PLOS ONE

RESEARCH ARTICLE

Unraveling water monitoring association towards weather attributes for response proportions data: A unit-Lindley learning

Paulo H. Ferreira[1]*, **Anderson O. Fonseca**[1], **Diego C. Nascimento**[2], **Estefania Bonnail**[3], **Francisco Louzada**[4]

1 Department of Statistics, Federal University of Bahia, Salvador, Bahia, Brazil, **2** Department of Mathematics, University of Atacama, Copiapó, Atacama, Chile, **3** Coastal Research Center, University of Atacama, Copiapó, Atacama, Chile, **4** Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, São Paulo, Brazil

* paulohenri@ufba.br

Abstract

Learning techniques involve unraveling regression structures, which aim to analyze in a probabilistic frame the associations across variables of interest. Thus, analyzing fraction and/or proportion data may not be adequate with standard regression procedures, since the linear regression models generally assume that the dependent (outcome) variable is normally distributed. In this manner, we propose a statistical model called unit-Lindley regression model, for the purpose of Statistical Process Control (SPC). As a result, a new control chart tool was proposed, which targets the water monitoring dynamic, as well as the monitoring of relative humidity, per minute, of Copiapó city, located in Atacama Desert (one of the driest non-polar places on Earth), north of Chile. Our results show that variables such as wind speed, 24-hour temperature variation, and solar radiation are useful to describe the amount of relative humidity in the air. Additionally, Information Visualization (InfoVis) tools help to understand the time seasonality of the water particle phenomenon of the region in near real-time analysis. The developed methodology also helps to label unusual events, such as *Camanchaca*, and other water monitoring-related events.

Citation: Ferreira PH, Fonseca AO, Nascimento DC, Bonnail E, Louzada F (2022) Unraveling water monitoring association towards weather attributes for response proportions data: A unit-Lindley learning. PLoS ONE 17(10): e0275841. https://doi.org/10.1371/journal.pone.0275841

Editor: Zaher Mundher Yaseen, TDTU: Ton Duc Thang University, VIET NAM

Received: May 26, 2022

Accepted: September 24, 2022

Published: October 14, 2022

Data Availability Statement: The adopted data set is available at: https://doi.org/10.34740/KAGGLE/DSV/4051087, while the developed R script is available at: https://github.com/ProfNascimento/ULreg.

Funding: Anderson O. Fonseca acknowledges the support from Bahia State Research Foundation (FAPESB Proc. 084.0508.2020.0002837-61). Francisco Louzada acknowledges support from the São Paulo State Research Foundation (FAPESP Processes 2013/07375-0) and CNPq (grant no.

1 Introduction

Data acquisition related to natural resources is, day to day, more and more abundant, due to the miniaturization and reduction of data storage costs. Additionally, adopting the Internet of Things (IoT) technology helps to connect countless of decentralized devices and sustainability sources in benefits of analyzing to enhance planning, delivery, and efficiency of existing sources. These elements foment the decision-making performed on processing in near real-time, which remains a big challenge, given the need to configure systems to run every few minutes or hours, which causes them to process only the most recent features stored using robust data analysis tools.

Environmental Indicators (EI) play an important role in the sustainability in order to disseminate global environment statistics, based on the wide range of data sources, streamlining

301976/2017-1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

processes with IoT. Combined with Information Visualization (InfoVis), visual techniques provide support for specialists to visually summarize, explore, and reveal trends and patterns within data sets. Nonetheless, they are still in an early stage of development in many countries, and data are often sparse.

For instance, water source patterns may be seen as a combination of attributes (such as radiation, temperature, wind information, etc.) that impact directly this water proportion/rate. Moreover, in Atacama Desert, the water resources limitation demands extra need to unravel such dependent structure for water particles monitoring [1] and watershed estimation/forecasting.

In the case of rates and proportions processes, whereas the observed variable assumes values in the range (0, 1), there is a well-represented class of models, the unit distributions family, which deals with this type of sensor data, but are often univariate and not extended, in their inference, to some regression structures (see, e.g., [2–5]). Regression structures, in probabilistic modeling, can provide a flexible set of tools for examining such associations, while enabling, either, potentially confounding effects of other factors, or interaction effects for Statistical Process Control (SPC) tools [6].

This study considers a statistical model called unit-Lindley regression model [7], in which through the logit transformation, new variables can be incorporated into the parameter estimation, and the adequacy of some statistical association across those explanatory variables on the response variable can be verified. Furthermore, for the purpose of SPC, a new control chart is proposed targeting the water monitoring dynamic in Copiapó city, located in Atacama Desert (one of the driest non-polar places on Earth; see [8, 9]), north of Chile.

The upcoming contents of this paper are organized as follows. Section 2 describes the practical motivation. In Section 3, we present the unit-Lindley learning model and some of its basic properties (Subsection 3.1), as well as the novel control chart based on it (Subsection 3.2). Section 4 provides simulation studies designed to assess the performance of the proposed unit-Lindley regression control chart. Section 5 illustrates some findings towards the water particles monitoring in Atacama Desert through EI relationship. Finally, Section 6 concludes the paper with a few remarks and discussions on future studies.

2 Motivation

Water is the most precious resource that enables us to maintain the fauna and flora of a region, thus the existence of water resources is conditioned for its under-existence. An important water source from the Andean range is the cryosphere defrost, which generates water flows and underwater basins (Fig 1). Additionally, a phenomenon called *Camanchaca* occurs, whereas marine stratocumulus cloud banks that form in the Chilean coast blow as a passageway of "low clouds", right after sunrise, in sequence for a couple of hours, creating a huge influence on the marks and infiltrating along the river valleys [10].

Relative humidity is the ratio of the partial pressure of water vapor in the air to reach an equilibrium in vapor pressure (of water), and there are three elements to be related with the water precipitation phenomenon: *temperature*, *wind movement*, and *solar radiation*. It is, therefore, the effective ratio of water content of the air in relation to the maximum water content that the air could contain (water in the form of vapor).

The maximum water content of the air depends directly on the temperature (the higher the temperature, the more water the air can contain, which is why when the temperature cools, the relative humidity increases, sometimes reaching saturation, which leads to the formation of fog or *Camanchaca*). Additionally, the wind moves the air masses and will, therefore, influence in their water content. The wind will also affect the evaporation of water, which will, again,

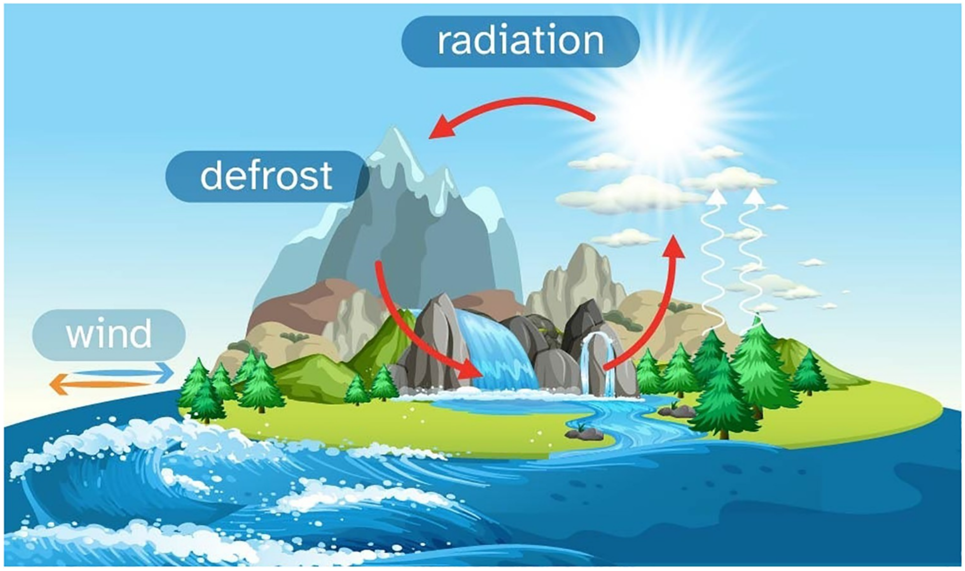

Fig 1. Water cycle sources (solid, liquid and vapor phases). Source: elaborated by the authors.

https://doi.org/10.1371/journal.pone.0275841.g001

modify the water content of the air masses (but directly linked to the moves of air masses). Finally, radiation affects the temperature and, therefore, the equilibrium in vapor pressure.

During precipitation, the falling water will partly evaporate, again modifying the water content of the air masses (but also decreasing the temperature since evaporation requires an energy input, drawn from the environment). The altitude from which precipitation is initiated is, therefore, important. The lower layers will gain water and lose temperature, but the cloud will lose water and warm up.

As for the *Camanchaca* phenomenon, fog often forms at sea level during the night, when the air temperature cools both because of the night and because of the contact with the cold waters of Humboldt Current [1]. The westerly winds, then, push these fog banks inland, preferably following the valleys, as the reliefs crossing implies in a loss of humidity by cooling (thus, a decrease in relative humidity, which potentially causes the fog to disappear).

The particular climatological conditions of the region often imply in a very low temperature inversion, normally present around 10,000 meters (m) over the Earth, but sometimes below 1,000 m in the Atacama region [11]. Due to the altitude, the temperature normally decreases till the tropopause (the boundary between the troposphere and the stratosphere), and then increases again in the stratosphere. In the Atacama region, the first temperature inversion around 1,000 m or less (and then another one, much more higher, that determines the tropopause), prevents the *Camanchaca* from going up in altitude (a temperature inversion is almost an insurmountable wall for air masses, with only the stratocumulus clouds having enough energy to cross this border).

3 Methodology

Human learning is majorly associated, directly and indirectly, with visual stimulation, whereas almost half of the neural tissue of the cognitive systems is related to pattern recognition through the vision [12]. The area of InfoVis aims to develop and apply visual representations towards the modeling and understanding of attribute values, relationships, and information extraction from data [13, 14]. InfoVis takes advantage of human cognition abilities to

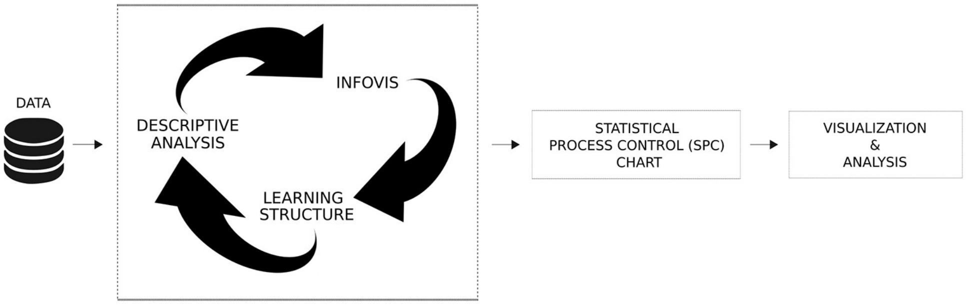

Fig 2. Visual representation of the adopted methodology.

https://doi.org/10.1371/journal.pone.0275841.g002

transform abstract data into visual information, since the effort to identify, interpret, and extract patterns is reduced when raw data is depicted in the form of graphical elements [15].

InfoVis techniques have been successfully applied to the analyses of different data sets (e.g., those related to business planning, social networks, climate, pollution, finances, criminal cases, etc.) [16–19] and in the identification of patterns and anomalies to be used as information to support decision-making [20], and can represent, both, unidimensional and multidimensional data (data sets with a large number of attributes related to each data record). InfoVis approach graphically represent information through computer systems, which show alternative data visions to describe their structures [21]. However, a large gap lies between the generation and storage of data, and the ability of analytical tools to process, organize and properly display the extracted information [22, 23].

In this manner, the following three-phase procedure is adopted towards the collected data: a Descriptive Analysis is combined with InfoVis, then an inference analysis is made through Learning Structure approach, with the unit-Lindley regression model (summarized in Fig 2).

In the following subsections, we will present, in details, the inferences related to the unit-Lindley distribution (Subsection 3.1), then its extension to SPC reasoning, as well as complementary regression (Subsection 3.2).

3.1 The unit-Lindley learning

Recently, [7] have introduced a one-parameter continuous probability distribution, defined by the interval (0, 1), called the unit-Lindley (UL) distribution. In this study, we shall consider a mean parameterized form of the UL distribution, also presented by the authors.

A random variable Y has a UL distribution with parameter $0 < \mu < 1$, denoted by $Y \sim$ UL (μ), if its cumulative distribution function is given by

$$F(y \mid \mu) = 1 - \left[1 - \frac{y(1-\mu)}{y-1}\right] \exp\{-\frac{y(1-\mu)}{\mu(1-y)}\}, \quad \text{for } 0 < y < 1.$$

The corresponding probability density function (PDF) is

$$f(y \mid \mu) = \frac{(1-\mu)^2}{\mu(1-y)^3} \exp\{-\frac{y(1-\mu)}{\mu(1-y)}\}, \quad \text{for } 0 < y < 1.$$

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

$$\mathbb{E}[Y] = \mu \quad \text{and} \quad \mathbb{V}\text{ar}[Y] = \mu\left[\left(\frac{1}{\mu} - 1\right)^2 \exp\left\{\frac{1}{\mu} - 1\right\} Ei\left(1, \left(\frac{1}{\mu} - 1\right)\right) - \frac{1}{\mu} + 2\right] - \mu^2,$$

in which $Ei(a, z) = \int_1^\infty y^{-a} e^{-yz}\, dy$ is the exponential integral function [24].

The quantile function of the UL(μ) distribution can be written as

$$Q(p \mid \mu) = F^{-1}(p \mid \mu) = \frac{\frac{1}{\mu} + W_{-1}\left(\frac{(p-1)}{\mu}\exp\left\{-\frac{1}{\mu}\right\}\right)}{1 + W_{-1}\left(\frac{(p-1)}{\mu}\exp\left\{-\frac{1}{\mu}\right\}\right)}, \quad \text{for } 0 < p < 1, \tag{1}$$

in which W_{-1} is the negative branch of the Lambert W function [25].

In a regression analysis, it is very common to model the mean of the response variable (the variable of interest) as a function of several other variables, also called explanatory variables or covariates. The UL distribution described above can be easily and promptly used in this context, as demonstrated by [7].

Let $Y_1, Y_2, \ldots, Y_n$ be n independent random variables, in which $Y_i \sim \text{UL}(\mu_i)$, for $i = 1, 2, \ldots, n$. The so-called UL regression model is defined assuming that the mean of Y_i satisfies the following functional relation (linear predictor):

$$g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{2}$$

in which $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)^\top \in \mathbb{R}^k$ denotes a k-dimensional vector of regression coefficients ($k < n$), $\boldsymbol{x}_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ik})$ represents the observations on k known covariates, and $g(\cdot)$ is a strictly monotonic and twice differentiable function that maps the interval (0, 1) into $\mathbb{R}$ (mean link function).

In this study, as in [7], we shall consider the logit link function, which ensures that the predicted mean stays within bounds (0, 1). Hence, the regression structure for μ_i is given by

$$\text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i^\top \boldsymbol{\beta}.$$

Among other possible choices for the mean link function, we should mention the probit, cauchit, log-log and complementary log-log link functions (for these and other useful link functions, see [26]).

Under a classical inference approach, the unknown parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)^\top$ can be estimated by maximizing the log-likelihood function:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell(\mu_i),$$

in which

$$\ell(\mu_i) = 2\log(1-\mu_i) - \log(\mu_i) - 3\log(1-y_i) - \frac{y_i(1-\mu_i)}{\mu_i(1-y_i)}$$

and

$$\mu_i = \text{logit}^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = \frac{\exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\}}.$$

Since the maximum likelihood (ML) estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ cannot be expressed in closed form, we must resort to iterative methods, such as the Newton-Raphson and Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithms (for further details, see, e.g., [27]), to obtain the parameter estimates.

3.2 The unit-Lindley regression control chart

Let us start this subsection with a brief review about the UL control chart [1], which is useful to statistically monitor variables of rate or proportion type, that are independent and with no control variables being present. Subsequently, we will present the proposed UL regression control chart.

Suppose that a process (e.g., a manufacturing or business process) generates outputs according to a UL(μ) distribution, and the probability of false alarm (or equivalently, type I error) is given by α. Then, the lower control limit (LCL), centerline (CL) and upper control limit (UCL) of the UL control chart are defined as follows [1]:

$$\mathrm{LCL} = Q(\alpha/2 \mid \mu), \qquad \mathrm{CL} = \mu, \qquad \mathrm{UCL} = Q(1-\alpha/2 \mid \mu),$$

in which $Q(.)$ is the quantile function presented in Eq (1).

It is worth mentioning that the use of the quantile function is justified by [28].

Nevertheless, the UL control chart does not consider situations in which the practitioner is required to impose a regression structure for the variable of interest. Our interest lies in cases in which the mean of the quality characteristic of interest (of rate or proportion type) is affected by control variables and can, then, be modeled as a function of them and unknown parameters.

Thus, considering the regression structure for μ_i defined in Eq (2) (e.g., using the logit link function), and a probability of false alarm equal to α (e.g., $\alpha = 0.0027$, which corresponds to the standard three-sigma rule or Six Sigma program), we have that the non-constant (or observation-specific) LCL, CL and UCL of the proposed UL regression control chart are given by

$$\mathrm{LCL}_i = Q(\alpha/2 \mid \mu_i), \qquad \mathrm{CL}_i = \mu_i, \qquad \mathrm{UCL}_i = Q(1-\alpha/2 \mid \mu_i),$$

for $i = 1, 2, \ldots, n$.

In practice, the ML estimator of μ_i is considered, with $\hat{\mu}_i = g^{-1}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})$, in which $\hat{\boldsymbol{\beta}}$ is the ML estimator of $\boldsymbol{\beta}$.

It is important to mention that there are two natural ways to visually represent and monitor this conditional structure: by adopting the *varying conditional mean* representation, or through its *residual* representation (that is, centralizing the process on zero by subtracting the expected value from the observed value).

4 Numerical evaluation

In this section, we conduct Monte Carlo (MC) simulations to assess and compare the performance of the proposed UL regression control chart with the existing beta regression control chart [28]. The performance comparisons are made in terms of the average run length (ARL), the median run length (MRL), and the standard deviation of the run length (SDRL). All statistical analyses were conducted using the R software version 3.6.3 [29].

The ARL is a popular measure used to assess the performance of control charts. The in-control and out-of-control ARLs are denoted as ARL_0 and ARL_1, respectively. The first one is defined as the average number of points plotted on the control chart until a signal occurs (that is, a single point falls beyond the control limits), assuming that the process is in control; whereas the second one represents the average number of observations that are taken before a mean shift is first detected when the process is out of control [30].

The MRL is the 50th percentage point of the run length (RL) distribution. In constrast to the ARL, the MRL is less affected by the skewness of the RL distribution [31]. The SDRL is a useful measure to estimate the dispersion of the RL distribution. Moreover, we will use the in-control (MRL_0 and SDRL_0) and out-of-control (MRL_1 and SDRL_1) versions of such metrics.

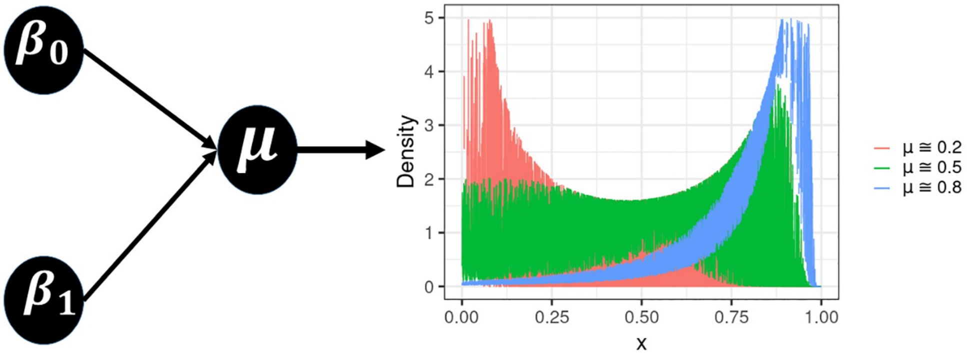

Fig 3. General description of the UL density function, followed by its illustration in different locations conditioned to the process mean parameter (μ).

https://doi.org/10.1371/journal.pone.0275841.g003

Let Y be the output of a process that follows a UL(μ) distribution, for instance. Also, let $\mu^{(0)}$ be the average of the process under control, and let $\mu^{(1)}$ be the average of the out-of-control process. Then, the ARL$_0$, MRL$_0$ and SDRL$_0$ are defined as

$$\mathrm{ARL}_0 = 1/\alpha, \qquad \mathrm{MRL}_0 = \log(0.5)/\log(1-\alpha), \qquad \mathrm{SDRL}_0 = \sqrt{(1-\alpha)/\alpha^2},$$

for $\alpha = \mathbb{P}(Y \notin [\mathrm{LCL}, \mathrm{UCL}] \mid \mu = \mu^{(0)})$. Whereas the ARL$_1$, MRL$_1$ and SDRL$_1$ are given by

$$\mathrm{ARL}_1 = 1/(1-\gamma), \qquad \mathrm{MRL}_1 = \log(0.5)/\log(\gamma), \qquad \mathrm{SDRL}_1 = \sqrt{\gamma/(1-\gamma)^2},$$

for $\gamma = \mathbb{P}(Y \in [\mathrm{LCL}, \mathrm{UCL}] \mid \mu = \mu^{(1)})$.

In the usual Six Sigma program, $\alpha = 0.0027$ and, therefore, ARL$_0$ = 1/0.0027 $\approx$ 370, MRL$_0$ = log(0.5)/log(1 – 0.0027) $\approx$ 256 and $\mathrm{SDRL}_0 = \sqrt{(1-0.0027)/0.0027^2} \approx 370$. This means, e.g. for the ARL$_0$, that even though the process is in control, on average, a false alarm (incorrect out-of-control signal) will be generated at every 370 points [32]. On the other hand, low (i.e., close to one) values of ARL$_1$ are desired, mainly for large-size shifts in the process mean.

4.1 General specifications

Without loss of generality, we may consider UL and beta processes with (in-control) mean parameter: $\mu^{(0)} \approx$ 0.2, 0.5 and 0.8, whose PDF plots are shown in Fig 3 (UL) and Fig 4 (beta), as

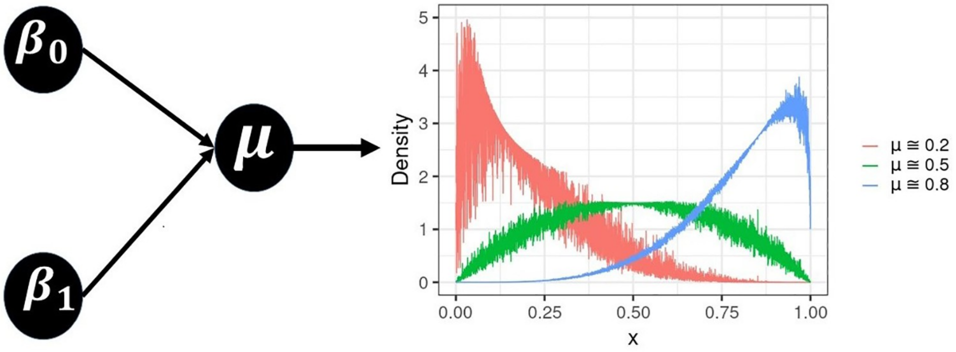

Fig 4. General description of the beta density function, followed by its illustration in different locations conditioned to the process mean parameter (μ).

https://doi.org/10.1371/journal.pone.0275841.g004

Table 1. Parameter values for scenarios considered in the simulation, when the true data-generating process is UL distributed (in-control condition).

Scenario	β_0	β_1	Characteristic
1	1.00	-6.16	$\mu \approx 0.2$
2	1.00	-2.00	$\mu \approx 0.5$
3	1.00	0.77	$\mu \approx 0.8$

https://doi.org/10.1371/journal.pone.0275841.t001

well as two distinct values for the probability of false alarm: $\alpha = 0.1$ (which corresponds to $\text{ARL}_0 = 10$, $\text{MRL}_0 \approx 6.579$ and $\text{SDRL}_0 \approx 9.487$) and 0.01 (which corresponds to $\text{ARL}_0 = 100$, $\text{MRL}_0 \approx 68.968$ and $\text{SDRL}_0 \approx 99.499$). We also use different sample sizes for each process: $n = 100, 200, 500$ and $1, 000$. These n observations are devoted to Phase I or retrospective analysis (process parameter estimation → assessment of process stability → control limits establishment). Whereas $n^* = 5, 000$ new observations are used to Phase II or prospective analysis (process monitoring → assessment of control chart performance). For the numerical evaluation, we consider 5,000 MC simulations (or replicates), which, according to [33], and also pointed out by other authors (e.g., [34]), is sufficient to obtain accurate results.

The generation of the data under control is based on the UL and beta regression models, with structure for the mean μ_i (around $\mu^{(0)}$) given by

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_i,$$

in which the values of the single covariate are drawn from the Uniform(0, 1) and Normal(1, 0.01) distributions, for the UL and beta processes, respectively, that is, $X_i \overset{\text{iid}}{\sim} \text{Uniform}(0, 1)$, for $i = 1, 2, \ldots, n, n + 1, \ldots, n + n^*$, when the true data-generating process is UL distributed, and $X_i \overset{\text{iid}}{\sim} \text{Normal}(1, 0.01)$, for $i = 1, 2, \ldots, n, n + 1, \ldots, n + n^*$, when the true data-generating process is beta distributed. Here, the abbreviation *iid* stands for "independent and identically distributed". The parameter values for the mean structure of each regression model are presented in Table 1 (UL) and Table 2 (beta).

For the estimation of beta regression model parameters, the R package `gamlss` [35] is used.

Finally, to compute the ARL_1, MRL_1 and SDRL_1, we consider shifts at different levels, representing percentage decreases and increases p in the process mean. The assumed levels are: $p = 1\%$ (down-shifted mean: $\mu^{(1)} \approx 0.198$, 0.495 and 0.792; up-shifted mean: $\mu^{(1)} \approx 0.202$, 0.505 and 0.808) to 10% (down-shifted mean: $\mu^{(1)} \approx 0.18$, 0.45 and 0.72; up-shifted mean: $\mu^{(1)} \approx$ 0.22, 0.55 and 0.88) and 20% (down-shifted mean: $\mu^{(1)} \approx 0.16$, 0.4 and 0.64; up-shifted mean: $\mu^{(1)} \approx 0.24$, 0.6 and 0.96). The parameter values for the mean structure of each regression model are shown in Table 3.

Table 2. Parameter values for scenarios considered in the simulation, when the true data-generating process is beta distributed (in-control condition).

Scenario	β_0	β_1	Characteristic
1	1.00	-2.39	$\mu \approx 0.2$ and $\sigma \approx 0.37$
2	1.00	-1.00	$\mu \approx 0.5$ and $\sigma \approx 0.45$
3	1.00	0.40	$\mu \approx 0.8$ and $\sigma \approx 0.38$

https://doi.org/10.1371/journal.pone.0275841.t002

Table 3. Parameter values for scenarios considered in the simulation (out-of-control condition).

UL		Beta		Characteristic
β_0	β_1	β_0	β_1	
0.48	-5.77	0.58	-2.26	$\mu \approx 0.160$
0.70	-5.89	0.91	-2.41	$\mu \approx 0.180$
0.91	-6.08	0.97	-2.37	$\mu \approx 0.198$
0.96	-6.11	1.02	-2.37	$\mu \approx 0.202$
1.18	-6.30	1.06	-2.35	$\mu \approx 0.220$
1.46	-6.57	1.35	-2.50	$\mu \approx 0.240$
0.47	-1.76	0.46	-0.87	$\mu \approx 0.400$
0.76	-1.91	0.67	-0.87	$\mu \approx 0.450$
1.05	-2.11	1.00	-1.02	$\mu \approx 0.495$
1.12	-2.16	1.07	-1.05	$\mu \approx 0.505$
1.48	-2.43	1.33	-1.14	$\mu \approx 0.550$
1.95	-2.88	1.69	-1.28	$\mu \approx 0.600$
0.37	0.40	0.37	0.22	$\mu \approx 0.640$
0.68	0.50	0.70	0.25	$\mu \approx 0.720$
0.99	0.67	1.02	0.32	$\mu \approx 0.792$
1.05	0.76	0.98	0.47	$\mu \approx 0.808$
1.39	1.22	1.44	0.57	$\mu \approx 0.880$
1.27	5.35	1.13	2.12	$\mu \approx 0.960$

https://doi.org/10.1371/journal.pone.0275841.t003

4.2 Simulation results

The results obtained with MC simulations, performed for each situation studied, are available in S1 Appendix. In short, such results seem to indicate (despite some slight to moderate discrepancies between the theoretical/target values of the performance measures, and the values calculated through simulations in some cases, which are, in fact, expected due to the effect of parameter estimation on control chart properties; see, for instance [36–38]) a good performance of the proposed regression control chart, mainly when the true data-generating process is UL distributed, and with increases in the process mean (this latter finding is in agreement with the results of [1]). In the cases in which data are generated from the beta distribution, the UL regression control chart tends to show slightly higher false alarm rates, whereas performing reasonably well when changes (increases and decreases) in the process mean occur. Once again, it is worth pointing out that the proposed regression control chart has its basis on a distribution with a single parameter (and, thus, more straightforward than the two-parameter beta distribution).

5 Application

The relative humidity of the air in the city of Copiapó, Chile, is particularly interesting to be monitored given that it is located in the heart of Atacama Desert, and also because it is an important northern Chilean city. The main economic source is related to the extraction of minerals and agriculture, both demanding great volume of water sources. Nevertheless, a natural water particle flux happens periodically given the geolocation of this city (placed in a valley, and about 60 kilometers open-field from the Chilean coast).

As a consequence, water events monitoring is needed, in relation to other natural phenomena, thus creating an opportunity to adopt the UL regression control chart for real data source.

The data set adopted in this study was acquired from the *Dirección General De Aeronáutica Civil, Dirección Meteorológica de Chile—Servicios Climáticos*, which provides several sets

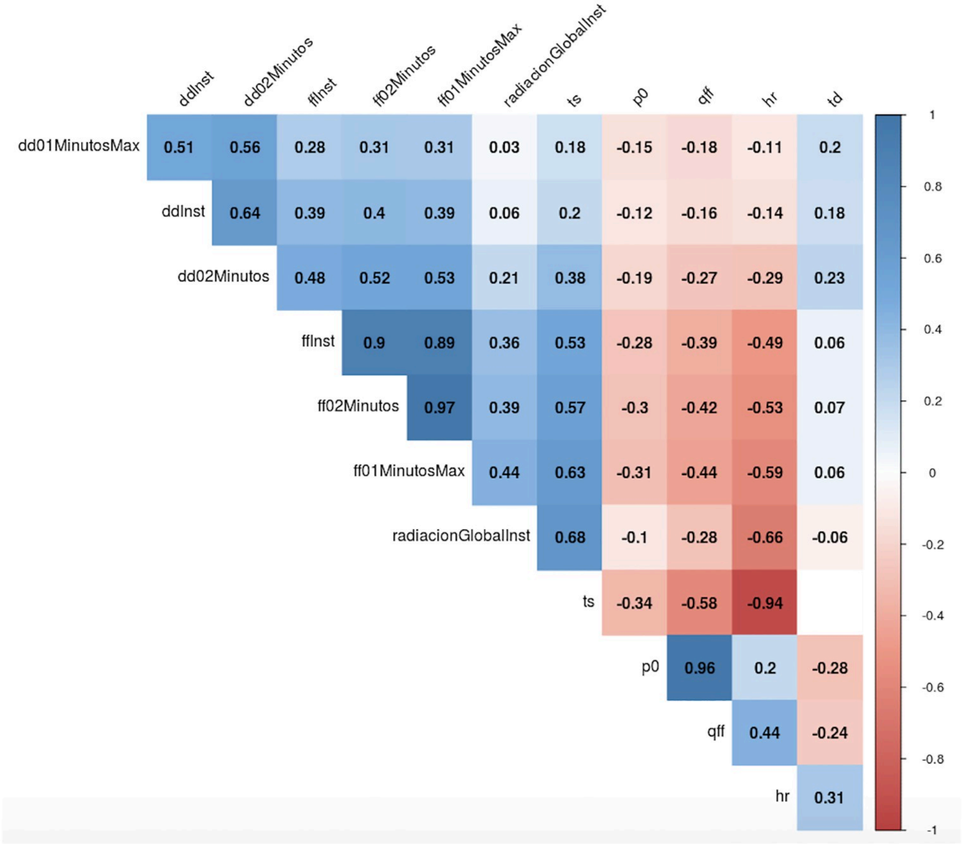

Fig 5. Correlation plot across weather variables.

https://doi.org/10.1371/journal.pone.0275841.g005

related to natural sources in Chile. Additionally, the water relative humidity-related *Wind*, *Temperature*, *Solar Radiacion*, and *Humidity Pressure* sets of the studied city. All records were taken from January 1st, 2019 to June 30th, 2021 (per minute), period in which some missing data are noticeable, resulting in a total of 1,237,596 data points. After data wrangling, we obtained a unified set containing 12 covariates: six related to the Wind dimension (*ddInst*, *ffInst*, *dd02Minutos*, *ff02Minutos*, *dd01MinutosMax*, *ff01MinutosMax*), two related to Temperature (*ts*, *td*), one related to Radiation (*radiacionGlobalInst*), and three related to Humidity (*hr*, *p0*, *qff*). The adopted data set is available at: https://doi.org/10.34740/KAGGLE/DSV/4051087, while the developed R script is available at: https://github.com/ProfNascimento/ULreg.

Since all variables are continuous, as a first relation metric, we used the Spearman's rank correlation coefficient. Fig 5 shows that, considering the relativity humidity (*hr*) variable, in module, the most associated covariates considering the same time period per dimension were: current temperature (*ts*), wind speed (*ffInst*), and solar radiation (*radiacionGlobalInst*). Moreover, in the Wind dimension, in module, the greatest correlated variable was *ff01MinutosMax*, thus highly positive related to the instantaneous speed (*ffInst*).

The first step before starting to seek some association across these explanatory variables (*X*'s), was to perform a multicollinearity checking. After adjusting an Ordinary Least Squares (OLS) model, three variables (*p0*, *qff*, *ts*) presented high Variance Inflation Factor (VIF), greater than 500 units in module, and were then excluded. Additionally, a presence of three clusters was noticed related to the response of the Wind, Temperature, and Radiation. Based

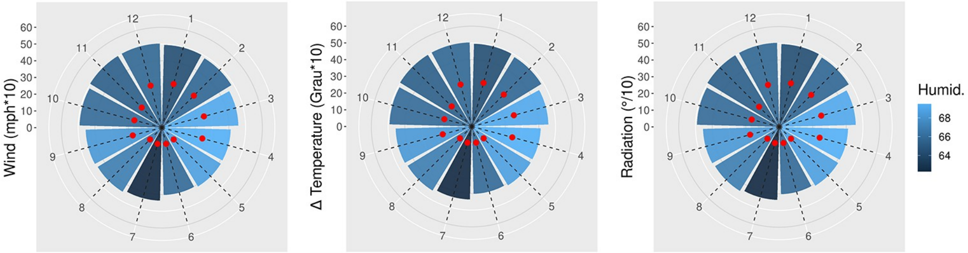

Fig 6. Radar plots: monthly wind variation (left-hand panel), 24-hour temperature variation (middle panel), and monthly solar radiation (right-hand panel). The red dots represent the minimum relative record for each month.

https://doi.org/10.1371/journal.pone.0275841.g006

on the expert knowledge, the selected variables to represent those clusters, and might be related to the relative humidity (theoretically), were the instantaneous wind speed (*ffInst*), instantaneous solar radiation (*radiacionGlobalInst*), and 24-hour temperature variation (*td*).

Looking closer into the instantaneous wind speed variable across time, Fig 6 depicts the monthly dynamic of this variable (left-hand panel), as well as the dynamic related to the instantaneous temperature (middle panel), and the solar radiation (right-hand panel). The lowest average records were from May to August, and the highest from December to February for all of these variables.

Fig 7 displays the variables' multidimensional relation, through the parallel plot, in which each line is a day-related record. Specifically, the top graphics are related to 10:00–10:01 a.m.

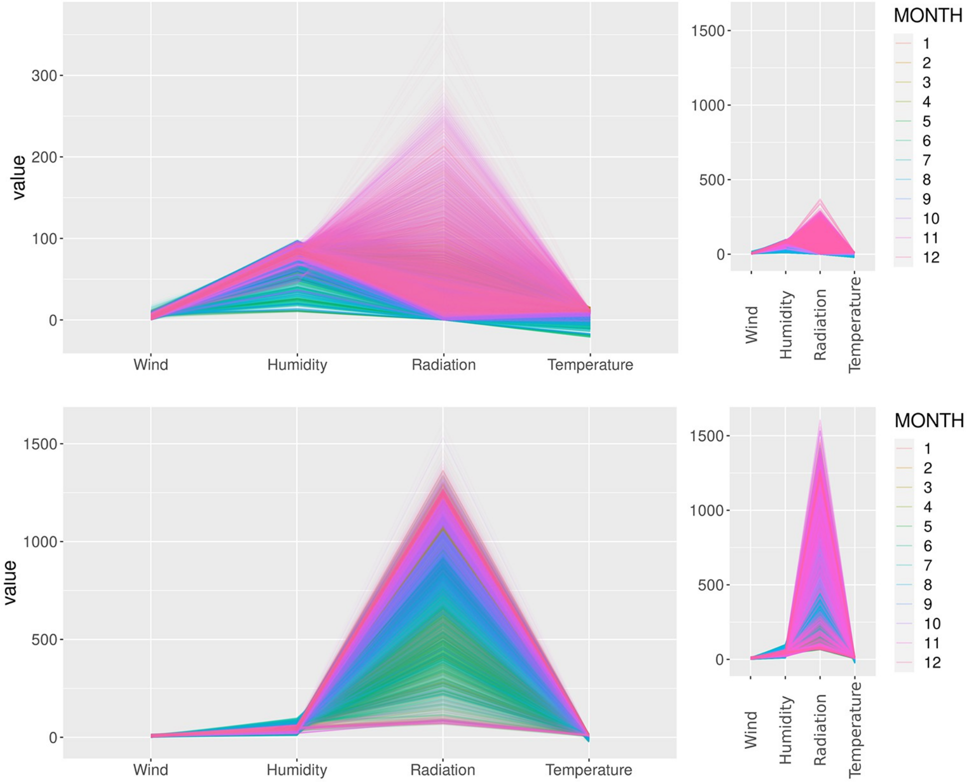

Fig 7. Parallel plots describing the records variation throughout time and across variables. The top graphics are related with Phase I, at 10 a.m., and, the bottom graphics, at 4 p.m. The summer period months (mostly pink tones) present great range of records across weather variables (wind, 24-hour temperature variation, solar radiation and relative humidity).

https://doi.org/10.1371/journal.pone.0275841.g007

Table 4. UL regression model adjusted for relative humidity data.

	Estimate	Std. error	*t* stat	*p*-value
Intercept	0.7103	0.0026	301.5	<0.0001
Wind Speed	-0.0695	0.0003	-259.7	<0.0001
ΔTemperature	0.0585	0.0002	266.7	<0.0001
Solar Radiation	-0.0013	<0.0001	-665.1	<0.0001

https://doi.org/10.1371/journal.pone.0275841.t004

(first minute), and the bottom graphics to 4:00–4:01 p.m. (first minute). Moreover, the bright colors are related to November until March (seasons in which the days receive sunlight earlier). For instance, higher values are noticeable in the summer time, during the morning (given the sunlight incidence), and vary more throughout the year in the afternoon.

The theoretical model adopted for this problematic was:

$$\text{Relative Humidity}_i \sim \text{UL}(\mu_i), \quad \text{in which}$$
$$\text{logit}\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1\,\text{Wind Speed}_i + \beta_2\,\Delta\text{Temperature}_i + \beta_3\,\text{Solar Radiation}_i,$$

for $i = 1, 2, \ldots, 1, 207, 079$ (Phase I).

Table 4 shows the estimates related to the three EI, which play an important role in the predictability, and association, of relative humidity. Since the link function related to the UL regression model is the logit function, further interpretation of β parameters requires to be transformed before. That is, for a unit increase in wind speed, on average, a change of $\exp\{-0.0695^* X\}/(1 + \exp\{-0.0695^* X\}) \rightarrow X = 1 \approx 0.483$ in the mean of the relative humidity, or the odds rate is to be expected. It is important to mention that a fraction contribution can also be adopted (based on proportion), since the response variable is a unit.

Thus, through the use of Phase II observation points, that is, the last 23 days, it is also possible to confirm the similarity across the association obtained from the UL regression results from Phase I. Fig 8 shows the pair relation across the studied variables, from which negative Spearman's rank correlation coefficients were obtained across Relative Humidity and Solar Radiation, and Relative Humidity and Wind Speed (similar to the estimated β coefficients from the UL regression model).

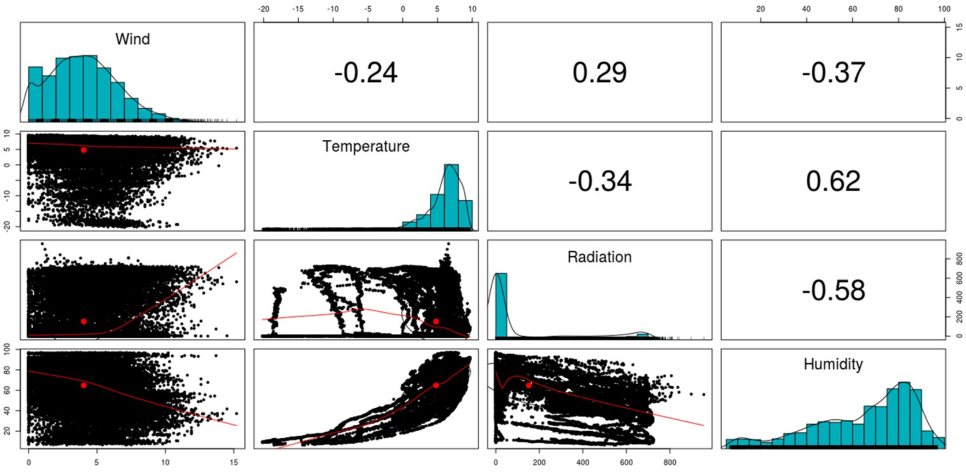

Fig 8. Pairwise representations adopting the Phase II observations, and their Spearman correlations.

https://doi.org/10.1371/journal.pone.0275841.g008

Table 5. Relative humidity's summary statistics obtained from the UL regression analysis, given time-varying SPC, for its expected mean and boundaries.

	LCL	CL (μ)	UCL
Minimum	0.014	**0.182**	0.436
1st Quartile	0.115	0.545	0.833
Median	0.223	0.669	0.898
Mean	0.207	**0.621**	0.863
3rd Quartile	0.289	0.718	0.918
Maximum	0.399	**0.779**	0.941

https://doi.org/10.1371/journal.pone.0275841.t005

The historical data were captured per minute, and the relative humidity, for Phase II, was observed during June 9th, 7:21 p.m. and July 1st, midnight. The total number of observation points was of 30,517, and the adoption of the UL regression model enables to estimate the process mean varying in time. Table 5 summarizes the expected air humidity, per minute, and its variation in Copiapó city. Considering a Six Sigma policy supervision, on average per minute, the relative humidity fluctuation is centered in 62.1%, with range between 18.2% and 77.9%. The maximum that could be observed was up to 94.1% of humidity (thus, some influx of water particles from the ocean happens, even though the analyzed city is placed at 350 m from the sea level and no elevation is placed between the coast and this city).

Moreover, the obtained SPC boundaries help to detect some massive water flux events, like *Camanchaca*, revealed by their discrepancy. Fig 9 shows the control chart in which the minutes

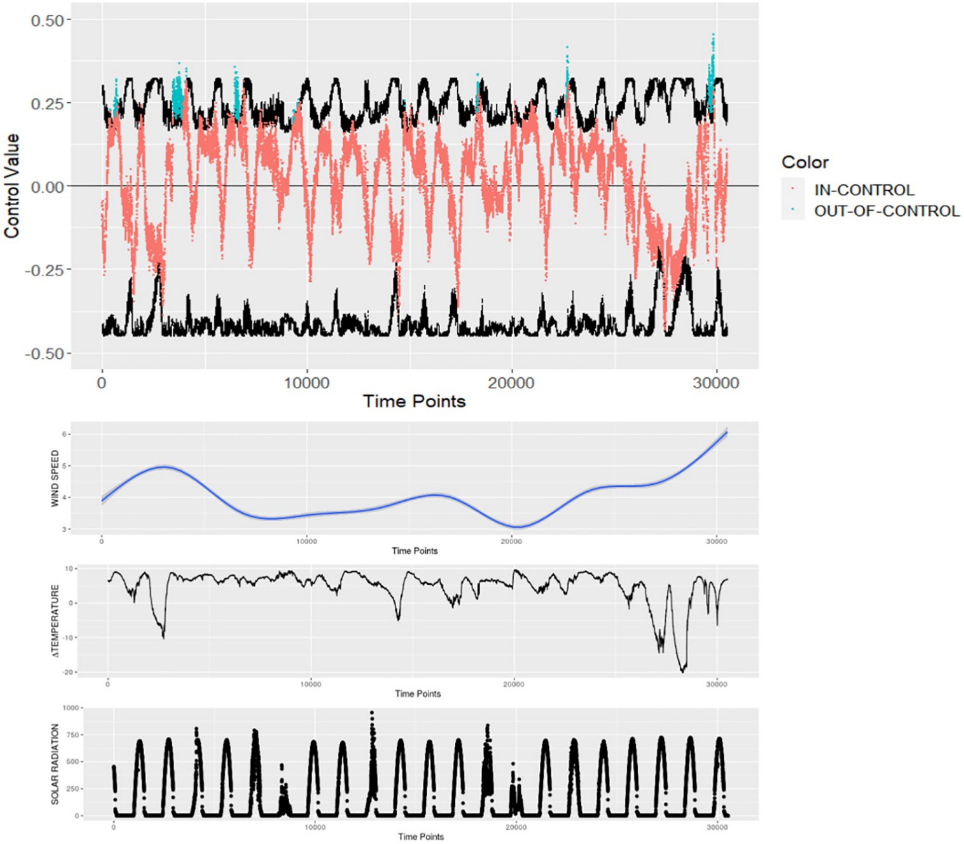

Fig 9. UL regression control chart (Phase II) adopting the residual representation (top panel), and the covariates' dynamic plots (lower panels).

https://doi.org/10.1371/journal.pone.0275841.g009

Table 6. Performance comparison of the selected regression models using AIC, BIC and RMSE (Phase I data).

	AIC	BIC	RMSE
UL	-1,296,760	-1,296,656	0.129
Beta 1	-1,727,663	-1,727,603	0.124
Beta 2	-1,777,595	-1,777,499	0.126

https://doi.org/10.1371/journal.pone.0275841.t006

when this suddenly discrepant water-related event started (top graphic, in blue colors), and the covariates' response during this event (lower graphics).

We also compared the performance of the proposed approach with the one presented in [28]. That is, we also built control charts based on the beta regression model with constant and varying dispersion parameter (hereafter, beta 1 and beta 2 models, respectively). Note that the beta 1 model was already considered in Section 4. First, a performance comparison of the three models (UL, beta 1 and beta 2), in terms of Akaike Information Criterion (AIC) [39], Bayesian or Schwarz Information Criterion (BIC) [40], and Root Mean Squared Error (RMSE), is presented in Table 6. Observe that these models are practically equivalent in terms of RMSE, although considering the AIC and BIC criteria, the beta 2 model performs the best (i.e., it provides the best fit). The complete estimation results for the beta models, as well as the corresponding control charts, are shown in S2 Appendix.

Taking a closer look into the predictive statistics (RMSE), for Phase II, the UL model showed 1,221 (4%) of out-of-control observations, whereas the beta 1 model showed 15 out-of-control observations (0.0005%) and 0 (0%) out-of-control observations adopting the beta 2 model. An important effect is that the observed month in Phase II is June (transition from Autumn to Winter season), which is characterized by lower temperatures through high changes in solar radiation, wind speed, and temperature variation. Therefore, further investigations can be performed towards verification of the quality of the adjusted models considering the time frequency per minute.

In addition, Table 7 shows the computational cost required to adjust and summarize the three models, considering Phase I data. Observe that the UL model showed to be three times faster than the beta 1 model, and five times faster than the beta 2 model. Therefore, the proposed UL model-based SPC approach has an advantage over the existing beta models-based SPC approach, in terms of the computational time and space required.

Furthermore, a lower Random-Access Memory (RAM) consumption is required by the flexible although simpler (i.e., with fewer parameters) UL model, allowing simple computers to process greater/large data sets. In the simulation studies, beta models took up over 20,000 times more memory space than the UL models. For the real data application, it was used a notebook with AMD Ryzen 5 3500U and 12 GB DDR4 RAM.

6 Conclusion

Learning structures are often used to describe association across variables. Nonetheless, inferential procedures are required to guarantee robustness on the analysis. For instance, the OLS

Table 7. Computational cost (time and memory space) required for each regression model (Phase I data).

	Fit (sec)	Model Summary (sec)	RAM Memory (bytes)
UL	108	0.01	60,048
Beta 1	324	213	1,149,182,240
Beta 2	456	374.4	1,207,127,200

sec = seconds.

https://doi.org/10.1371/journal.pone.0275841.t007

regression, with normal assumption for the response variable (*Y*). This is not the case of proportional/rate data, which present truncated and skewed information in a bounded (0, 1) interval. Therefore, this study considered a flexible UL regression model, which can model symmetric, right- and left-asymmetric truncated unit data, despite having a single parameter.

Variables integrated in the current study were previously chosen by other authors in the study of fog/humidity distribution. [6] developed a method for calculating diurnal patterns of air temperature, wind speed, global radiation and relative humidity, and validated it with data from different countries. Some other studies, as [41], have also demonstrated the relevance of scheduling appropriately the sampling frequency of climatic variables, in order to adequately estimate land surface fluxes. A study based on solar radiation, air temperature, relative humidity and dew point [42], daily and monthly reported over a year, has revealed the minimum relative humidity coinciding with the driest month of the year.

The process of reference evapotranspiration (ET) calculation is commonly estimated through Penman-Monteith ET [43]. This equation, based on the original [44]'s equation, determines evaporation based on the combination of energy balance and aerodynamic formula; and the [45]'s modification, that includes the surface resistance denominator. Finally, the FAO adapted the formula for crops [46]. This ET estimating reference has previously used daily weather forecast [47]. The solar radiation provides energy to vaporize water and heat up the atmosphere and ground. So, a day scale monitoring for wind speed, temperature, and solar radiation was used to describe the relative humidity in the air. Accordingly, to the FAO formulation, for hourly periods, the soil heat flux (G) can be daylight periods estimated with net radiation (0.1*Rn) for night-time (0.5*Rn).

Occurrence and distribution of *Camanchaca* along northern Chile had been previously described [48–51]. It is characterized by diurnal and interannual variability with dependence on atmospheric conditions at regional and global scales [52]. It is necessary to highlight that flora and fauna distribution in the arid area is fog-dependent [53]. Therefore, this fog plays a key role in maintaining the assemblage of animal species of the ecosystem, especially during adverse climatic periods. But it also supposes an important water resource for human settlement [48].

Results from the current study agree with [54]'s study, which determined that *Camanchaca* derived from the marine inversion layer from the Atacama Desert was more persistent, though weaker, during summer months (November-March), but greater condensed and shallower in winter months, with uncharacteristically dry air and high temperatures occurring at and above 400 m above sea level. The authors explained that the stability of the temperature inversion depends on a seasonal consistent high-humidity, onshore breeze. On the other hand, diurnal variations in wind speed and direction and moisture content and temperature, show that, during summer, there is almost no offshore breeze and that the humidity of the air mass over that site is nearly constant. However, the land-sea breeze cycle is enhanced in winter, in a way that there is considerable diurnal variation in specific humidity, correlated with the night-time breeze from the inland desert. But when night falls, wind begins to blow from the east, which lowers the atmospheric humidity. At day break, winds shift to the west and humidity rises as marine air moves east. In other words, winter-fog is more intense and shallower, in comparison to summer-fog.

It is important to highlight that, as a model, it is subjected to limitations. So, further statistical models should include some interannual variations or distinguish patterns affected by the climatic conditions, such as the ENSO (El Niño-Southern Oscilation); for example, La Niña conditions promote a lower cloud amount [52]. In this manner, SPC models can assess the weather monitoring, whenever its suppositions are carefully adopted. Further studies shall explore more the data structure dependence in the statistical inference procedure, such as spatial-temporal memory.

Supporting information

S1 Appendix. Graphical visualization of simulation results.
(PDF)

S2 Appendix. Beta regression models' estimation results and control charts.
(PDF)

Author Contributions

Conceptualization: Paulo H. Ferreira, Diego C. Nascimento.

Formal analysis: Anderson O. Fonseca.

Funding acquisition: Diego C. Nascimento.

Investigation: Anderson O. Fonseca.

Methodology: Paulo H. Ferreira, Anderson O. Fonseca, Diego C. Nascimento.

Project administration: Paulo H. Ferreira, Francisco Louzada.

Software: Anderson O. Fonseca.

Supervision: Paulo H. Ferreira, Francisco Louzada.

Validation: Estefania Bonnail.

Visualization: Anderson O. Fonseca.

Writing – original draft: Paulo H. Ferreira, Anderson O. Fonseca, Diego C. Nascimento.

Writing – review & editing: Estefania Bonnail, Francisco Louzada.

References

1. Fonseca A, Ferreira PH, Nascimento DC, Fiaccone R, Ulloa-Correa C, García-Piña A, et al. Water Particles Monitoring in the Atacama Desert: SPC Approach Based on Proportional Data. Axioms. 2021; 10 (3):154. https://doi.org/10.3390/axioms10030154
2. Mazucheli J, Menezes A, Dey S. The unit-Birnbaum-Saunders distribution with applications. Chilean Journal of Statistics. 2018; 9(1):47–57.
3. Ho LL, Fernandes FH, Bourguignon M. Control charts to monitor rates and proportions. Quality and Reliability Engineering International. 2019; 35(1):74–83. https://doi.org/10.1002/qre.2381
4. Bantan RAR, Chesneau C, Jamal F, Elgarhy M, Tahir MH, Ali A, et al. Some new facts about the unit-Rayleigh distribution with applications. Mathematics. 2020; 8(11):1954. https://doi.org/10.3390/math8111954
5. Bakouch HS, Nik AS, Asgharzadeh A, Salinas HS. A flexible probability model for proportion data: Unit-half-normal distribution. Communications in Statistics: Case Studies, Data Analysis and Applications. 2021; 7(2):271–288.
6. Aher S, Shinde S, Guha S, Majumder M. Identification of drought in Dhalai river watershed using MCDM and ANN models. Journal of Earth System Science. 2017; 126(2):1–4. https://doi.org/10.1007/s12040-017-0795-1
7. Mazucheli J, Menezes AFB, Chakraborty S. On the one parameter unit-Lindley distribution and its associated regression model for proportion data. Journal of Applied Statistics. 2019; 46(4):700–714. https://doi.org/10.1080/02664763.2018.1511774
8. Grosjean M, Veit H. In: Huber UM, Bugmann HKM, Reasoner MA, editors. Water Resources in the Arid Mountains of the Atacama Desert (Northern Chile): Past Climate Changes and Modern Conflicts. Dordrecht: Springer Netherlands; 2005. p. 93–104. Available from: https://doi.org/10.1007/1-4020-3508-X_10.
9. Bull AT, Andrews BA, Dorador C, Goodfellow M. Introducing the Atacama desert. Springer; 2018.

10. Bonnail E, Lima RC, Turrieta GM. Trapping fresh sea breeze in desert? Health status of Camanchaca, Atacama's fog. Environmental Science and Pollution Research. 2018; 25(18):18204–18212. https://doi.org/10.1007/s11356-018-2278-6 PMID: 29797192

11. García A, Ulloa C, Amigo G, Milana JP, Medina C. An inventory of cryospheric landforms in the arid diagonal of South America (high Central Andes, Atacama region, Chile). Quaternary International. 2017; 438:4–19. https://doi.org/10.1016/j.quaint.2017.04.033

12. Hoffman DD. Visual Intelligence: How We Create What We See. 1st ed. New York, NY, USA: W. W. Norton and Company, Inc.; 1998.

13. Chi EH. A framework for Visualization Information. Springer; 2002.

14. Telea AC. Data visualization: principles and practice. 2nd ed. CRC Press; 2014.

15. Ward M, Grinstein G, Keim D. Interactive Data Visualization: Foundations, Techniques, and Applications. 1st ed. Natick, MA, USA: A. K. Peters, Ltd.; 2010.

16. Silva SF, Catarci T. Visualization of Linear Time-Oriented Data: A Survey. In: Proceedings of the First International Conference on Web Information Systems Engineering. vol. 1 of WISE'00. Washington, DC, USA: IEEE Computer Society; 2000. p. 310–319.

17. Thakur S, Hanson AJ. A 3D Visualization of Multiple Time Series on Maps. In: Proceedings of the 2010 14th International Conference Information Visualisation. IV'10. Washington, DC, USA: IEEE Computer Society; 2010. p. 336–343.

18. Aigner W, Miksch S, Schumann H, Tominski C. Visualization of Time-Oriented Data. 1st ed. New York, NY, USA: Springer Publishing Company, Incorporated; 2011.

19. Kehrer J, Hauser H. Visualization and Visual Analysis of Multifaceted Scientific Data: A Survey. IEEE Transactions on Visualization and Computer Graphics. 2013; 19(3):495–513. https://doi.org/10.1109/TVCG.2012.110 PMID: 22508905

20. McLachlan P, Munzner T, Koutsofios E, North S. LiveRAC: Interactive Visual Exploration of System Management Time-series Data. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI'08. New York, NY, USA: ACM; 2008. p. 1483–1492.

21. Ward MO, Grinstein G, Keim D. Interactive data visualization: foundations, techniques, and applications. 2nd ed. CRC Press; 2015.

22. Keim DA. Information visualization and visual data mining. IEEE transactions on Visualization and Computer Graphics. 2002; 8(1):1–8. https://doi.org/10.1109/2945.981847

23. Alexandrina EC, Ortigossa ES, Lui ES, Gonçalves JAS, Corrêa NA, Nonato LG, et al. Analysis and visualization of multidimensional time series: Particulate matter (PM_{10}) from São Carlos-SP (Brazil). Atmospheric Pollution Research. 2019; 10(4):1299–1311. https://doi.org/10.1016/j.apr.2019.03.001

24. Abramowitz M, Stegun IA. Handbook of mathematical functions with formulas, graphs, and mathematical tables. vol. 55. US Government Printing Office; 1964.

25. Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ, Knuth DE. On the LambertW function. Advances in Computational Mathematics. 1996; 5(1):329–359. https://doi.org/10.1007/BF02124750

26. Bazán J, Torres-Avilés F, Suzuki A, Louzada F. Power and reversal power links for binary regressions: An application for motor insurance policyholders. Applied Stochastic Models in Business and Industry. 2017; 33(1):22–34. https://doi.org/10.1002/asmb.2215

27. Nocedal J, Wright SJ. Numerical Optimization. 2nd ed. Springer Science & Business Media; 2006.

28. Bayer FM, Tondolo CM, Müller FM. Beta regression control chart for monitoring fractions and proportions. Computers & Industrial Engineering. 2018; 119:416–426. https://doi.org/10.1016/j.cie.2018.04.006

29. R Core Team. R: A Language and Environment for Statistical Computing; 2020. Available from: https://www.R-project.org/.

30. Saghir A, Lin Z. Control charts for dispersed count data: an overview. Quality and Reliability Engineering International. 2015; 31(5):725–739. https://doi.org/10.1002/qre.1642

31. Riaz M, Ajadi JO, Mahmood T, Abbasi SA. Multivariate mixed EWMA-CUSUM control chart for monitoring the process variance-covariance matrix. IEEE Access. 2019; 7:100174–100186. https://doi.org/10.1109/ACCESS.2019.2928637

32. Montgomery DC. Introduction to statistical quality control. John Wiley & Sons; 2020.

33. Schaffer JR, Kim M-J. Number of replications required in control chart Monte Carlo simulation studies. Communications in Statistics—Simulation and Computation. 2007; 36(5):1075–1087. https://doi.org/10.1080/03610910701539963

34. Lima-Filho LMA, Pereira TL, Souza TC, Bayer FM. Inflated beta control chart for monitoring double bounded processes. Computers & Industrial Engineering. 2019; 136:265–276. https://doi.org/10.1016/j.cie.2019.07.017

35. Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape (with discussion). Applied Statistics. 2005; 54(3):507–554.

36. Jensen WA, Jones-Farmer LA, Champ CW, Woodall WH. Effects of parameter estimation on control chart properties: A literature review. Journal of Quality Technology. 2006; 38(4):349–364. https://doi.org/10.1080/00224065.2006.11918623

37. Moraes D, Oliveira FLPd, Quinino RdC, Duczmal LH. Self-oriented control charts for efficient monitoring of mean vectors. Computers & Industrial Engineering. 2014; 75:102–115. https://doi.org/10.1016/j.cie.2014.06.008

38. Paroissin C, Penalva L, Pétrau A, Verdier G. New control chart for monitoring and classification of environmental data. Environmetrics. 2016; 27(3):182–193. https://doi.org/10.1002/env.2382

39. Akaike H. On entropy maximization principle. In: Krishnaiah, P.R. (ed.), Applications of Statistics. North-Holland, Amsterdam. 1977;27–41.

40. Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978; 6(2):461–464. https://doi.org/10.1214/aos/1176344136

41. Hupet F, Vanclooster M. Effect of the sampling frequency of meteorological variables on the estimation of the reference evapotranspiration. Journal of Hydrology. 2001; 243(3-4):192–204. https://doi.org/10.1016/S0022-1694(00)00413-3

42. Shrestha AK, Thapa A, Gautam H. Solar radiation, air temperature, relative humidity, and dew point study: Damak, Jhapa, Nepal. International Journal of Photoenergy. 2019; 2019. https://doi.org/10.1155/2019/8369231

43. Zotarelli L, Dukes MD, Romero CC, Migliaccio KW, Morgan KT. Step by step calculation of the Penman-Monteith Evapotranspiration (FAO-56 Method). Institute of Food and Agricultural Sciences University of Florida. 2010.

44. Penman HL. Natural evaporation from open water, bare soil and grass. Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences. 1948; 193(1032):120–145.

45. Monteith JL. Evaporation and environment. In: Symposia of the society for experimental biology. vol. 19. Cambridge University Press (CUP) Cambridge; 1965. p. 205–234.

46. Allen RG, Pereira LS, Raes D, Smith M. FAO Irrigation and drainage paper No. 56. Rome: Food and Agriculture Organization of the United Nations. 1998; 56(97):e156.

47. Cai J, Liu Y, Lei T, Pereira LS. Estimating reference evapotranspiration with the FAO Penman–Monteith equation using daily weather forecast messages. Agricultural and Forest Meteorology. 2007; 145(1-2):22–35. https://doi.org/10.1016/j.agrformet.2007.04.012

48. Schemenauer RS, Fuenzalida H, Cereceda P. A neglected water resource: The Camanchaca of South America. Bulletin of the American Meteorological Society. 1988; 69(2):138–147. https://doi.org/10.1175/1520-0477(1988)069%3C0138:ANWRTC%3E2.0.CO;2

49. Cereceda P, Schemenauer RS. The occurrence of fog in Chile. Journal of Applied Meteorology and Climatology. 1991; 30(8):1097–1105. https://doi.org/10.1175/1520-0450(1991)030%3C1097:TOOFIC%3E2.0.CO;2

50. Cereceda P, Osses P, Larrain H, Farías M, Lagos M, Pinto R, et al. Advective, orographic and radiation fog in the Tarapacá region, Chile. Atmospheric Research. 2002; 64(1-4):261–271. https://doi.org/10.1016/S0169-8095(02)00097-2

51. Larrain H, Velásquez F, Cereceda P, Espejo R, Pinto R, Osses P, et al. Fog measurements at the site "Falda Verde" north of Chañaral compared with other fog stations of Chile. Atmospheric Research. 2002; 64(1-4):273–284. https://doi.org/10.1016/S0169-8095(02)00098-4

52. Garreaud R, Barichivich J, Christie DA, Maldonado A. Interannual variability of the coastal fog at Fray Jorge relict forests in semiarid Chile. Journal of Geophysical Research: Biogeosciences. 2008; 113 (G4). https://doi.org/10.1029/2008JG000709

53. del Val E, Armesto JJ, Barbosa O, Christie DA, Gutiérrez AG, Jones CG, et al. Rain forest islands in the Chilean semiarid region: fog-dependency, ecosystem persistence and tree regeneration. Ecosystems. 2006; 9(4):598–608. https://doi.org/10.1007/s10021-006-0065-6

54. Thompson MV, Palma B, Knowles JT, Holbrook NM. Multi-annual climate in Parque Nacional Pan de Azúcar, Atacama Desert, Chile. Revista Chilena de Historia Natural. 2003; 76(2):235–254. https://doi.org/10.4067/S0716-078X2003000200009