Physica A 393 (2014) 222–234

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Macroscopic characterization of data sets by using the average absolute deviation

Diógenes Campos *,[1]

Faculty of Natural Sciences and Engineering, University ''Jorge Tadeo Lozano'', Bogotá, Colombia

HIGHLIGHTS

- A method for the macroscopic characterization of time series data is presented.
- Thermodynamics-like functions are used to measure macroscopic properties contained in a dataset.
- The chronological information available in the Vostok records is characterized from a thermodynamics-like point of view.

ARTICLE INFO

Article history:
Received 28 March 2013
Received in revised form 8 July 2013
Available online 11 September 2013

Keywords:
Average absolute deviation
Mean deviation
Thermodynamic description
Macroscopic description of data
Time series
Vostok ice cores

ABSTRACT

This paper describes a method for getting a *synthesis* of the knowledge about a given system, assuming that a data set x of measurements of a variable Ξ is known: i.e., separate data are combined in order to form a coherent whole, à la thermodynamics. For getting the macroscopic characterization of time series data, one takes advantage of the average absolute deviation concept together with an already known thermodynamic-like approach.

The method is applied to the macroscopic characterization of existing time series for the Vostok Antarctica ice cores (deuterium content of the ice and temperature variation) for a depth range between 0 and 3,310 meters, and the age of the ice between 0 and 422,766 years before the present.

1. Introduction

Consider a sequence of observations taken sequentially in time, i.e., a time series $x := \{x_1, x_2, \ldots, x_{\mathcal{N}}\}$, where x_n is the result of the measurement of a given variable Ξ at time t_n. The purpose of this paper is to propose a theoretical framework for the characterization of the sequence x by using the thermodynamics-like method presented in Refs. [1,2].

Different from the thermodynamics that has a phenomenological origin involving e.g. heat engines or the principle of Thomson and Clausius of the impossibility of perpetual motion machines of the second kind [3], the thermodynamics-like method in Refs. [1,2] is a purely theoretical formulation that is based on a quantitative measure of information developed by Hartley in 1928 and applied to communications problems [4]. Given an alphabet with s symbols and a message of length n, the amount of information contained in this message is equal to the amount of information of n messages of length one. This requirement is satisfied only by the logarithm function to base b (i.e., $\log_b$) and, therefore, the quantitative measure of information is $H_n := n \log_b(s) = \log_b(s^n)$. If $b = 2$ and $s = 2$, the unit of information is the bit, and $H_2 = \log_2(2^n)$ gives the bit-number [5].

* Tel.: +57 1 6194054.
E-mail addresses: dcamposr@cable.net.co, dcamposr@utadeo.edu.co.
URL: http://orcid.org/0000-0003-1990-3990.

[1] Member of the Colombian Academy of Sciences (ACCEFYN).

0378-4371/$ – see front matter
http://dx.doi.org/10.1016/j.physa.2013.09.013

The Hartley measure of information allows a probabilistic interpretation based on the uncertainty of an event n, which has probability P_n of occurring [5, Section 4.1]. The so-called Hartley formula $\mathcal{E}_n(P_n) := -\ln_b(P_n)$ gives a number which measures the amount of information, that one gains when one learns of the occurrence of this event [6,7]. In what follows the natural logarithm is used ($b = e$, $\ln := \ln_e$), and the unit of information is the nat.

In Ref. [2] a quantum mechanical system with Hamiltonian $\hat{H}$ and eigenvalue equation $\hat{H}|\phi_n\rangle = E_n|\phi_n\rangle$ is considered, and the following truncated canonical probability distribution is obtained:

$$P_n(\beta) = \frac{\omega_{\mathcal{N}}(P(\beta))}{\mathbb{Z}_{\mathcal{N}}(\beta)} \exp(-\beta E_n), \quad \text{for } n = 1, 2, 3, \ldots, \mathcal{N}. \tag{1}$$

Here, $\mathcal{N}$ states are taken into account ($\mathcal{N} < \infty$) and $\beta > 0$ is a characteristic parameter of the system that is associated with a "virtual temperature" scale $T := 1/(k_B\beta)$, where k_B is some conventional positive constant, such that $k_B T$ has dimensions of energy. If the system is in thermal equilibrium with a thermostat adjusted at the "real temperature" T, β can be interpreted as the thermodynamic $\beta = 1/(k_B T)$, where k_B is the Boltzmann constant, and T is the absolute temperature of the system (Kelvin degrees). The quantity $\omega_{\mathcal{N}} = \omega_{\mathcal{N}}(P(\beta)) := \sum_{m=1}^{\mathcal{N}} P_m(\beta)$ is the degree of completeness of the set $P(\beta) := \{P_n(\beta)|n = 1, 2, 3, \ldots, \mathcal{N}\}$, and $\mathbb{Z}_{\mathcal{N}}(\beta)$ is the partition function of the quantum statistical mechanics, i.e.,

$$\mathbb{Z}_{\mathcal{N}}(\beta) := \sum_{m=1}^{\mathcal{N}} \exp(-\beta E_m). \tag{2}$$

Thus, according to Eq. (1), given the energy eigenvalues $\{E_1, E_2, \ldots, E_{\mathcal{N}}\}$ and the parameter β, one can construct the truncated canonical probability distribution $P_n(\beta)$, for $n = 1, 2, 3, \ldots, \mathcal{N}$.

Once the probability distribution P is given, the Hartley information or pseudo-energy $\mathcal{E}_n(P(\beta)) := -\ln(P_n(\beta))$ provided by the observation of the n-th state becomes a basic quantity of the theory. The second central quantity in the theory is the escort probability set $\{p_n(P(\beta), q), \ n = 1, 2, \ldots, \mathcal{N}\}$, a concept created by Beck and Schögl [5] and later used by the nonextensive statistical mechanics [8, and references therein], where the entropic index q ($q \in [0, \infty)$) is an arbitrary real nonnegative parameter. Expressions (1) satisfy the important property that $p_n(P(\beta), q) = P_n(q\beta)/\omega_{\mathcal{N}}(P(q\beta))$, i.e., the escort probability $p_n(P(\beta), q)$ coincides with the real probability $P_n(q\beta)$, with the exception of the normalization factor $1/\omega_{\mathcal{N}}(P(q\beta))$ (see Eq. (11) in Ref. [2]). Consequently, the probability distribution given by Eq. (1) allows us to overcome the problems pointed out in the paper entitled "The role of constraints within generalized nonextensive statistics" [9,10].

At this point one has all the essential ingredients for evaluating "average values" of physical entities with respect to the escort probability set and, in this way, a macro-description of the system arises by taking advantage of some q-dependent thermodynamic-like functions [2]. The most relevant functions are classified as analogous to the familiar description of the standard quantum statistical mechanics (SQSM) or as new entities belonging to the modified quantum statistical mechanics (MQSM):

- SQSM. Average energy $\mathbb{U}(P(\beta), q)$, Shannon entropy $\mathbb{S}(P(\beta), q)$, Helmholtz free energy $\mathbb{F}(P(\beta), q)$, and heat-capacity $\mathbb{C}(P(\beta), q)$.
- MQSM. In this case, the corresponding quantities are named pseudo-…(e.g. pseudo-energy), and they are denoted as $\mathcal{U}(P(\beta), q)$, $\mathcal{S}(P(\beta), q)$, $\mathcal{F}(P(\beta), q)$, and $\mathcal{C}(P(\beta), q)$, respectively.

Whereas classical thermodynamics deals with systems consisting of a larger number of particles or subsystems (i.e., atoms or molecules), the method presented in Refs. [1,2], and also used in this paper, can also be applied to systems composed of a small number of elements. Thus, one can theorize about the possibility of associating to a data set x a probability distribution P. In this way, the method summarized in this section allows a macroscopic description of the data set X.

The structure of the paper is the following. Section 2 describes a method for assigning a probability distribution P to the sequence x composed by $\mathcal{N}$ nonvanishing elements x_n, for $n = 1, 2, \ldots, \mathcal{N}$. In Section 3, the method proposed in Refs. [1,2] is applied directly to P, and the basic thermodynamic-like functions required for the macroscopic description of the set x are established. At this stage, one takes advantage of the average absolute deviation of the set x.

In Section 4, the method is applied to the characterization of historical records obtained by researchers from France, Russia and the USA for the Vostok Antarctica ice cores [11–14]. The core is analyzed with a spatial distance of 1 m between the samples, and a time scale is linked to the depth. The data collected in Ref. [15] includes: (1) the depth range between 0 and 3,310 meters (m), (2) the age of the ice between 0 and 422,766 years before the present (years BP), (3) the deuterium content of the ice (δD) in ‰ SMOW (Standard Mean Ocean Sea Water), and (4) the isotope temperature record variation (ΔT, in °C) with reference to the mean recent time value. The variable x could be chosen as δD or ΔT, both of them characterized by time series with 3311 known values [15].

Finally, Section 5 concludes with some remarks.

2. Assembling an energy probability distribution

In this section, Eqs. (1) are considered from a point of view that differs from the quantum mechanical approach reviewed in Section 1. Assume that a hypothetical observer (i) knows the time series x and decides to describe the system by assigning