Article

Water Particles Monitoring in the Atacama Desert: SPC Approach Based on Proportional Data

Anderson Fonseca [1], Paulo Henrique Ferreira [1,*], Diego Carvalho do Nascimento [2], Rosemeire Fiaccone [1], Christopher Ulloa-Correa [3], Ayón García-Piña [3] and Francisco Louzada [4]

[1] Department of Statistics, Federal University of Bahia, Salvador 40170110, Brazil; andersonof@ufba.br (A.F.); fiaccone@ufba.br (R.F.)

[2] Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó 1530000, Chile; diego.nascimento@uda.cl

[3] Laboratorio de Investigación de la Criósfera y Aguas, IDICTEC, Universidad de Atacama, Copiapó 1530000, Chile; christopher.ulloa@uda.cl (C.U.-C.); ayon.garcia@uda.cl (A.G.-P.)

[4] Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos 13566590, Brazil; louzada@icmc.usp.br

* Correspondence: paulohenri@ufba.br

Abstract: Statistical monitoring tools are well established in the literature, creating organizational cultures such as Six Sigma or Total Quality Management. Nevertheless, most of this literature is based on the normality assumption, e.g., based on the law of large numbers, and brings limitations towards truncated processes as open questions in this field. This work was motivated by the register of elements related to the water particles monitoring (relative humidity), an important source of moisture for the Copiapó watershed, and the Atacama region of Chile (the Atacama Desert), and presenting high asymmetry for rates and proportions data. This paper proposes a new control chart for interval data about rates and proportions (symbolic interval data) when they are not results of a Bernoulli process. The unit-Lindley distribution has many interesting properties, such as having only one parameter, from which we develop the unit-Lindley chart for both classical and symbolic data. The performance of the proposed control chart is analyzed using the average run length (ARL), median run length (MRL), and standard deviation of the run length (SDRL) metrics calculated through an extensive Monte Carlo simulation study. Results from the real data applications reveal the tool's potential to be adopted to estimate the control limits in a Statistical Process Control (SPC) framework.

Keywords: Symbolic Data Analysis (SDA) in Statistical Process Control (SPC); rates and proportions data; unit-Lindley distribution; relative air humidity monitoring; Monte Carlo simulation

1. Introduction

Control charts are often applied to monitor processes in many fields, including ecology [1], health [2] and industry [3]. This primary SPC tool is used for various types of data, such as count [4], attributes [5] and rates/proportions [6]. The latter is a widespread type of data, having applications in the most diverse areas, for example, climate [7] and industry [8]. The most widely used control charts for this situation are the p and np charts, but to use them, the process needs to be completed from Bernoulli experiments [9].

However, there are cases (i.e., processes) where rates and proportions are not results from Bernoulli experiments (e.g., come from individual measures or continuous number ratio), despite assuming values in the range $(0, 1)$. For these processes, the p and np charts are not applicable [7,10]. Therefore, it seeks alternatives, some of which are the beta [11], Kumaraswamy [7], simplex and unit-gamma [10] charts for monitoring fraction data, although not suitable in all cases (e.g., overdispersion). Thus, this article will propose an interesting and useful alternative to the previously mentioned control charts for monitoring processes

with continuous data in the unit interval (e.g., rates, proportions or indices), the so-called unit-Lindley chart. Other works, such as [12–15], explored the unit transformation and showed the importance of the unit distributions class. Moreover, stochastic phenomena can be approximated, with a clear advantage and simplicity grounds, by summarizing a large quantity of data by a few numerical values, in the form of parametric models, enabling the modeller to learn about observable phenomena [16].

Recently, [17] introduced a one-parameter continuous probability distribution defined on the range $(0, 1)$, which was named the unit-Lindley distribution. Such a model can describe processes involving data on rates and proportions. It also features many attractive properties, for example, a single parameter distribution (thus, more straightforward than the previously mentioned two-parameter distributions, namely the beta, Kumaraswamy, simplex, and unit-gamma models), a convenient reparameterization of the mean and closed-form expression for the maximum likelihood estimator. Thus, the unit-Lindley distribution motivated us to develop a new statistical control chart to monitor rates and proportions. It is worth pointing out that the main aim of the proposed unit-Lindley chart is to detect significant shifts in the process parameter (mean).

Our practical motivation came from monitoring the relative humidity in arid conditions of the Atacama Desert, which presents extreme weather with a small precipitation rate and high-temperature variation in the day regardless of the season. Three main climatic features of this region can be highlighted: (i) the (cold) Humboldt Current in the Pacific Ocean at the west; (ii) two mountain ranges: the Coastal range to the west, a part of blocking the moisture coming from the Pacific Ocean, and the long and high Andes mountain range to the east, which blocks the moisture coming from the Amazon Basin or the Atlantic Ocean; and (iii) its localization in the air convection cells close to the tropic of Capricorn (southern or tropical limit of the southern Hadley Cell, characterized by dry air).

Despite this arid environment, there is a particular phenomenon that will allow the formation of the *Camanchaca* in some area (in the cities of Copiapó and Huasco): morning and night sea mist (or fog), which will advance inland and penetrate deeply into the valleys and can be a source of water [18] in a region that suffers scarcity of this resource. In a simplified way, marine stratocumulus form over the Pacific Ocean, presenting, on the Chilean coast, cold temperatures linked to the Humboldt Current [19]. This humid air is limited in its convection by a low thermal inversion linked to a warm and dry air linked to the tropical component of the Hadley Cell and located higher up than the colder air close to the sea surface. Such phenomenon is characteristic of the morning and night, as the temperatures are lower during that time and pass more easily below the dew point, which corresponds to the condensation point of the water contained in the atmosphere. Then, the clouds formed at low altitude can then, under the trade winds, enter deeply into the land following the west-east oriented valleys.

A classical data representation is infeasible, requiring a symbolic data representation, set-valued (interval or multi-valued) or modal (weight or probability distribution), containing more complex information of the phenomenon (humidity in hyper-arid conditions). These variables are called "symbolic", account for variability or uncertainty, making the symbolic data analysis more comprehensive than classical data analyses [20,21], and shall be extended for SPC reasoning. Thus, this paper proposes a new statistical methodology to overcome the complexity of monitoring the humidity in Copiapó city, Chile, due it is located in the Atacama Desert and most days present the *Camanchaca* phenomenon.

The remainder of this paper is organized as follows. Section 2 describes the practical motivation and the data acquisition. In Section 3, we first revise the unit-Lindley distribution and some of its basic properties (Section 3.1), and then present the new control chart based on this distribution (Section 3.2). Section 4 provides simulation studies designed to assess the performance of the proposed unit-Lindley chart. Section 5 illustrates the usefulness of the unit-Lindley chart through several examples. Finally, Section 6 concludes the paper with a few remarks and discussions on future works.

2. The Data

In the Atacama Desert, a north-south geographic band located mainly in northern Chile, precipitation is only a few millimeters per year or sometimes non-existent, making it one of the driest places on Earth [22,23]. However, the vast expanses of the desert are punctuated by fertile valleys with rivers originating in the central Andes and flowing into the Pacific Ocean. Along these rivers, human populations have settled historically, exploiting more and more, this rare and precious water, especially with the growing development of the monoculture and mining industry, logically interested in the special weather conditions and the mineral resources of the Cordillera.

The hydrological regime of the principal rivers of Atacama is characterized by ice sources: water flows from the peaks following the melting of snowfall, glaciers and permafrost located in the upper parts of the Andes range [24,25]. In the context of climate change, it is therefore essential to understand the hydrological cycle of these regions in order to set up a sustainable management policy [26]. Understand the hydrological cycle requires the implementation of tools for forecasting river flows, relative humidity, groundwater reservoirs or any other water-related quantity monitoring, which inevitably needs an in-depth knowledge of the physical phenomena that govern the entire hydrological cycle and, more precisely, the complex interaction between atmosphere, climate, landforms, ice, snow and river flows.

Whereas, an important event occurs in this area, where marine stratocumulus cloud banks that form on the Chilean coast, called *Camanchaca*, which is daily the passageway of "low clouds", right after sunrise, sequentially for a couple of hours. This event is the source of water for many types of flora and fauna in the Atacama desert.

For illustrating the relative humidity dynamics in the Atacama, were processed satellite images acquired using MODIS (Moderate Resolution Imaging Spectroradiometer) sensor from 2000 to 2020, this sensor is available in two satellites, these are Terra (daily around 11:00 a.m. local time), as well as Aqua (around 3:00 p.m. local time). Figure 1 shows the statistical representation of the cloud occurrence over the Atacama region (Terra MODIS in panel B, and Aqua MODIS in panel C, adopting the same colour scale), shaded in dark-red for a high probability of cloud occurrence, and in beige for low probability. At the same time, a high observation of cloud (moisture source) is noticeable from the Chilean coast till the beginning of the highlands (this limit is evident in yellow), shown through the digital elevation model for part of the Copiapó watershed presented in panel A.

Regarding the *Camanchaca* event, when monitoring the humidity in the Atacama region of Chile, moreover, in Copiapó city, it is relevant to know that water is a scarce element (as liquid or vapour). Nonetheless, water vapour flows through the city in the mornings, almost daily, caused by the dominant winds coming from the west.

The weather station from which the relative humidity data used in this study are coming from is located on the campus of the University of Atacama (27.359 S/70.353 W) and belongs to the Chilean Meteorological Directorate (DMC). In addition to relative humidity, the data measured are atmospheric pressure, temperature, global solar radiation, precipitation, and wind direction and speed, the last both are installed at 10 meters above ground level. These data covering a period from 2016 to 2021 (at the time of writing this article) with, at best, a one-minute recording format. The relative humidity is recorded thanks to a Vaisala HMP155A-L17-PT probe protected from solar radiation and including an air filter with a Teflon membrane installed at 2 meters above ground level. The weather stations managed by the DMC are part of a Chilean national weather data network.

Relative humidity monitoring done for the meteorological purpose could also be used to assess the potential of the *Camanchaca* as a source of water in a region that suffers substantial scarcity of this essential resource.
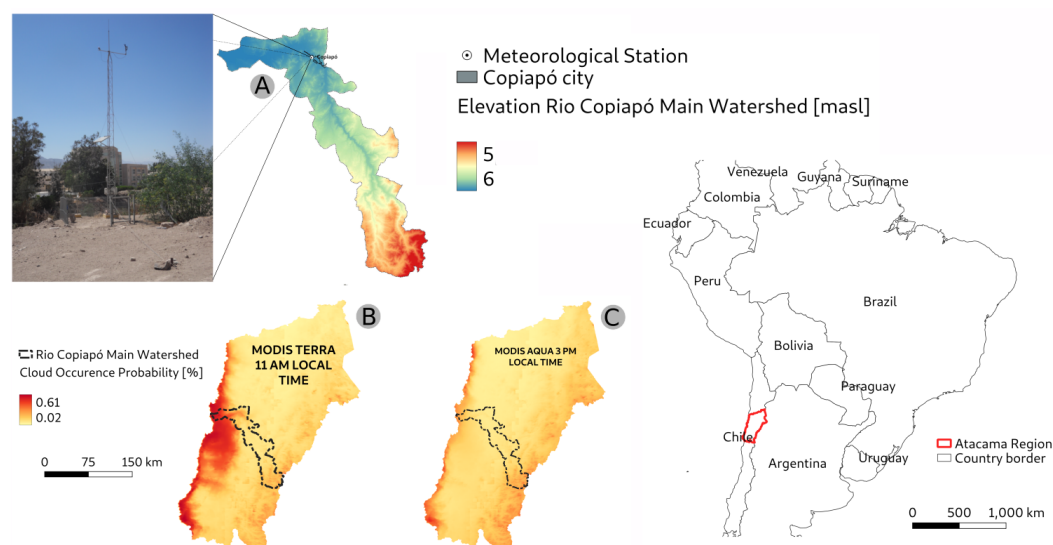
Figure 1. Statistics of the cloud occurrence over the Atacama region (panels B and C), shaded from high probability of cloud occurrence in dark-red to low probability of cloud occurrence in beige. The bottom left-hand map represents Terra MODIS (panel B), and Aqua MODIS (panel C) is in the bottom center map. The transaction of the dark-red area, which occurs mainly during the dawn and morning and is most associated with the *Camanchaca* increasing the humidity of the Chilean third region coast up to the beginning of the highlands, turns into the low scale of humidity right in the afternoons, represented by the full red map. Two bays are to be noticeable: in Copiapó and Huasco, as convergence points. In panel A, the digital elevation model for part of the Copiapó watershed is presented. Our data acquisition came from a weather station located at the University of Atacama in Copiapó, Chile (in the top left-hand picture).

3. Methodology

In this section, we first revise the unit-Lindley distribution and some of its basic properties (Section 3.1). Then, we present a new control chart based on this distribution (Section 3.2).

Figure 2 summarizes visually the adopted methodology, where we first transform the data set records from hourly into four periods per day (considering their minimum and maximum values). That is, the symbolic representation is given by transforming a time window of every 6 h observation points into a pairwise (MIN, MAX) representation. Then, we show the proposed new statistical methodology, which will be adopted as a symbolic interval data approach, through the unit-Lindley distribution and the control chart based on it. In this way, we intend to contribute to the understanding of the complexity of monitoring the humidity in Copiapó city.
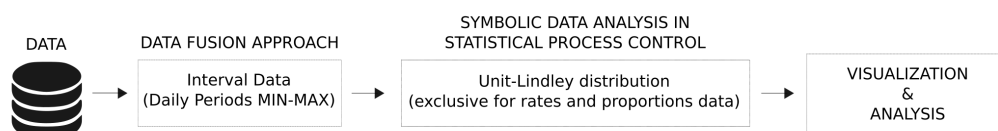
Figure 2. Visual representation of the adopted methodology.

3.1. The Unit-Lindley Distribution

Introduced by [17], a random variable Y is said to be unit-Lindley distributed with parameter $\theta > 0$, denoted by $Y \sim \mathrm{UL}(\theta)$, if its cumulative distribution function (CDF) is given by

$$F(y;\theta) = 1 - \left(1 - \frac{\theta y}{(1+\theta)(y-1)}\right) \exp\left\{-\frac{\theta y}{1-y}\right\}, \qquad \text{for } 0 < y < 1.$$

The corresponding probability density function (PDF) is

$$f(y; \theta) = \frac{\theta^2}{1+\theta} (1-y)^{-3} \exp\left\{-\frac{\theta y}{1-y}\right\}, \qquad \text{for } 0 < y < 1,$$

which is unimodal with maximum at $Y_{\max} = 1 - \theta/3$ for $\theta < 3$, and $Y_{\max} = 0$ for $\theta \geq 3$.

If $Y \sim \mathrm{UL}(\theta)$, then the mean and variance of Y are given, respectively, by

$$E[Y] = \frac{1}{1+\theta} \qquad \text{and} \qquad Var[Y] = \frac{1}{1+\theta}\left(\theta^2 e^\theta Ei(1, \theta) - \theta + 1\right) - \left(\frac{1}{1+\theta}\right)^2,$$

where $Ei(a, z) = \int_1^\infty x^{-a} e^{-xz}\, dx$ is the exponential integral function [27], which can be computed using the $expint(\cdot)$ function of the `expint` package [28] in R.

The quantile function, $Q(p; \theta) = F^{-1}(p; \theta)$, can be written as

$$Q(p; \theta) = \frac{1 + \theta + W_{-1}((1+\theta)(p-1)\exp\{-(1+\theta)\})}{1 + W_{-1}((1+\theta)(p-1)\exp\{-(1+\theta)\})}, \qquad \text{for } 0 < p < 1,$$

where W_{-1} denotes the negative branch of the Lambert W function [29], which can be computed via the $lambertWn(\cdot)$ function of the `pracma` package [30] in R.

We can easily estimate the parameter θ using the maximum likelihood method. By considering the observed random sample $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ of size n from $Y \sim \mathrm{UL}(\theta)$, we obtain the likelihood function

$$L(\theta; \mathbf{y}) \propto \left(\frac{\theta^2}{1+\theta}\right)^n e^{-\theta\, t(\mathbf{y})},$$

where $t(\mathbf{y}) = \sum_{i=1}^n \frac{y_i}{1-y_i}$. Mazucheli et al. [17] showed that the maximum likelihood estimator (MLE) $\hat{\theta}$ of θ has a closed-form expression and is given by

$$\hat{\theta} = \frac{1}{2\, t(\mathbf{y})}\left(n - t(\mathbf{y}) + \sqrt{[t(\mathbf{y})]^2 + 6\, n\, t(\mathbf{y}) + n^2}\right).$$

In order to achieve substantial bias reduction, especially for small and moderate sample sizes, the authors derived a bias-corrected MLE $\tilde{\theta}$ of θ through the methodology proposed by Cox and Snell [31]. This estimator is given by

$$\tilde{\theta} = \hat{\theta} - \frac{\hat{\theta}^5 + 7\,\hat{\theta}^4 + 12\,\hat{\theta}^3 + 8\,\hat{\theta}^2 + 2\,\hat{\theta}}{n\left(\hat{\theta}^2 + 4\,\hat{\theta} + 2\right)^2}.$$

Mazucheli et al. [17] also presented an alternative and useful reparameterization of the unit-Lindley distribution, where $\mu = E[Y] = 1/(1+\theta)$ and, thus, $\theta = 1/\mu - 1$. In this case, the CDF and PDF of $Y \sim \mathrm{UL}(\mu)$, $0 < \mu < 1$, are written, respectively, as

$$F(y; \mu) = 1 - \left(1 - \frac{y(1-\mu)}{y-1}\right)\exp\left\{-\frac{y(1-\mu)}{\mu(1-y)}\right\}, \qquad \text{for } 0 < y < 1,$$

and

$$f(y; \mu) = \frac{(1-\mu)^2}{\mu(1-y)^3}\exp\left\{-\frac{y(1-\mu)}{\mu(1-y)}\right\}, \qquad \text{for } 0 < y < 1. \tag{1}$$

The mean and variance of the reparameterized unit-Lindley distribution are given, respectively, by

$$E[Y] = \mu \qquad \text{and}$$

$$Var[Y] = \mu\left[\left(\frac{1}{\mu} - 1\right)^2 \exp\left\{\frac{1}{\mu} - 1\right\} Ei\left(1, \left(\frac{1}{\mu} - 1\right)\right) - \frac{1}{\mu} + 2\right] - \mu^2.$$

The quantile function, $Q(p; \mu) = F^{-1}(p; \mu)$, can be written as

$$Q(p; \mu) = \frac{\frac{1}{\mu} + W_{-1}\left(\frac{(p-1)}{\mu} \exp\left\{-\frac{1}{\mu}\right\}\right)}{1 + W_{-1}\left(\frac{(p-1)}{\mu} \exp\left\{-\frac{1}{\mu}\right\}\right)}, \qquad \text{for } 0 < p < 1. \tag{2}$$

Furthermore, the authors showed that the MLE $\hat{\mu}$ of μ is given by

$$\hat{\mu} = -\frac{1}{2n}\left(n + t(\boldsymbol{y}) - \sqrt{[t(\boldsymbol{y})]^2 + 6\,n\,t(\boldsymbol{y}) + n^2}\right)$$

and the corresponding bias-corrected MLE $\tilde{\mu}$ of μ is

$$\tilde{\mu} = \hat{\mu} - \frac{2\hat{\mu}^2(2\hat{\mu} - 2)}{n(\hat{\mu}^2 - 2\hat{\mu} - 1)^2}. \tag{3}$$

It is worth noting that, in the original work of [17], there was a typo in the $\hat{\mu}$ expression, with $t(\boldsymbol{y})$ instead of n in the denominator.

Due to its simplicity and better interpretability, which makes it arguably a more appealing model to use in practice, we shall hereafter consider this reparameterized version of the unit-Lindley distribution, as well as the control chart limits expressed in terms of the mean parameter μ.

3.2. Proposed Unit-Lindley Chart

Suppose that a process (e.g., a hydrological or environmental process) generates data (e.g., rates or proportions) according to a unit-Lindley distribution. That is, if Y denotes the monitored variable, then the PDF of Y is given by (1). Also, consider that the probability of false alarm (or type I error) is α. Thus, we have

$$P(Y < \text{LCL} \mid \mu) = P(Y > \text{UCL} \mid \mu) = \alpha/2,$$

where μ is the in-control process parameter (that is, the mean value of the quality characteristic based on the in-control state), LCL and UCL are the lower and upper control chart limits, respectively.

Following [32], the control limits and centerline (CL) of the proposed unit-Lindley chart are given by

$$\text{LCL} = Q(\alpha/2; \mu), \qquad \text{CL} = \mu, \qquad \text{UCL} = Q(1 - \alpha/2; \mu), \tag{4}$$

where $Q(.)$ is the quantile function presented in (2).

Table 1 shows these control limits for several values of μ, considering $\alpha = 0.1$, 0.01 and 0.0027. Note that the latter α value corresponds to the standard three-sigma rule (Six Sigma program).

When the parameter μ is unknown/unspecified, it can be estimated using all the available observations (data). Thus, we can replace μ by, e.g., its bias-corrected MLE (3) in the expressions of the control limits shown in (4), obtaining the so-called "trial control limits" [9].

Table 1. Control limits of the unit-Lindley chart for some values of μ and α.

μ	$\alpha = 0.1$			$\alpha = 0.01$			$\alpha = 0.0027$		
	LCL	CL	UCL	LCL	CL	UCL	LCL	CL	UCL
0.08	0.0048	0.08	0.2190	0.0005	0.08	0.3303	0.0001	0.08	0.3802
0.12	0.0079	0.12	0.3124	0.0008	0.12	0.4428	0.0002	0.12	0.4965
0.16	0.0115	0.16	0.3954	0.0011	0.16	0.5320	0.0003	0.16	0.5846
0.20	0.0158	0.20	0.4688	0.0016	0.20	0.6038	0.0004	0.20	0.6530
0.24	0.0208	0.24	0.5335	0.0021	0.24	0.6623	0.0006	0.24	0.7072
0.28	0.0269	0.28	0.5905	0.0027	0.28	0.7107	0.0007	0.28	0.7512
0.32	0.0341	0.32	0.6407	0.0035	0.32	0.7512	0.0009	0.32	0.7873
0.36	0.0428	0.36	0.6851	0.0044	0.36	0.7854	0.0012	0.36	0.8174
0.40	0.0534	0.40	0.7244	0.0055	0.40	0.8146	0.0015	0.40	0.8429
0.44	0.0662	0.44	0.7592	0.0070	0.44	0.8397	0.0019	0.44	0.8646
0.48	0.0819	0.48	0.7902	0.0088	0.48	0.8616	0.0024	0.48	0.8834
0.52	0.1012	0.52	0.8179	0.0112	0.52	0.8807	0.0030	0.52	0.8997
0.56	0.1250	0.56	0.8426	0.0142	0.56	0.8975	0.0039	0.56	0.9139
0.60	0.1545	0.60	0.8648	0.0183	0.60	0.9124	0.0050	0.60	0.9265
0.64	0.1912	0.64	0.8848	0.0240	0.64	0.9256	0.0066	0.64	0.9377
0.68	0.2366	0.68	0.9029	0.0319	0.68	0.9375	0.0089	0.68	0.9477
0.72	0.2927	0.72	0.9193	0.0433	0.72	0.9481	0.0122	0.72	0.9566
0.76	0.3612	0.76	0.9341	0.0607	0.76	0.9578	0.0174	0.76	0.9647
0.80	0.4433	0.80	0.9477	0.0881	0.80	0.9665	0.0260	0.80	0.9720
0.84	0.5393	0.84	0.9600	0.1339	0.84	0.9744	0.0417	0.84	0.9786
0.88	0.6475	0.88	0.9713	0.2151	0.88	0.9817	0.0739	0.88	0.9847
0.92	0.7642	0.92	0.9817	0.3645	0.92	0.9883	0.1522	0.92	0.9902

4. Statistical Performance

In this section, we use Monte Carlo (MC) simulation studies to evaluate the unit-Lindley chart's statistical performance measured in terms of the average run length (ARL), median run length (MRL), and standard deviation of the run length (SDRL). All computational routines were implemented using the R software version 3.6.3 [33].

The ARL is a metric widely used to evaluate the efficiency of control charts. The in-control ARL (or ARL_0) is defined as the average number of observations (or monitoring points) before a signal is given (that is, a single point falls outside the control limits), assuming that the process is in control. In contrast, the out-of-control ARL (or ARL_1) is the average number of observations that are taken until a mean shift is identified when the process is out of control [34].

Since the run length (RL) is high asymmetrically distributed, other metrics than the ARL, such as the SDRL and MRL, can be considered [10]. The SDRL is a useful measure used to assess the spread (or dispersion) of the RL distribution, whereas the MRL refers to the midpoint of the RL distribution and is a more credible measure of a chart's performance since it is less affected by the skewness of the RL distribution [35]. In addition, we will also use the in-control ($SDRL_0$ and MRL_0) and out-of-control ($SDRL_1$ and MRL_1) versions of these metrics.

Let Y be the result or output of a process that follows a unit-Lindley distribution reparameterized by its mean: $Y \sim UL(\mu)$. Also, let μ_s be the shifted mean proportion parameter after a change occurs in μ, that is, $Y \sim UL(\mu_s)$.

For the proposed unit-Lindley chart, the in-control ARL, SDRL and MRL are defined as

$$ARL_0 = 1/\alpha, \qquad SDRL_0 = \sqrt{(1-\alpha)/\alpha^2}, \qquad MRL_0 = \log(0.5)/\log(1-\alpha),$$

for $\alpha = 1 - P(LCL < Y < UCL \mid \mu)$. While the out-of-control metrics are given by

$$ARL_1 = 1/(1-\beta), \qquad SDRL_1 = \sqrt{\beta/(1-\beta)^2}, \qquad MRL_1 = \log(0.5)/\log(\beta),$$

for $\beta = P(\text{LCL} < Y < \text{UCL} \mid \mu_s)$.

Moreover, we will also use the down and up versions of the in-control ARL, MRL and SDRL, which consider the occurrence of false alarms at the LCL (i.e., sample points falling below the lower limit) and UCL (i.e., sample points falling above the upper limit), respectively. These metrics are given by

$$\text{ARL}_0^{\text{down}} = \text{ARL}_0^{\text{up}} = \frac{1}{\alpha/2},$$

$$\text{SDRL}_0^{\text{down}} = \text{SDRL}_0^{\text{up}} = \sqrt{\frac{1 - \alpha/2}{(\alpha/2)^2}},$$

$$\text{MRL}_0^{\text{down}} = \text{MRL}_0^{\text{up}} = \frac{\log(0.5)}{\log(1 - \alpha/2)}.$$

In the usual Six Sigma program, $\alpha = 0.0027$ and, therefore, $\text{ARL}_0 = 1/0.0027 \approx 370$, $\text{SDRL}_0 = \sqrt{(1 - 0.0027)/0.0027^2} \approx 370$, and $\text{MRL}_0 = \log(0.5)/\log(1 - 0.0027) \approx 256$. This means, e.g., for the first measure, that even though the process is in control, an incorrect out-of-control signal (or false alarm) will be generated every 370 samples, on the average [9]. On the other hand, values of $\text{ARL}_1 \approx 1$ are desired, mainly for large-size shifts in the process mean parameter.

4.1. In-Control Processes

Without loss of generality, in this subsection we consider unit-Lindley processes with mean parameter: $\mu = 0.2$, 0.5 and 0.8 (whose PDF plots are shown in Figure 3), as well as two distinct values for the probability of false alarm: $\alpha = 0.1$ (which corresponds to $\text{ARL}_0 = 10$, $\text{SDRL}_0 \approx 9.487$, $\text{MRL}_0 \approx 6.579$, $\text{ARL}_0^{\text{down}} = \text{ARL}_0^{\text{up}} = 20$, $\text{SDRL}_0^{\text{down}} = \text{SDRL}_0^{\text{up}} \approx 19.494$, and $\text{MRL}_0^{\text{down}} = \text{MRL}_0^{\text{up}} \approx 13.513$) and 0.01 (which corresponds to $\text{ARL}_0 = 100$, $\text{SDRL}_0 \approx 99.499$, $\text{MRL}_0 \approx 68.968$, $\text{ARL}_0^{\text{down}} = \text{ARL}_0^{\text{up}} = 200$, $\text{SDRL}_0^{\text{down}} = \text{SDRL}_0^{\text{up}} \approx 199.499$, and $\text{MRL}_0^{\text{down}} = \text{MRL}_0^{\text{up}} \approx 138.283$). We also use different sample sizes (i.e., different numbers of Phase 1 observations) for each process: $n = 10$, 30, 50, 100 and 200. According to [9], in the Phase I study (or retrospective analysis), a process data set is collected and analyzed at once, building trial control limits to decide whether the process was under control when the first n observations were gathered. Whereas in the Phase II study (prospective analysis or process monitoring), the control chart constructed from a process "clean" data set showing control (reliable control limits) is used for monitoring future production.

The results obtained from 5000 MC simulations (or replicates) with $n^* = 5000$ Phase 2 observations each, performed for each scenario studied (that is, by varying the number of Phase 1 observations, the mean parameter of the unit-Lindley distribution, and the probability of false alarm), and further information regarding these results, please contact the correspondence author. Despite some slight to moderate discrepancies between the theoretical (target) values of the performance measures, and the values calculated through MC simulations in some cases, which are expected due to the effect of parameter estimation on control chart properties (see, e.g., [36–38]), the obtained results seem to indicate the good performance of the proposed control chart.
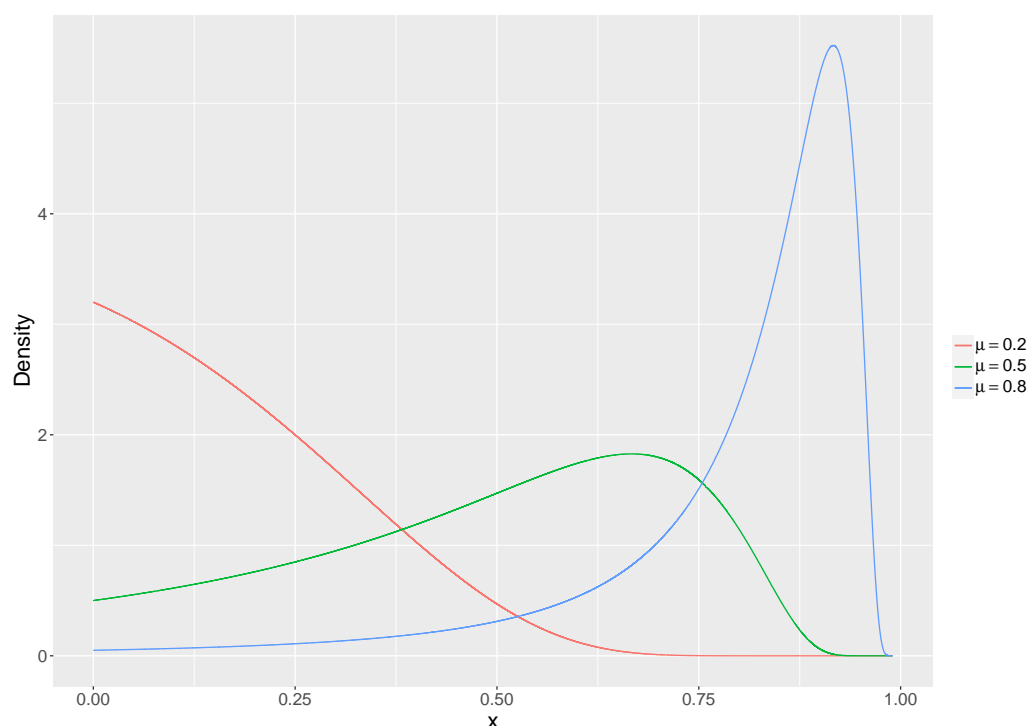
Figure 3. Unit-Lindley density function for the different parameter values considered in this study.

4.2. Out-of-Control Processes

In this subsection, we assess the shift-detection ability of the proposed unit-Lindley chart in terms of ARL_1, $SDRL_1$ and MRL_1, for the same scenarios as before. We consider shifts at different levels, representing percentage decreases and increases p in the process mean parameter μ. The assumed levels are: $p = 1\%$ (down-shifted mean: $\mu_s = 0.198$, 0.495 and 0.792; up-shifted mean: $\mu_s = 0.202$, 0.505 and 0.808), 10% (down-shifted mean: $\mu_s = 0.18$, 0.45 and 0.72; up-shifted mean: $\mu_s = 0.22$, 0.55 and 0.88) and 20% (down-shifted mean: $\mu_s = 0.16$, 0.4 and 0.64; up-shifted mean: $\mu_s = 0.24$, 0.6 and 0.96).

The results obtained from 5000 MC simulations with $n^* = 5000$ Phase 2 observations. Note that the values of the performance measures fall faster the higher the mean, especially with increases. With the 1% change, there is practically no difference in the metrics. At 10%, one can already see a significant drop, which is more robust for $\mu = 0.8$. With 20%, the estimated values are closer to one.

4.3. Comparison with Some Standard Control Charts

The unit-Lindley control chart theory introduced here may be applied in practical situations as a valuable and exciting alternative to, e.g., the well-known beta [11], simplex [10] and Kumaraswamy [7] charts, when the process data are continuous in the interval $(0, 1)$. Thus, in-control proportion/rate/index data can be modelled well via the proposed unit-Lindley chart.

In this subsection, we apply the four above-mentioned control charts to sample data generated from the unit-Lindley, beta, simplex [39] and Kumaraswamy [40] distributions. The aim is to investigate, through simulations, the performance (in terms of the same in-control and out-of-control metrics used before) of these control charts when they are applied to process data that come from different distributions defined on the range $(0, 1)$.

The simulations were carried out using the same settings as described in the previous subsections. Nevertheless, we consider only $n = 200$, $n^* = 5000$ and $\alpha = 0.1$. For the true data-generating process (which can be unit-Lindley, beta, simplex or Kumaraswamy distributed), we set the (in-control) mean parameter: $\mu = 0.2$ (case 1), 0.5 (case 2) and 0.8 (case 3). It can be seen, among others, that the proposed unit-Lindley chart has a good

performance in all scenarios, despite being based on a distribution with a single parameter (and, thus, more straightforward than the other two-parameter distributions).

5. Application

In this section, we apply the novel unit-Lindley chart to real data on the relative humidity of the air in the city of Copiapó, Chile. Located in the Atacama Desert, this important northern Chilean city has 16,681.3 square kilometers and 158,438 inhabitants [41]. Copiapó's economy is based on mineral exploitation and agriculture, which demands a significant volume of water in both activities.

The acquired data set is from the Copiapó Station, located at the University of Atacama. Data were obtained hourly from 21 December 2021 (day/month/year) to 2 February 2021 ($n = 35{,}047$ records). Figure 4 shows the dynamic of the data set, where, through the bimodality of the histogram, at first sight, it is noticeable that at least two phenomena are happening at once.
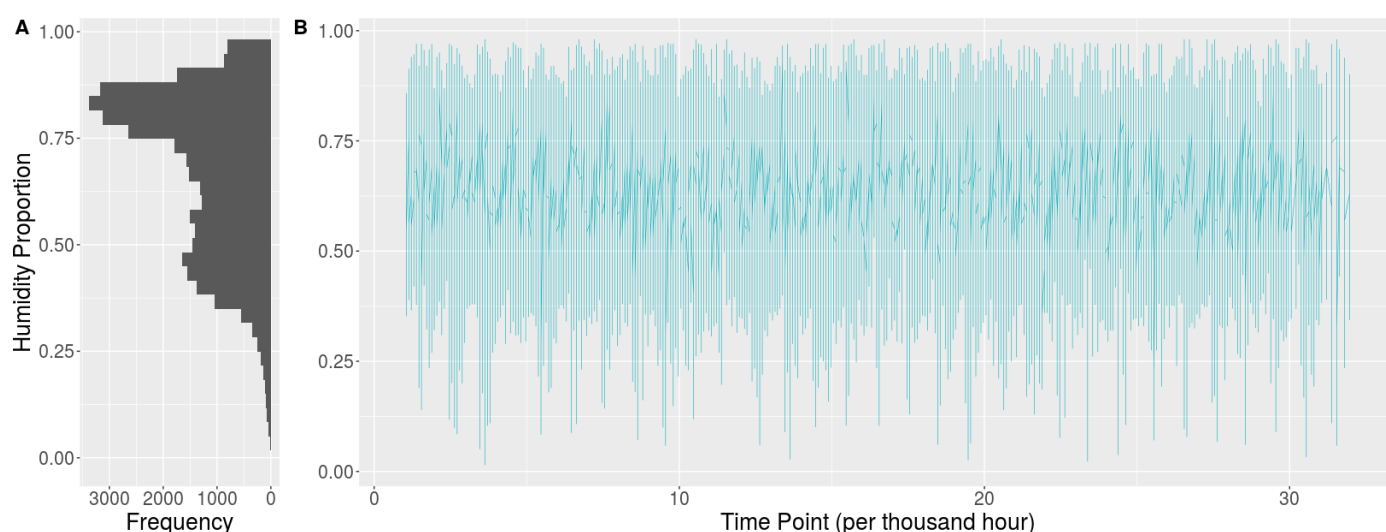
Figure 4. Humidity variation, collected per hour, from Copiapó (Chile) in the last five years. Panel A presents the histogram of the time series (TS), and panel B shows the dynamic of this series in light blue. Also, the solid line represents the TS average using a LOESS (an acronym for "Locally Estimated Scatterplot Smoothing") smoothing method [42].

For many real applications, finding hidden structures or properties in complex data can be computationally costly or problematic due to the messy sets, latent patterns, and noisily signals. Symbolic Data Analysis (SDA) is a paradigm of Machine Learning and Statistics areas aiming to build, describe, analyze and extract new knowledge from more complex data structures. The SDA intends to study and propose methodologies that can handle more complex data (or symbolic data), such as intervals, sets or histograms, in order to consider variability and uncertainty that is often inherent to the data [43–47]. SDA starts summarizing massive classical data and describes the new units of more minor and smoother data sets by symbolic variables.

Nascimento et al. [21] summarized a database containing more than 1.5 billion observations that are signals of a multi-channel electroencephalogram (EEG) corresponding to a frequency over time. A dynamic linear model for EEG interval data was introduced and applied to a database which, after being condensed, resulted in 15,000,000 total observations (that is, only 0.98% of the size of the original data set). This kind of model can incorporate dynamic events regarding more complex data, showing a competitive alternative to modelling a problem, given its flexibility and speed in data convergence. Moreover, questions of dynamic accommodation in neuroscience research could be resolved to reveal brain activity patterns.

Thus, to adopt the SDA towards the meteorological data, this work uses them as interval data, minimum-maximum (MIN-MAX) period, holding a physical interpretation regardless of the data compression. We aim first as a data fusion step to transform the data granularity by taking the mean of every 6 h into daily periods: day part 1 (from midnight to 5:59 a.m. UTC), day part 2 (from 6:00 a.m. to 11:59 a.m. UTC), day part 3 (from midday to 5:59 p.m. UTC), and day part 4 (from 6:00 p.m. to 11:59 p.m. UTC). This is in order to reduce the data noise (justified in terms of increasing the signal-to-noise ratio). So, the SDA representation transformed 35,047 original records into 11,736 new ones (5868 observations for the MIN period, and 5868 observations for the MAX period). Table 2 outlines the statistical descriptions per year and highlights in bold the highest values per statistic.

Table 2. Statistical summary description, per year, highlighting in bold the highest values per statistic.

		Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
Minimum	2016	**0.33**	0.398	0.525	0.544	0.66	0.81
	2017	0.10	**0.43**	**0.58**	**0.581**	0.74	**0.98**
	2018	0.072	0.418	0.57	0.575	**0.741**	0.965
	2019	0.015	0.408	0.557	0.563	0.722	0.963
	2020	0.059	0.413	0.571	0.571	0.74	0.957
	2021	0.295	0.39	0.567	0.559	0.731	0.873
Maximum	2016	**0.50**	0.61	0.77	0.726	0.81	0.89
	2017	0.24	**0.69**	0.81	**0.774**	0.87	**0.98**
	2018	0.182	0.674	**0.816**	0.771	**0.878**	0.973
	2019	0.079	0.652	0.808	0.755	0.866	0.972
	2020	0.216	0.668	0.815	0.77	0.873	0.973
	2021	0.449	0.686	0.805	0.753	0.847	0.958

Figure 5 shows the histograms of the acquired minimum (left-hand panel) and maximum (right-hand panel) daily period of humidity, where the solid black line represents the unit-Lindley PDF adjusted for these data. It is noticeable the asymmetry present in these data and the nonexistence of a whole period raining (that is, 6 h of raining, which would imply observation in the minimum relative humidity equal to one).
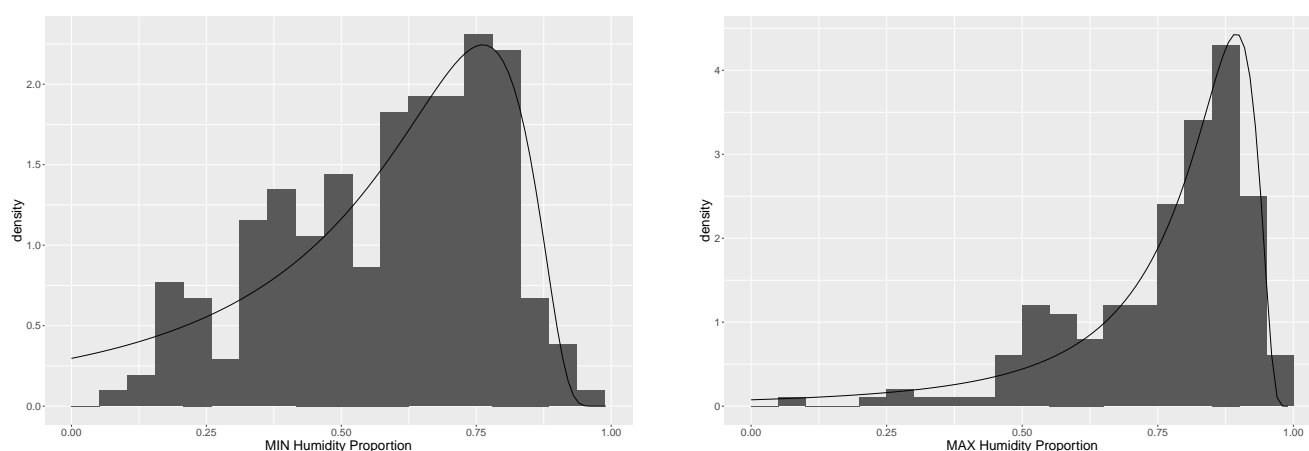
Figure 5. Histogram of the minimum (**left-hand panel**) and maximum (**right-hand panel**) observations, after aggregating the daily humidity representation of the data in day periods (parts 1–4), followed by a black solid line representing the estimated PDF of the unit-Lindley distribution.

As the first SPC analysis, we devoted the records from December 2016 to December 2020 for the Phase 1 study (process parameter estimation/assessment of process stability/control limits establishment). After that, we reserved all the records of 2021 for doing the Phase 2 study (online process monitoring). In Phase 1, we randomly selected 200 observation points to perform statistical inferences, and results suggested that, for both processes (minimum and maximum daily humidity monitoring), the unit-Lindley distribution assumption is valid (with estimated mean parameter values: $\hat{\mu} = 0.584$ and 0.760 for the MIN and MAX periods/processes, respectively). Also, in comparison with other used distributions (e.g., beta, simplex and Kumaraswamy), the Kolmogorov-Smirnov goodness-of-fit test (for details on such a test, see, e.g., [48]) corroborates with the unit-Lindley adoption, for both the minimum and maximum TS, as shown in Table 3.

Table 3. The *p*-values from the Kolmogorov-Smirnov goodness-of-fit test for some continuous distributions defined on the range $(0, 1)$ adjusted to the minimum and maximum daily relative humidity data.

| Distribution	Minimum	Maximum
Unit-Lindley	0.769	0.797
Beta	0.104	0.012
Simplex	0.089	0.038
Kumaraswamy	0.176	0.015

Table 4 shows the control limits of both studied processes (using data from December 2016 to December 2020 only), based on the unit-Lindley distribution (unit-Lindley chart), and considering different tolerances to false alarms or type 1 error (α).

Table 4. Control limits of the unit-Lindley chart for the minimum and maximum daily humidity monitoring.

| Tolerance	Minimum			Maximum		
(α)	LCL	CL ($\hat{\mu}$)	UCL	LCL	CL ($\hat{\mu}$)	UCL
0.15	0.197	0.584	0.840	0.447	0.760	0.927
0.10	0.142	0.584	0.856	0.361	0.760	0.934
0.01	0.017	0.584	0.906	0.061	0.760	0.958

In Figure 6, we present the control charts for each period of the day (parts 1–4), adopting the unit-Lindley distribution for the minimum and maximum observations (as interval data in blue), considering those daily period records (solid black lines, MIN-MAX respectively), with three tolerance (α) levels (15% as red thick dashed line, 10% as red thin dashed line, and 1% as solid red line). These control charts were developed for interval data in SPC, considering the estimated unit-Lindley LCL of the minimum humidity TS and the estimated unit-Lindley UCL of the maximum TS.

As a next step, we developed a three-dimensional (3D) visualization, considering both control limits for the maximum/minimum daily humidity. Figure 7 shows an SDA bivariate control chart for the humidity monitoring, adopting all the 2021 data records as online process monitoring (i.e., for Phase 2 analysis). This plot contains the maximum humidity values in the z-axis, the minimum humidity values in the y-axis, and the time observation points in the x-axis (as dots). Moreover, the z- and x-axes (background series) project the maximum humidity TS, while the y- and x-axes (bottom series) project the minimum humidity TS.
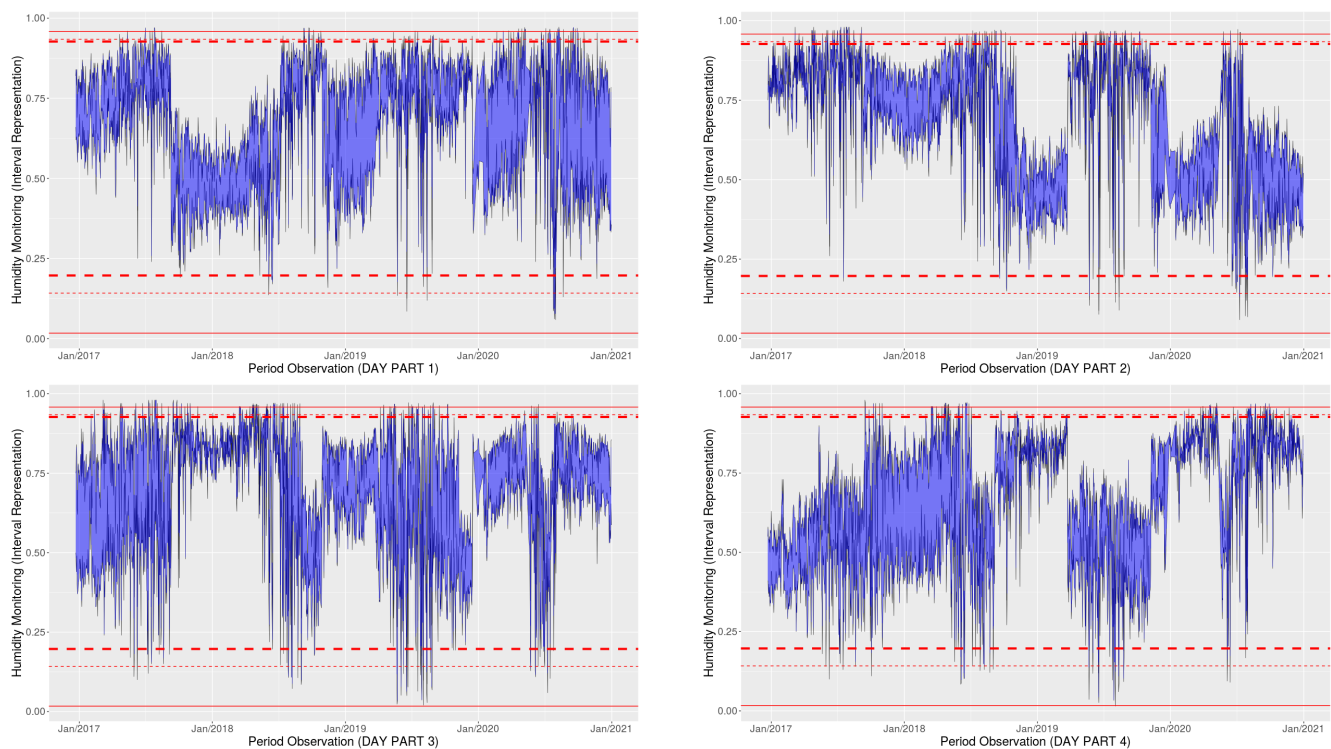
Figure 6. A visualization of the dynamic of Phase 1 (minimum and maximum TS), considering the observation points as four periods of the day, one in each graphic. Top-left panel: the day period part 1 (from midnight to 5:59 a.m. UTC); top-right panel: the day period part 2 (from 6:00 a.m. to 11:59 a.m. UTC); bottom-left panel: the day period part 3 (from midday to 5:59 p.m. UTC); bottom-right panel: the day period part 4 (from 6:00 p.m. to 11:59 p.m. UTC). Thus, the red lines represent three tolerance (α) levels (15% as thick dashed line, 10% as thin dashed line, and 1% as solid line) for each estimated control limit (UCL for the maximum TS, and LCL for the minimum TS).

The bivariate plot (minimum-maximum control chart) results in a rectangular box as a control/expected behaviour. Therefore, any observation point that out-patterns the expected dynamic (also called out-of-control) will be coloured in red during 3D plotting (Phase 2), followed by its projections. For instance, by considering the maximum TS (UCL = 0.927), it is notable that days 8 January 2021 (night), 11 January 2021 (dawn), 11 January 2021 (night), 20 January 2021 (dawn) and 20 January 2021 (night) were highlighted, being more significant than the tolerance level of 15%. Whereas, for the minimum TS (UCL = 0.840), 12 days (all during the night period) were extrapolated to the control limits (with ten observation points being from January 2021). Further investigations need to be conducted to appoint a stable tolerance level (α) to estimate these LCL and UCL, although the presented methodology shows to be competitive to make inferences about this process.
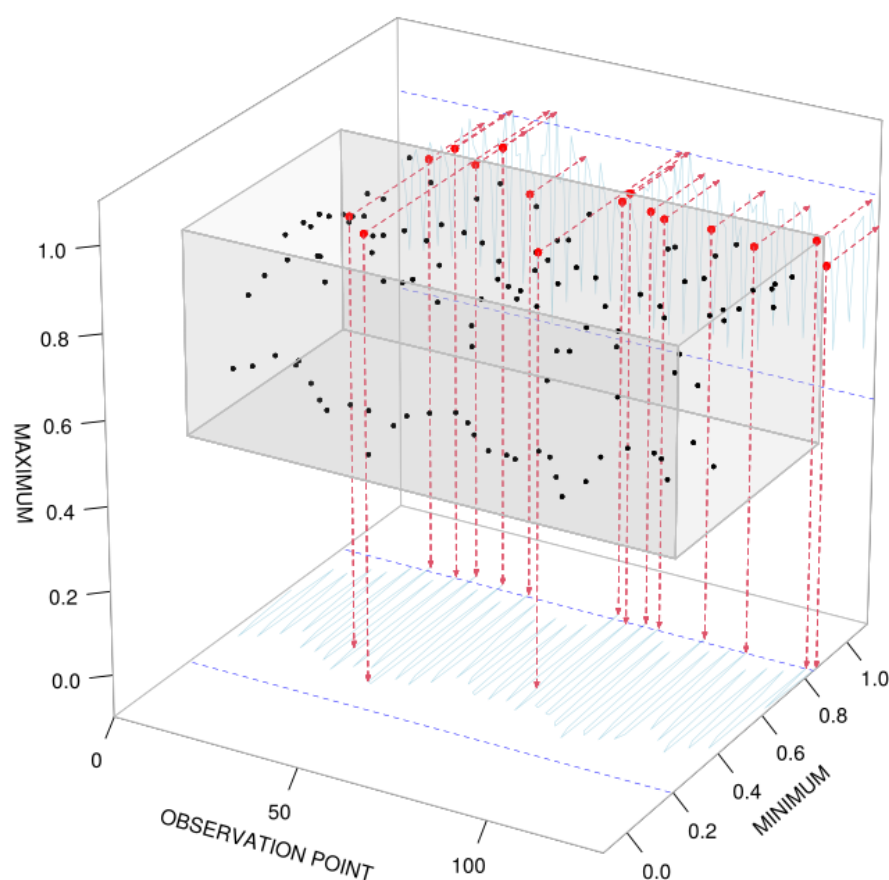
Figure 7. SDA bivariate control chart for the daily humidity in Phase II monitoring (records from 2021). Through the 3D plot, the z-axis represents the maximum upper bound and the y-axis the minimum lower bound from the daily humidity (aggregated per periods), adopting a certain tolerance level ($\alpha = 0.15$ or 15%), whereas the x-axis is related to the time observation points (as dots). The estimated control limits are represented as a shaded box, observed out-of-control points are highlighted as red points and their projections placed in the control chart projections. Thus, the TS projection on the bottom (x- and y-axes) is the control chart related to the minimum daily humidity, and the TS projection on the background (x- and z-axes) is the control chart of the maximum daily humidity.

6. Concluding Remarks and Future Prospects

In this paper, we developed a new control chart based on the unit-Lindley distribution by [17], named as unit-Lindley chart, and its inferential properties. Moreover, we also showed the competitiveness of working with interval data representation (as a SDA) to contour the missing data problem and the presence of noise.

As demonstrated by the simulation and empirical studies, the proposed control chart can be an efficient, exciting and valuable alternative to some well-known SPC tools when dealing with continuous process data in the interval $(0, 1)$, e.g., indices, rates and proportions, not resulting from Bernoulli experiments. For instance, the most common control charts used to monitor this kind of data are based on the beta, simplex and Kumaraswamy distributions, which present more parameters and show some limits towards overdispersed data, confirmed through an extensive MC simulation study (using the ARL, SDRL and MRL metrics).

The developed parametric SPC tool enlightens the prediction and opens new doors to discuss extreme events in the Atacama water particles monitoring through probabilistic reasoning. Further works shall explore the extension of this work to a unit-Lindley regression model (which enables to include trend and season components and spatial dependence).

References

1. Petitgas, P. The CUSUM out-of-control table to monitor changes in fish stock status using many indicators. *Aquat. Living Resour.* **2009**, *22*, 201–206. [CrossRef]
2. Hanslik, T.; Boelle, P.Y.; Flahault, A. The control chart: An epidemiological tool for public health monitoring. *Public Health* **2001**, *115*, 277–281. [CrossRef]
3. Khan, Z.; Gulistan, M.; Hashim, R.; Yaqoob, N.; Chammam, W. Design of S-control chart for neutrosophic data: An application to manufacturing industry. *J. Intell. Fuzzy Syst.* **2020**, *38*, 4743–4751. [CrossRef]
4. Sellers, K.F. A generalized statistical control chart for over-or under-dispersed data. *Qual. Reliab. Eng. Int.* **2012**, *28*, 59–65. [CrossRef]
5. Woodall, W.H. Control charts based on attribute data: Bibliography and review. *J. Qual. Technol.* **1997**, *29*, 172–183. [CrossRef]
6. Joekes, S.; Barbosa, E.P. An improved attribute control chart for monitoring non-conforming proportion in high quality processes. *Control. Eng. Pract.* **2013**, *21*, 407–412. [CrossRef]
7. Lima-Filho, L.M.d.A.; Bayer, F.M. Kumaraswamy control chart for monitoring double bounded environmental data. *Commun. Stat.-Simul. Comput.* **2019**, 1–16. [CrossRef]
8. Abbas, Z.; Nazir, H.Z.; Akhtar, N.; Abid, M.; Riaz, M. On designing an efficient control chart to monitor fraction nonconforming. *Qual. Reliab. Eng. Int.* **2020**, *36*, 547–564. [CrossRef]
9. Montgomery, D.C. *Introduction to Statistical Quality Control*; John Wiley & Sons: Hoboken, NJ, USA, 2020.
10. Lee Ho, L.; Fernandes, F.H.; Bourguignon, M. Control charts to monitor rates and proportions. *Qual. Reliab. Eng. Int.* **2019**, *35*, 74–83. [CrossRef]
11. Sant'Anna, Â.M.O.; Ten Caten, C.S. Beta control charts for monitoring fraction data. *Expert Syst. Appl.* **2012**, *39*, 10236–10243. [CrossRef]
12. Korkmaz, M.Ç.; Chesneau, C.; Korkmaz, Z.S. On the arcsecant hyperbolic normal distribution. Properties, quantile regression modeling and applications. *Symmetry* **2021**, *13*, 117. [CrossRef]
13. Bakouch, H.S.; Nik, A.S.; Asgharzadeh, A.; Salinas, H.S. A flexible probability model for proportion data: Unit-half-normal distribution. *Commun. Stat. Case Stud. Data Anal. Appl.* **2021**, *7*, 271–288.
14. Bantan, R.A.R.; Chesneau, C.; Jamal, F.; Elgarhy, M.; Tahir, M.H.; Ali, A.; Zubair, M.; Anam, S. Some new facts about the unit-Rayleigh distribution with applications. *Mathematics* **2020**, *8*, 1954. [CrossRef]
15. Mazucheli, J.; Menezes, A.; Dey, S. The unit-Birnbaum-Saunders distribution with applications. *Chil. J. Stat.* **2018**, *9*, 47–57.
16. Zellner, A.; Keuzenkamp, H.A.; McAleer, M. *Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple*; Cambridge University Press: Cambridge, UK, 2001.
17. Mazucheli, J.; Menezes, A.F.B.; Chakraborty, S. On the one parameter unit-Lindley distribution and its associated regression model for proportion data. *J. Appl. Stat.* **2019**, *46*, 700–714. [CrossRef]
18. Bonnail, E.; Lima, R.C.; Turrieta, G.M. Trapping fresh sea breeze in desert? Health status of Camanchaca, Atacama's fog. *Environ. Sci. Pollut. Res.* **2018**, *25*, 18204–18212. [CrossRef]
19. Schemenauer, R.S.; Fuenzalida, H.; Cereceda, P. A neglected water resource: The Camanchaca of South America. *Bull. Am. Meteorol. Soc.* **1988**, *69*, 138–147. [CrossRef]
20. Diday, E. Thinking by classes in data science: The symbolic data analysis paradigm. *Wiley Interdiscip. Rev. Comput. Stat.* **2016**, *8*, 172–205. [CrossRef]

21. Nascimento, D.C.; Pimentel, B.; Souza, R.; Leite, J.P.; Edwards, D.J.; Santos, T.E.; Louzada, F. Dynamic time series smoothing for symbolic interval data applied to neuroscience. *Inf. Sci.* **2020**, *517*, 415–426. [CrossRef]

22. Bull, A.T.; Andrews, B.A.; Dorador, C.; Goodfellow, M. *Introducing the Atacama Desert*; Springer: Berlin/Heidelberg, Germany, 2018.

23. Grosjean, M.; Veit, H. Water Resources in the Arid Mountains of the Atacama Desert (Northern Chile): Past Climate Changes and Modern Conflicts. In *Global Change and Mountain Regions: An Overview of Current Knowledge*; Huber, U.M., Bugmann, H.K.M., Reasoner, M.A., Eds.; Springer: Dordrecht, The Netherlands, 2005; pp. 93–104. [CrossRef]

24. García, A.; Ulloa, C.; Amigo, G.; Milana, J.P.; Medina, C. An inventory of cryospheric landforms in the arid diagonal of South America (high Central Andes, Atacama region, Chile). *Quat. Int.* **2017**, *438*, 4–19. [CrossRef]

25. Donoso, G.; Lictevout, E.; Rinaudo, J.D. Groundwater management lessons from Chile. In *Sustainable Groundwater Management*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 481–509.

26. Suárez, F.; Muñoz, J.F.; Fernández, B.; Dorsaz, J.M.; Hunter, C.K.; Karavitis, C.A.; Gironás, J. Integrated water resource management and energy requirements for water supply in the Copiapó river basin, Chile. *Water* **2014**, *6*, 2590–2613. [CrossRef]

27. Abramowitz, M.; Stegun, I.A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*; US Government Printing Office: Washington, DC, USA , 1964; Volume 55.

28. Goulet, V. Expint: Exponential Integral and Incomplete Gamma Function; R Package, 2016. Available online: https://cran.r-project.org/package=expint (accessed on 26 January 2021).

29. Corless, R.M.; Gonnet, G.H.; Hare, D.E.G.; Jeffrey, D.J.; Knuth, D.E. On the LambertW function. *Adv. Comput. Math.* **1996**, *5*, 329–359. [CrossRef]

30. Borchers, H.W. Pracma: Practical Numerical Math Functions; R Package Version 2.2.9; 2019. Available online: https://cran.r-project.org/package=pracma (accessed on 29 January 2021).

31. Cox, D.R.; Snell, E.J. A general definition of residuals. *J. R. Stat. Soc. Ser. B (Methodol.)* **1968**, *30*, 248–265. [CrossRef]

32. Bayer, F.M.; Tondolo, C.M.; Müller, F.M. Beta regression control chart for monitoring fractions and proportions. *Comput. Ind. Eng.* **2018**, *119*, 416–426. [CrossRef]

33. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.

34. Saghir, A.; Lin, Z. Control charts for dispersed count data: an overview. *Qual. Reliab. Eng. Int.* **2015**, *31*, 725–739. [CrossRef]

35. Riaz, M.; Ajadi, J.O.; Mahmood, T.; Abbasi, S.A. Multivariate mixed EWMA-CUSUM control chart for monitoring the process variance-covariance matrix. *IEEE Access* **2019**, *7*, 100174–100186. [CrossRef]

36. Jensen, W.A.; Jones-Farmer, L.A.; Champ, C.W.; Woodall, W.H. Effects of parameter estimation on control chart properties: A literature review. *J. Qual. Technol.* **2006**, *38*, 349–364. [CrossRef]

37. Moraes, D.; Oliveira, F.L.P.d.; Quinino, R.d.C.; Duczmal, L.H. Self-oriented control charts for efficient monitoring of mean vectors. *Comput. Ind. Eng.* **2014**, *75*, 102–115. [CrossRef]

38. Paroissin, C.; Penalva, L.; Pétrau, A.; Verdier, G. New control chart for monitoring and classification of environmental data. *Environmetrics* **2016**, *27*, 182–193. [CrossRef]

39. Jorgensen, B. *The Theory of Dispersion Models*; CRC Press: Boca Raton, FL, USA, 1997.

40. Kumaraswamy, P. A generalized probability density function for double-bounded random processes. *J. Hydrol.* **1980**, *46*, 79–88. [CrossRef]

41. Wikipedia. Copiapó—Wikipedia, The Free Encyclopedia, 2021. Available online: https://en.wikipedia.org/w/index.php?title=Copiapó&oldid=1013845587 (accessed on 26 April 2021).

42. Cleveland, W.S. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **1979**, *74*, 829–836. [CrossRef]

43. Bock, H.H.; Diday, E. *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*; Springer: Berlin/Heidelberg, Germany, 2000.

44. Billard, L.; Diday, E. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*; John Wiley: Hoboken, NJ, USA, 2006.

45. Diday, E.; Noirhomme-Fraiture, M. *Symbolic Data Analysis and the SODAS Software*; John Wiley & Sons: Hoboken, NJ, USA, 2008.

46. Billard, L.; Diday, E. *Clustering Methodology for Symbolic Data*; John Wiley & Sons: Hoboken, NJ, USA, 2019.

47. Diday, E.; Guan, R.; Saporta, G.; Wang, H. *Advances in Data Science: Symbolic, Complex, and Network Data*; John Wiley & Sons: Hoboken, NJ, USA, 2020.

48. Conover, W.J. *Practical Nonparametric Statistics*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 1999.