[5] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 22–32, 2014.

[6] S. Greengard, "Weathering a New Era of Big Data," *Association for Computing Machinery. Communications of the ACM*, vol. 57, no. 9, p. 12, sep 2014.

[7] J. A. Stankovic, "Research Directions for the Internet of Things," *Internet of Things Journal, IEEE*, vol. 1, no. 1, pp. 3–9, feb 2014.

[8] J. B. Borges Neto, T. H. Silva, R. M. Assunção, R. A. F. Mini, and A. A. F. Loureiro, "Sensing in the collaborative internet of things," *Sensors (Switzerland)*, vol. 15, no. 3, pp. 6607–6632, mar 2015.

[9] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering - A decade review," *Information Systems*, vol. 53, pp. 16–38, 2015.

[10] X. Wang, K. Smith, and R. Hyndman, "Characteristic-Based Clustering for Time Series Data," *Data Mining and Knowledge Discovery*, vol. 13, no. 3, pp. 335–364, sep 2006.

[11] P. Montero and J. A. Vilar, "TSclust : An R Package for Time Series Clustering," *Journal of Statistical Software*, vol. 62, pp. 1–43, 2014.

[12] B. D. Fulcher, M. a. Little, and N. S. Jones, "Highly comparative time-series analysis: the empirical structure of time series and their methods," *Journal of The Royal Society Interface*, vol. 10, no. 83, apr 2013.

[13] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery - DMKD '03*. New York, New York, USA: ACM Press, 2003, p. 2.

[14] O. A. Rosso, H. A. Larrondo, M. T. Martin, A. Plastino, and M. A. Fuentes, "Distinguishing Noise from Chaos," *Physical Review Letters*, vol. 99, no. 15, p. 154102, oct 2007.

[15] L. Zunino, M. C. Soriano, I. Fischer, O. A. Rosso, and C. R. Mirasso, "Permutation-information-theory approach to unveil delay dynamics from time-series analysis," *Physical Review E*, vol. 82, no. 4, 2010.

[16] O. A. Rosso, L. Zunino, D. G. Pérez, A. Figliola, H. A. Larrondo, M. Garavaglia, M. T. Martín, and A. Plastino, "Extracting features of Gaussian self-similar stochastic processes via the Bandt-Pompe approach," *Physical Review E*, vol. 76, no. 6, p. 061114, dec 2007.

[17] A. L. Aquino, H. S. Ramos, A. C. Frery, L. P. Viana, T. S. Cavalcante, and O. A. Rosso, "Characterization of electric load with Information Theory quantifiers," *Physica A: Statistical Mechanics and its Applications*, vol. 465, pp. 277–284, jan 2017.

[18] B. A. Gonçalves, L. Carpi, O. A. Rosso, and M. G. Ravetti, "Time series characterization via horizontal visibility graph and Information Theory," *Physica A: Statistical Mechanics and its Applications*, vol. 464, no. September, pp. 93–102, dec 2016.

[19] C. Bandt and B. Pompe, "Permutation Entropy: A Natural Complexity Measure for Time Series," *Physical Review Letters*, vol. 88, no. 17, p. 174102, 2002.

[20] L. Zunino, M. C. Soriano, and O. A. Rosso, "Distinguishing chaotic and stochastic dynamics from time series by using a multiscale symbolic approach," *Physical Review E*, vol. 86, no. 4, 2012.

[21] P. Lamberti, M. Martin, A. Plastino, and O. Rosso, "Intensive entropic non-triviality measure," *Physica A: Statistical Mechanics and its Applications*, vol. 334, no. 1-2, pp. 119–131, mar 2004.

[22] P. Barnaghi, W. Wang, L. Dong, and C. Wang, "A Linked-Data Model for Semantic Sensor Streams," in *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, 2013, pp. 468–475.

[23] C. Perera, A. Zaslavsky, C. H. Liu, M. Compton, P. Christen, and D. Georgakopoulos, "Sensor Search Techniques for Sensing as a Service Architecture for the Internet of Things," *IEEE Sensors Journal*, vol. 14, no. 2, pp. 406–420, feb 2014.

[24] Y. Qin, Q. Z. Sheng, and E. Curry, "Matching Over Linked Data Streams in the Internet of Things," *IEEE Internet Computing*, vol. 19, no. 3, pp. 21–27, 2015.

[25] S. Maurus and C. Plant, "Skinny-dip: Clustering in a Sea of Noise Samuel," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. New York, New York, USA: ACM Press, 2016, pp. 1055–1064.