Statistical modelling of space weather extremes and process monitoring of rates and proportions

Cristine Rauber Oliveira, B.Sc., M.Sc.

Submitted for the degree of Doctor of Philosophy at Lancaster University.

February 2024

Abstract

The first part of this work has been motivated by a multivariate space weather dataset. When the Sun releases high-energy particles into space due to a solar explosion, we experience magnetic storms, that once reaching Earth can last hours or days. Although solar flares emmited by the Sun cannot cause any harm to humans on Earth, if they are too severe they can damage machinery and techonology, such as satellites and radio communication. Thus, the modelling of extreme solar activity is important so we can be prepared for undesireable extreme events. Extreme value analysis can help professionals to understand the risks that severe geomagnetic field fluctuations can pose to Earth. For example, we can characterise the tail of the distribution of geomagnetic disturbances and the probability of extreme events. Hence, we perform a pairwise analysis for modelling the extremes of multiple bivariate processes of geomagnetic activity considering two copula models. The aim is to model the joint extremal probability and depict the pairwise extremal dependence structure between pairs of sites in two regions in Europe. The results show that the dependence structure differs in Northern and Southern Europe and that the dependence weakens as the distance increases.

The second part of this work proposes a control chart for detecting small shifts in the mean of a double-bounded process, such as fractions or proportions, in the presence of control variables. For this purpose, we consider the cumulative sum control chart applied to different residuals of the beta regression model. We conduct an extensive Monte Carlo simulation study to evaluate and compare the performance of the proposed

control chart with two other control charts in the literature in terms of run length analysis. The numerical results show that the proposed control chart is more sensitive to changes in the process than its competitors and that the quantile residual is the most suitable residual to be used in our proposal. Finally, based on the quantile residual, we present and discuss applications to real and simulated data to show the applicability of the proposed control chart.

Acknowledgements

Undertaking this work has been a life-changing experience for me and it would not have been possible without the support and guidance that I received from many people.

First and foremost, I would like to thank my supervisors, Emma Eastoe and Jennifer Wadsworth, for their help during my studies. I am also grateful to my defense committee for their thoughtful insights and suggestions, contributing significantly to the development of this work. My gratitude extends to the Department of Mathematics and Statistics for providing the funding that facilitated my studies at Lancaster University.

Furthermore, I wish to acknowledge with immense gratitude the guidance offered by Luiz Medeiros de Araújo Lima Filho and Fábio Mariano Bayer, whose collaboration enriched my research. Their belief in me and our friendship have been a constant source of motivation.

I am profoundly grateful to every individual who generously contributed to my crowdfunding campaign, easing the financial burden of my student visa. Your support has been invaluable.

I am extremely lucky to have a network of beautiful and loving friends who supported me in countless ways throughout my academic journey and life. While it's impossible to name you all, I want you to know how deeply I appreciate your presence and the countless memorable moments we have shared - whether through evenings out, video calls, or visits. Thank you for making my academic life less lonely. I extend special

thanks to Rafael and Natalie, who graciously opened their home and hearts to me upon my arrival in the United Kingdom.

Lastly, I would be remiss not to mention my mother, my steadfast pillar of support. Her unwavering love has been my constant source of strength and motivation. To my partner, I offer my heartfelt thanks for the unwavering support and continuous encouragement throughout this remarkable journey.

Declaration

I declare that the work in this thesis is my own, except where stated otherwise.

Chapter 1 is entirely devoted to the statistical modelling of space weather extremes using extreme value theory. In Section 1.1, we provide an introduction to space weather. Section 1.2 describes the data used in the analysis, whilst Section 1.3 gives a brief literature review of existing analyses of geomagnetic activity and indices. Section 1.4 introduces the bivariate copula models used to model the extremes of geomagnetic activity. Finally, in Section 1.5, we offer some concluding remarks.

In Chapter 2, we exclusively develop a new control chart methodology for double bounded processes. An introduction to the topic is given in Section 2.1, and a literature review is presented in Section 2.2. The new control chart is presented in Section 2.3. Section 2.4 provides an extensive Monte Carlo simulation study to evaluate the performance of the proposed control chart. In Section 2.5, we present and discuss applications to simulated and real data to show the applicability of the proposal. Finally, in Section 2.6, we offer some conclusions.

Chapter 2 is a joint work with Luiz Medeiros de Araújo Lima Filho and Fábio Mariano Bayer, and has been published in the *Quality and Reliability Engineering International* journal on 9$^{\text{th}}$ September, 2022 (`https://doi.org/10.1002/qre.3140`). I would also like to thank two anonymous referees for their thoughtful comments and suggestions which improved the quality of the work in this chapter.

All the computing code is my own and was carried out using the `R` statistical com-

puting environment (R Core Team, 2022) for the Linux operating system. All figures presented in this thesis were produced using `R` base or the library `ggplot2` (Wickham, 2016). This thesis was typeset using LaTeX. In Chapter 1, the main `R` libraries used are `ismev` and `evd` written by Alec Stephenson, and `SpatialADAI` written by Jennifer L. Wadsworth and Raphaël Huser. In Chapter 2, I used the `R` libraries `betareg` written by Francisco Cribari-Neto and Achim Zeileis, `qcc` written by Luca Scrucca, `extraDistr` written by Tymoteusz Wolodzko, `rattle` written by Graham Williams, and `lubridate` written by Garrett Grolemund and Hadley Wickham.

Cristine Rauber Oliveira

Contents

List of Figures

List of Tables

List of Abbreviations

CUSUM	Cumulative Sum
SD	Standard Deviation
SE	Standard Error
CI	Confidence Interval
AD	Asymptotically Dependent
AI	Asymptotically Independent
QQ	Quantile-Quantile
CDF	Cumulative Distribution Function
PDF	Probability Density Function
HW	Huser-Wadsworth
EVT	Extreme Value Theory
GIC	Geomagnetically Induced Currents
POT	Peaks-Over-Threshold
MLT	Magnetic Local Lime
GEV	Generalised Extreme Value
GP	Generalised Pareto
PIT	Probability Integral Transform
SPC	Statistical Process Control
MLE	Maximum Likelihood Estimator
EWMA	Exponentially Weighted Moving Average

BRCC	Beta Regression Control Chart
BCC	Beta Control Chart
CL	Center Line
UCL	Upper Control Limit
LCL	Lower Control Limit
ARL	Average Run Length
MRL	Median Run Length
SDRL	Standard Deviation Run Length
RH	Relative Humidity
IAGA	International Association of Geomagnetism and Aeronomy

List of Symbols

$\max(\cdot)$	Maximum.
$\mathbb{E}[\cdot]$	Expectation.
$\mathrm{Var}(\cdot)$	Variance.
$\phi(\cdot)$	Gaussian probability density function.
$\Phi^{-1}(\cdot)$	Gaussian quantile function.
$\Phi(\cdot)$	Gaussian cumulative distribution function.
$V(\cdot)$	Variance-Covariance matrix.
$\mathbb{1}(\cdot)$	Indicator function.
$L(\cdot)$	Likelihood function.
$\ell(\cdot)$	Log-likelihood function.
$\mathbb{P}(\cdot)$	Probability.
$F(\cdot)$	Distribution function.
$\tilde{F}(\cdot)$	Empirical distribution function.
$C(\cdot,\cdot)$	Copula function.
$\bar{C}(\cdot,\cdot)$	Joint survivor function.
$f(\cdot)$	Density function.
n	Sample size.
$\sup\{\cdot\}$	Supremum.
$\min(\cdot)$	Minimum.
d	Dimension.

u	Threshold.
h	Distance between two locations.
Σ	Correlation matrix.
$\boldsymbol{\Theta}$	Parameter vector.
$\boldsymbol{\psi}$	Parameter vector.

Chapter 1

Statistical modelling of space weather extremes

1.1 Introduction

Space weather can be described as the result of interactions between the behaviour of the sun and the Earth's magnetic field and atmosphere. During periods of high solar activity, large quantities of energised particles are released from the sun into Earth's magneto- and ionospheres at high speeds, causing instability in these parts of the atmosphere. These solar storms can miss Earth completely, for example, Baker et al. (2013) states that the powerful solar storm that occurred on 23rd July 2012 would have had devastating consequences if it was Earth-directed (Ngwira et al., 2013). However, when they do strike Earth in a direct hit, they have the potential to affect conditions in the Earth's atmosphere, and sometimes, on the Earth's surface. As Earth's magnetic field redirects the particles towards the polar caps, the astonishing phenomenon known as auroras (Northern and Southern lights) starts to form. The red, green, and pink lights flying on the sky at high or low latitudes are the most well-known and visible effect of intense solar disturbances, and although it fascinates skywatchers, the effects of ex-

treme space weather can be threatening to infrastructure, technology, communications systems, railway signalling systems, pipelines, and personal health (Boteler et al., 1998; Trichtchenko and Boteler, 2001; Wik et al., 2009; Eroshenko et al., 2010; Roy and Paul, 2013; Thaduri et al., 2020; Boteler, 2021).

The study of the sun-Earth interaction has gained attention in the last few decades due to a series of events that took place in the past. The largest known solar-induced disturbance that affected Earth directly was the solar storm of September 1859 (Green and Boardsen, 2006; Cliver, 2006). This event was first observed by Carrington (1859) and Hodgson (1859) and 17 hours later the Earth experienced a huge, red auroral display from a coronal mass ejection, which was visible from within 23° of the geomagnetic equator in both Northern and Southern hemispheres (Kimball, 1960).

The aurora of 1859 is known to be the first observation of a solar flare (Cliver and Svalgaard, 2004). Other massive solar storms happened on August 4^{th} , 1972 (Anderson et al., 1974) and March 13^{th} , 1989 (Allen et al., 1989; Czech et al., 1992; Bolduc, 2002). The major geomagnetic disturbance in 1972 caused an outage of a communication cable system on the Plano, Illinois, to Cascade, Iowa, and geomagnetically induced currents (GIC) from the extreme solar event in 1989 caused the failure of the Hydro-Quebec power system and a blackout in the Quebec Province that lasted up to nine hours. Besides millions of people going without energy power that morning, many other consequences on and near-Earth were reported, including out-of-control satellites for several hours worldwide and difficulties in high-frequency radio communication. Geomagnetic disturbances like the ones that happened in 1972 and 1989 are rare, so learning about space weather is vital to prepare us for the next large geomagnetic storm when it arrives.

As the impacts caused by GIC only occur when there are extreme changes in the electromagnetic field, extreme value models can help to characterise the behaviour of extreme space weather events and also describe the extremal dependence structure

between measurements from ground-based magnetometers.

Among the several ways to measure changes in the geomagnetic field, we shall study the absolute change in time of the horizontal component in the Earth's geomagnetic field, known as $|\,\mathrm{d}B_H/\mathrm{d}t\,|$. This work explains the extremal dependence structure of daily maximum absolute one-minute changes in $\mathrm{d}B_H/\mathrm{d}t$ through a pairwise approach. This type of analysis is useful because we can estimate the probability of joint events for two given locations and obtain a partial summary of the spatial extremal dependence, which is not possible to do marginally. However, at the same time that a pairwise approach is simple, yet powerful, it does not allow for global dependence inference and interpolation, which would require a full spatial model applied to all sites together.

To the best of our knowledge, there is no work in the literature aimed to describe the pairwise extremal behaviour of geomagnetic activity. In this regard, the chief contribution of this work is to measure the tendency for large geomagnetic field fluctuations at separate locations to occur simultaneously and the strength of extremal dependence in the joint upper tail as the geodesic distance between sites is increased.

1.2 Space weather data

Data on space weather events is currently available through the SuperMAG initiative, a project led by Johns Hopkins University (Gjerloev, 2009), which consists of a collection of organisations and national agencies that work together to provide data on Earth's geomagnetic field to study the behaviour of the ionospheric and magnetospheric current systems. The project currently operates more than 300 ground-based magnetometers worldwide and full details on the extreme value theory (EVT) data processing can be found in Gjerloev (2012).

In this work, we shall study the extreme behaviour of geomagnetic field fluctuation measurements from the stations in Figure 1.2.1 (information on the derivation of the

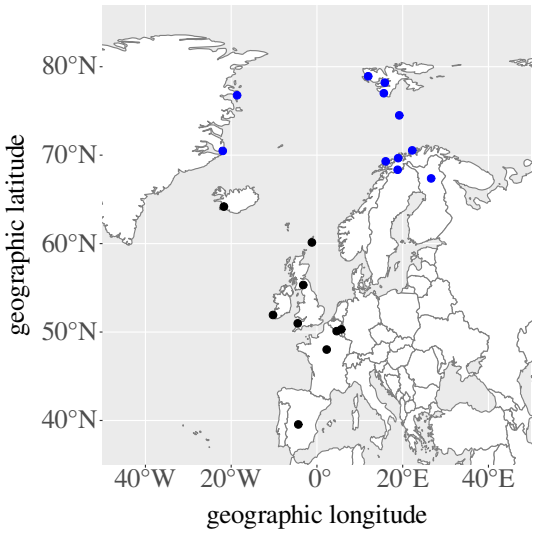


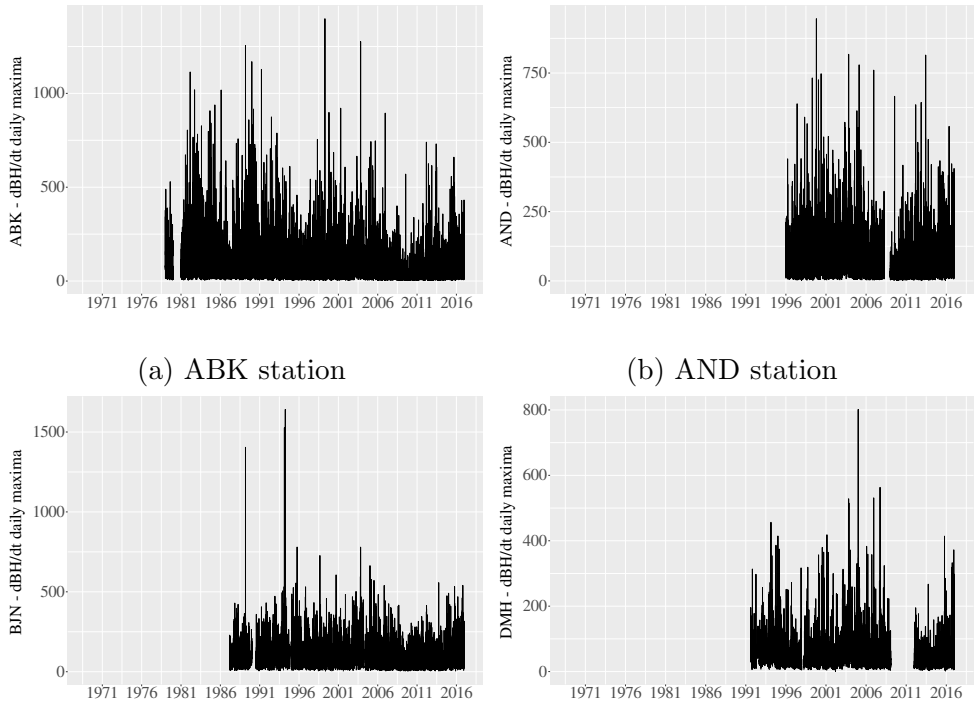
Figure 1.2.1: Location of the SuperMAG sites considered in this work; blue dots represent the Northern dataset and black dots the Southern dataset.

measurements can be found in Rogers et al. (2020)). Table 1.2.1 gives the International Association of Geomagnetism and Aeronomy (IAGA) code of the observatories, location and geographic coordinates. This region comprises sites in Northern and Southern Europe, totaling 20 observatories covering a range of latitudes. Blue dots represent the Northern dataset, where we have stations in the auroral ring ($65°\text{N} - 70°\text{N}$) and north pole ($70°\text{N} - 90°\text{N}$) zones. In contrast, the Southern dataset (black dots) has stations in the subauroral zone ($60°\text{N} - 65°\text{N}$) and lower latitudes ($39°\text{N} - 60°\text{N}$).

To illustrate the nature of the space weather data we studied in this work, Figures 1.2.2 and 1.2.3 provide some plots of the original data for select observatories in the Northern and Southern datasets. We observe some gaps in the data when considering the whole span period, and the extent of missingness varies across the datasets. This variability can lead to biased estimates of extreme events, making it difficult to predict the impact of future space weather rare events.

Table 1.2.1: IAGA code, location and geographic coordinates of the 20 sites considered in this work.

IAGA code	Location	Latitude	Longitude
ABK	Abisko, Sweden	68.35	18.82
AND	Andenes, Norway	69.30	16.03
BJN	Bjørnøya, Svalbard	74.50	19.20
CLF	Chambon-la-forêt, France	48.02	2.27
DMH	Danmarkshavn, Greenland	76.77	−18.63
DOU	Dourbes, Belgium	50.10	4.60
ESK	Eskdalemuir, Scotland	55.32	−3.20
HAD	Hartland, England	50.98	−4.48
HRN	Hornsund, Svalbard	77.00	15.60
LER	Lerwick, Scotland	60.13	−1.18
LRV	Leirvogur, Iceland	64.18	−21.70
LYR	Longyearbyen, Svalbard	78.20	15.83
MAB	Manhay, Belgium	50.30	5.68
NAL	Ny Ålesund, Svalbard	78.92	11.95
SCO	Ittoqqortoormiit, Greenland	70.48	−21.97
SOD	Sodankylä, Finland	67.37	26.63
SOR	Sørøya, Norway	70.54	22.22
SPT	San Pablo Toledo, Spain	39.55	−4.35
TRO	Tromsø, Norway	69.66	18.94
VAL	Valentia, Ireland	51.93	−10.25



(a) ABK station

(b) AND station

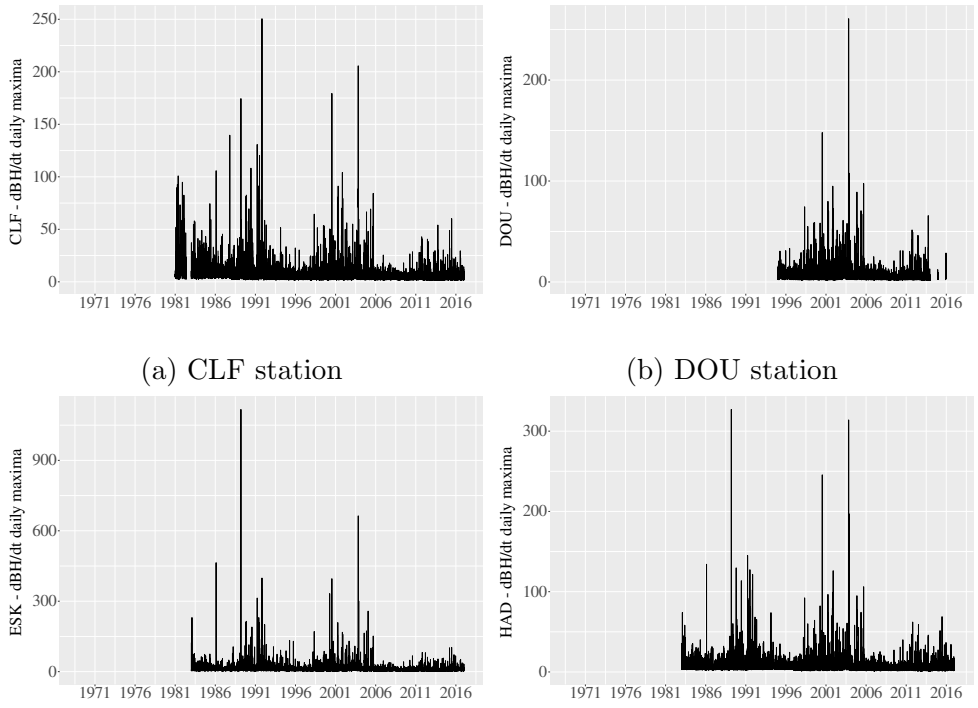
(c) BJN station

(d) DMH station

Figure 1.2.2: Daily maxima geomagnetic field fluctuation measurements from 1969 to 2016 at stations ABK, AND, BJN, and DMH from the Northern dataset.

1.3 On the literature review of space weather extremes modelling

In the literature, there is limited work regarding extreme value analysis of the SuperMAG data and those available only apply univariate modelling approaches. The first attempts to apply extreme value analysis to geomagnetic data used datasets comprising geomagnetic indices, rather than raw activity measurements. Tsubouchi and Omura (2007) and Silrergleit (1996) modelled the absolute value of the disturbance storm time (D_{st}) index proposed by Sugiura (1964). Such an index gives information on the geomagnetic activity at four stations near Earth's equator, and is used to analyse the strength and duration of geomagnetic storms. Values less than -50 nanotesla suggests high geomagnetic activity. Siscoe (1976) analysed the average "half-daily" *aa* index compiled by Mayaud (1973), which represents the geomagnetic activity level at an in-



(a) CLF station (b) DOU station

(c) ESK station (d) HAD station

Figure 1.2.3: Daily maxima geomagnetic field fluctuation measurements from 1969 to 2016 at stations CLF, DOU, ESK, and HAD from the Southern dataset.

variant latitude of about 50 degrees. Silrergleit (1999) studied the maximum average 24–hour global disturbance AA^* index, derived from the *aa* index, to predict when the next large geomagnetic storms could occur. Koons (2001) studied the annual maxima of the magnetic index A_p, which is a measure of the geomagnetic activity level over the world for a given day. Although these indices were designed to measure geomagnetic activity of the Earth's ionized environment field caused by irregular current systems, they usually underestimate large geomagnetic storms because of the inadequate distribution of the stations or insufficient observatories to compute the indices (Kozyreva et al., 2018). Thus, analysing geomagnetic activity directly can give more realiable estimates of the probability of occurence of large geomagnetic field variations and is more relevant to the assessment of space weather hazard to Earth and technology.

The first attempt to use EVT to analyse geomagnetic measurements was by Thomson et al. (2011). They analysed geomagnetic activity from 28 European sites, across

a range of latitudes, by fitting the generalised Pareto (GP) distribution to one-minute geomagnetic time series of the horizontal field and declination. A reasonable threshold chosen by the authors was the 99.97^{th} quantile for each station and they also used a declustering method to separate clustered extremes before fitting the distributions. Using the fitted distributions, they predicted return levels for each site for return periods as long as 200 years. In conclusion, both measured and extrapolated extreme values generally increase with latitude.

Wintoft et al. (2016) also considered a subset of European sites, analysing extremes of geomagnetic activity by fitting a generalised extreme value (GEV) distribution to the annual maxima. Their results showed that at higher latitudes, stations present higher probability of large values of geomagnetic perturbances compared to stations in lower latitudes. In addition, they also found that the tail distribution of observatories in high latitudes decays to zero more quickly and that this transition occurres around $59°\text{N} - 61°\text{N}$ latitudes.

More recently, Rogers et al. (2020) used univariate EVT to model the probability of occurrences of $|\,\mathrm{d}B_H/\mathrm{d}t\,|$ from 125 sites across the globe. The authors fitted the GP distribution to observations exceeding the 99.97^{th} quantile, presented distribution-based predictions of return levels for return periods ranging from 5 to 500 years, and examined the probability of large values of geomagnetic field fluctuation as a function of month, magnetic local time (MLT), and the direction of the fluctuation. The authors findings state that the occurrence of large geomagnetic field fluctuations is strongly dependent on latitude and MLT. For example, sites in the auroral zone are more likely to experience extreme values at 0300-1100 MLT. The use of MLT in distributions of the occurrence probability of extreme geomagnetic measurements is useful to refine return levels of estimates of $|\,\mathrm{d}B_H/\mathrm{d}t\,|$ for operations limited to certain times of the day.

In addition, Rogers et al. (2021) studied extreme geomagnetic fluctuations over a range of periods in a smaller timescale, from 1 to 60 minutes, to understand the causes

and impacts of GIC on these fluctuations. The authors described the variations in the geomagnetic field with geomagnetic latitude and MLT. The GP tail distribution was fitted to the data above the 99.97^{th} quantile and return levels were predicted for both the ramp changes and the root-mean-square of fluctuations over the periods.

1.4 Pairwise modelling of space weather extremes

EVT is widely used in many environmental and geophysical contexts where the interest is in the tails of the distribution rather than in the body. However, events occurring in the tails are, by nature, rare, often leading to a characterisation of the tail behaviour based on a few data points. EVT provides asymptotic distributions so we are able to extrapolate beyond the largest observations and predict extreme events. Univariate EVT describes the tail behaviour of univariate distributions and motivates a statistical distributions for a single response, for example, analysing extremes of geomagnetic activity from a single site in a particular location as in Rogers et al. (2020). Multivariate EVT involves the analysis of the joint tail behaviour of two or more random variables, such as analysing the joint distribution of geomagnetic field fluctuations from two or more observatories in different locations.

Before undertaking a full spatial analysis, a pairwise analysis of each pair of sites in a pre-determined region can help to obtain a general summary of the dependence structures that exist in that area. This information helps to understand in more detail the underlying process, such as spatial non-stationarity and the degree of dependence between sites in different latitudes, so that appropriate spatial models can be identified and extrapolation is more reliable. Thus, in this work, we shall focus on the special case of a bivariate model fitted to the pairwise extremes of geomagnetic field fluctuations in Europe for several pairs of sites. In this section, we introduce the copula function, the extremal dependence measure often applied in exploratory bivariate ex-

treme analyses, and two copula models - the bivariate Gaussian copula and the bivariate Huser-Wadsworth (HW) copula (Huser and Wadsworth, 2019). The copula models are outlined in Section 1.4.2.

1.4.1 Extremal dependence

Modelling the extremal dependence structure between random variables in a multivariate analysis is made complex when the variables do not have common marginal distributions. One useful and well-known method that can be used to depict such structures is the copula function, which places all variables on common margins.

The copula function

Let F_{Y_1} and F_{Y_2} be the cumulative distribution functions (CDFs) of two random variables Y_1 and Y_2, respectively. By the probability integral transform (PIT), the random variables $U_1 = F_{Y_1}(Y_1)$ and $U_2 = F_{Y_2}(Y_2)$ each follow a uniform distribution on the unit interval $[0, 1]$. Given $(u_1, u_2) \in [0, 1]^2$, the copula function is defined as the joint distribution of U_1 and U_2, i.e.

$$\begin{aligned} C(u_1, u_2) &= \mathbb{P}\left(U_1 \leq u_1, U_2 \leq u_2\right) \\ &= \mathbb{P}\left(F_{Y_1}(Y_1) \leq u_1, F_{Y_2}(Y_2) \leq u_2\right). \end{aligned}$$

The definition above motivates the following theorem.

Theorem 1.4.1 (Sklar's theorem (Sklar, 1959))**.** *Consider a two-dimensional joint distribution function F, with marginal distributions F_{Y_1} and F_{Y_2}, where $F(y_1, y_2) = \mathbb{P}\left(Y_1 \leq y_1, Y_2 \leq y_2\right)$. Then, there exists a copula C such that*

$$F(y_1, y_2) = C(F_{Y_1}(y_1), F_{Y_2}(y_2)). \tag{1.4.1}$$

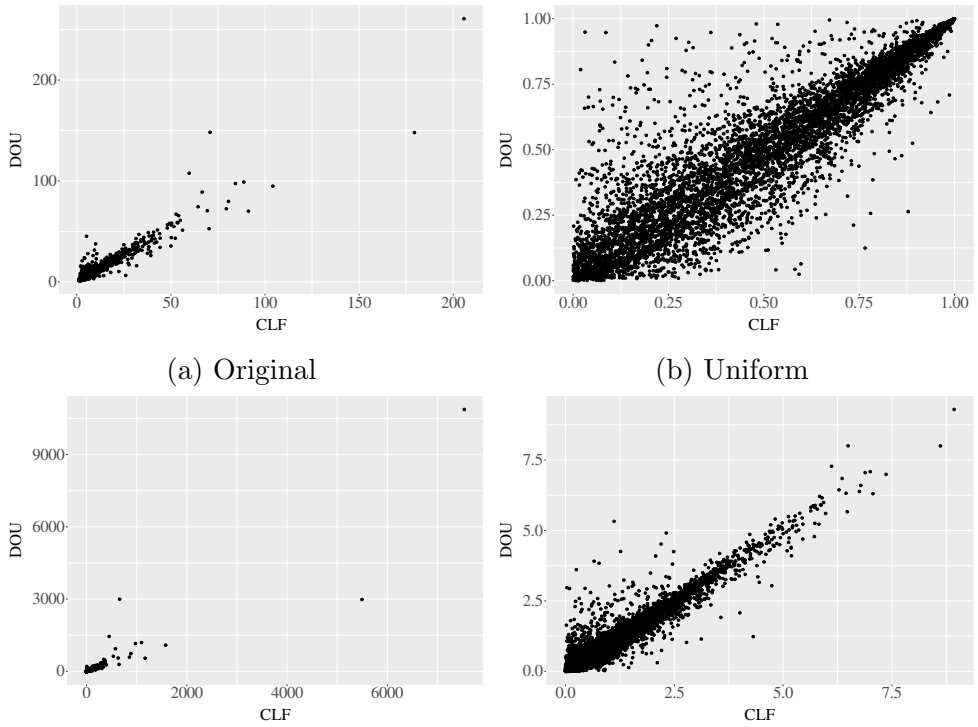
If F_{Y_1} and F_{Y_2} are continuous, then the copula C is unique. Conversely, if we have a copula $C : [0,1]^2 \to [0,1]$ and marginals $F_{Y_1}(y_1) = F(y_1, \infty)$ and $F_{Y_2}(y_2) = F(\infty, y_2)$, then $F(y_1, y_2)$ in (1.4.1) is a bivariate CDF with marginals F_{Y_1} and F_{Y_2}.

It is noteworthy that the copula function can be extended to d-dimensional distributions. For independent random variables, we have $C(u_1, u_2) = u_1 u_2$, whilst if the random vector (Y_1, Y_2) is perfectly dependent, we have $C(u_1, u_2) = \min(u_1, u_2)$. Further details about copula theory and multivariate models can be found in Nelsen (2007) and Joe (1997).

As a simple example, in Figure 1.4.1(a), geomagnetic field fluctuations from sites CLF (2.27°W, 48.02°N) and DOU (4.60°W, 50.10°N) are plotted on the original scale. In Figures 1.4.1(b), 1.4.1(c), and 1.4.1(d), the data have been transformed to three different marginal distributions, namely uniform, Frechét, and exponential, respectively. In order to transform the original data to uniform margins we used the peaks-over-threshold (POT) approach described in (1.4.7) to model exceedances above a 95^{th} quantile for each margin, and the marginal empirical distribution below this threshold. From the uniform margins we can then transform to any marginal distribution using the PIT. Certain choices of margins can be helpful for visualising the extremal dependence, e.g. Frechét and exponential.

Tail correlation

In EVT, the tails of the distributions are of most interest, and analysing their joint dependence structures is one important step that must be considered in multivariate extremes. Suppose we have two random variables, Y_1 and Y_2, and we want to know whether they are dependent in the tails. In the case of Y_1 and Y_2 being non-identically distributed, we first transform (Y_1, Y_2) to common standard uniform margins using the PIT. One possibility is then to study $\mathbb{P}(U_2 > u \mid U_1 > u)$ for large u.



(a) Original (b) Uniform

(c) Fréchet (d) Exponential

Figure 1.4.1: Daily maxima geomagnetic field fluctuation measurements from 1995 to 2015 at stations CLF in France (2.27°W, 48.02°N) and DOU in Belgium (4.60°W, 50.10°N). Original data (a), transformed to uniform (b), Fréchet (c), and exponential (d) margins using the PIT method.

More generally, considering a large value of u, we define

$$\chi = \lim_{u \to 1} \mathbb{P}\left(U_2 > u \mid U_1 > u\right), \tag{1.4.2}$$

where $0 \leq \chi \leq 1$ is called the tail correlation. The quantity χ represents the probability of one variable being extreme given that the other is extreme. We now define a sub-asymptotic version of (1.4.2) as

$$\begin{aligned} \chi(u) &= \mathbb{P}\left(U_2 > u \mid U_1 > u\right) \\ &= \frac{\mathbb{P}\left(U_1 > u, U_2 > u\right)}{\mathbb{P}\left(U_1 > u\right)} \\ &= \frac{\bar{C}(u,u)}{1-u}, \end{aligned} \tag{1.4.3}$$

for $0 \leq u \leq 1$, where $\bar{C}(\cdot, \cdot)$ is the joint survivor function of U_1 and U_2. It follows that $\chi = \lim_{u \to 1} \chi(u)$.

When $\chi > 0$, we say the variables are asymptotically dependent (AD) in the extremes, whereas $\chi = 0$ defines variables that are asymptotically independent (AI). The advantage of using χ is that it provides the relative strength of dependence for AD variables, with higher values of χ corresponding to stronger dependence in the joint extremes. The limitation of χ is that it does not discriminate between different strengths of extremal dependence for AI data ($\chi = 0$).

The limit in (1.4.2) cannot be estimated exactly from a finite sample size, so estimation of χ consists of examining the behaviour of $\chi(u)$ as $u \to 1$. The simplest estimation method uses the empirical distribution of (Y_{i1}, Y_{i2}), $i = 1, \ldots, n$, for which we obtain the empirical estimate of (1.4.3) as

$$\hat{\chi}(u) = \frac{\sum_{i=1}^{n} \mathbb{1}(\tilde{F}_{Y_2}(Y_{i2}) > u, \tilde{F}_{Y_1}(Y_{i1}) > u)}{\sum_{i=1}^{n} \mathbb{1}(\tilde{F}_{Y_1}(Y_{i1}) > u)},$$

where

$$\tilde{F}_{Y_1}(y_1) = \sum_{i=1}^{n} \frac{\mathbb{1}(Y_{i1} \leq y_1)}{n+1}, \quad \text{and} \quad \tilde{F}_{Y_2}(y_2) = \sum_{i=1}^{n} \frac{\mathbb{1}(Y_{i2} \leq y_2)}{n+1}.$$

When the focus is on the modelling of the dependence between the response at two or more spatial locations, we can also estimate $\chi(u)$ as a function of a distance. We define this as $\chi_h(u)$, where h is the distance between two locations s_1 and s_2, i.e. $h = |s_1 - s_2|$. In this spatial setting, the conditional probability distribution of U_1 and U_2 varies as a function of the distance h, and as well as analysing the behaviour of $\chi(u)$ for $u \to 1$, we are now also interested in examining the behaviour of $\chi_h(u)$ as h increases. Because we expect observations recorded farther apart to exhibit a weaker dependence, we usually observe $\chi_h(u)$ decreasing with h for a given u. By analysing

multiple pairs of sites in a region, the pairwise estimates viewed as a function of distance can be used to understand how the extremal dependence changes with distance.

1.4.2 Copula models

The extremal dependence measure $\chi_h(u)$ defined in Section 1.4.1 provides a description of the extremal dependence but does not give a basis for prediction or extrapolation. In a spatial context, such a measure helps us understand the dependence structure across a range of locations. However, for more complete inference on the extremal dependence we need to fit some models to the data. In this section, we introduce the bivariate Gaussian and the bivariate HW copula models which will subsequently be fitted to the extremes of the geomagnetic field fluctuations. One advantage of the bivariate HW model is that it can capture both asymptotic dependence and asymptotic independence, in contrast with the Gaussian copula which only captures asymptotic independence. The justification for using these models is that the Gaussian copula is nested in the HW copula, thus aiming to verify whether a more complex but flexible model is needed.

Bivariate Gaussian copula

In spatial applications, multivariate Gaussian distributions form the basic building block of many spatial models due to their attractive properties and mathematical tractability. The bivariate Gaussian distribution summarizes joint dependence behaviour through a correlation parameter. The study of how estimates of these correlation parameters evolve over space provides insights into a spatial dependence structure. Although often viewed as providing a good fit to the body of the data rather than the tails, we can still use a Gaussian model to explore extremal dependence by adapting it appropriately to focus only on those parts of the distribution where at least one variable is extreme, as in Bortot et al. (2000).

Let (Z_1, Z_2) be a pair of random variables following a standard bivariate Gaussian

distribution. The probability density function of (Z_1, Z_2) is

$$\phi_{Z_1,Z_2}(z_1, z_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{z_1^2 - 2\rho z_1 z_2 + z_2^2}{2\left(1-\rho^2\right)}\right\}, \tag{1.4.4}$$

where ρ is the correlation parameter and $-1 \leq \rho \leq 1$. The bivariate Gaussian copula is given by:

$$C(u_1, u_2) = \Phi_{Z_1,Z_2}\left[\Phi^{-1}(u_1), \Phi^{-1}(u_2); \Sigma\right], \tag{1.4.5}$$

where Φ^{-1} is the Gaussian quantile function and Φ_{Z_1,Z_2} is the joint bivariate distribution function of a Gaussian random variable with mean vector zero and correlation matrix

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{1.4.6}$$

In order to estimate the correlation ρ of the bivariate Gaussian copula to the extremes of geomagnetic field fluctuations, we adopt the censored likelihood approach used by Huser and Wadsworth (2019) and several earlier articles on extremal dependence modelling, which considers the full contribution of the values higher than a threshold only.

We assume that our data comprise n independent observations of a pair of random variables (Y_1, Y_2) that come from two different locations. The ith observation at the jth location is denoted by $Y_{ij}, i = 1 \ldots, n, j = 1, 2$. As the marginal distributions might not be the same and to ensure that they have the required exponential upper tails, we use a semi-parametric approach proposed by Keef et al. (2013) to transform the replicates at each station $s_j, j = 1, 2$. This procedure uses the GP distribution function for values above a high threshold u, and the empirical distribution function, which we denote $\tilde{F}(\cdot)$, otherwise. Thus, let Y_j be the geomagnetic fluctuations at station s_j, the

distribution function is given by

$$F(y_j) = \begin{cases} 1 - \lambda_{u_j} \left\{ 1 + \dfrac{\xi_j (y_j - u_j)}{\sigma_{u_j}} \right\}_+^{-1/\xi_j} & y_j \geq u, \\ \tilde{F}(y_j) & y_j < u, \end{cases} \tag{1.4.7}$$

where $\lambda_{u_j} = 1 - F(u_j)$, $\sigma_{u_j} > 0$, and $y_+ = \max(y, 0)$. Herein, u_j is a high, user-selected quantile of the data. Then, we can transform each random variable to uniform margins by doing $U_{ij} = \hat{F}_{s_j}(Y_{ij})$.

After transformation of the variables to uniform margins, the correlation needs to be estimated. The censored log-likelihood of the Gaussian copula model based on n independent observations from a pair of random variables (Y_1, Y_2) is given by:

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log L(\boldsymbol{\psi})_i, \tag{1.4.8}$$

where $\boldsymbol{\psi} = \rho$. For chosen high quantiles $u_1^\star$ and $u_2^\star$ on the unit interval $(0, 1)$, the contributions of the likelihood are defined as

$$L(\boldsymbol{\psi})_i = \begin{cases} C_{1,2}(u_1^\star, u_2^\star; \boldsymbol{\psi}) & U_{i1} < u_1^\star, U_{i2} < u_2^\star, \\ c_{1,2}(U_{i1}, U_{i2}; \boldsymbol{\psi}) & U_{i1} > u_1^\star, U_{i2} > u_2^\star, \\ C_1(U_{i1}, u_2^\star; \boldsymbol{\psi}) & U_{i1} > u_1^\star, U_{i2} < u_2^\star, \\ C_2(u_1^\star, U_{i2}; \boldsymbol{\psi}) & U_{i1} < u_1^\star, U_{i2} > u_2^\star. \end{cases} \tag{1.4.9}$$

The cases above correspond to an exceedance at none, both, at site 1 but not site 2, and at site 2 but not site 1, respectively. The contribution of the censored observations (both measurements below the threshold) $C_{1,2}(u_1^\star, u_2^\star; \boldsymbol{\psi})$ in (1.4.9) is given by (1.4.5),

and for the remaining it follows that

$$c_{1,2}(u_1, u_2; \boldsymbol{\psi}) = \phi_{Z_1,Z_2}\left[\Phi^{-1}(u_1), \Phi^{-1}(u_2); \Sigma\right]\left\{\phi\left[\Phi^{-1}(u_1)\right]\right\}^{-1}\left\{\phi\left[\Phi^{-1}(u_2)\right]\right\}^{-1},$$
$$C_1(u_1, u_2; \boldsymbol{\psi}) = \Phi\left[\frac{\Phi^{-1}(u_2) - \rho\Phi^{-1}(u_1)}{\sqrt{1-\rho^2}}\right],$$
$$C_2(u_1, u_2; \boldsymbol{\psi}) = \Phi\left[\frac{\Phi^{-1}(u_1) - \rho\Phi^{-1}(u_2)}{\sqrt{1-\rho^2}}\right],$$

where ϕ_{Z_1,Z_2} is the bivariate Gaussian density function in (1.4.4), with correlation matrix (1.4.6) and mean vector zero, and ϕ is the marginal Gaussian density. Likelihood (1.4.8) is maximised to give a maximum likelihood estimate $\widehat{\boldsymbol{\psi}}$.

Bivariate Huser-Wadsworth copula

In some environmental applications, the dependence weakens as we increase the distance between random variables but $\chi(u)$ does not reach zero (asymptotic independence). It can be difficult to determine whether data are fully AD or AI at finite levels, but most models only exhibit one type of dependence or the other. Huser and Wadsworth (2019) introduced a model for bivariate or spatial data that covers both cases.

Let (W_1, W_2) be a random vector with Gaussian copula and standard Pareto margins, and R an independent standard Pareto random variable. The authors define the dependence model through the following construction:

$$(X_1, X_2) = R^{\delta}(W_1, W_2)^{1-\delta}, \quad \delta \in [0, 1]. \tag{1.4.10}$$

When $\delta > 1/2$, (X_1, X_2) exhibits asymptotic dependence, and when $\delta \leq 1/2$ it exhibits asymptotic independence. The case $\delta = 0$ corresponds to the Gaussian copula, and $\delta = 1$ to perfect dependence. The parameter that describes the relationship between W_1 and W_2 is θ. The copula defined by construction (1.4.10) can be fitted to data as described below.

The parameter estimation method for the bivariate HW copula follows the same approach used for the bivariate Gaussian copula. Let $\boldsymbol{\psi} = (\delta, \theta)^\top$ be the parameter vector. The censored log-likelihood based on n independent observations from the copula of (1.4.10) is given by (1.4.8) and (1.4.9), where each component of the likelihood is given by:

$$
\begin{aligned}
C_{1,2}(u_1, u_2; \boldsymbol{\psi}) &= F_{X_1,X_2}\left[F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)\right], \\
c_{1,2}(u_1, u_2; \boldsymbol{\psi}) &= f_{X_1,X_2}\left[F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)\right]\left\{f_{X_1}\left[F_{X_1}^{-1}(u_1)\right]\right\}^{-1}\left\{f_{X_2}\left[F_{X_2}^{-1}(u_2)\right]\right\}^{-1}, \\
C_1(u_1, u_2; \boldsymbol{\psi}) &= F_{X_1,X_2}^{[1]}\left[F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)\right]\left\{f_{X_1}\left[F_{X_1}^{-1}(u_1)\right]\right\}^{-1}, \\
C_2(u_1, u_2; \boldsymbol{\psi}) &= F_{X_1,X_2}^{[2]}\left[F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)\right]\left\{f_{X_2}\left[F_{X_2}^{-1}(u_1)\right]\right\}^{-1},
\end{aligned}
$$

where F_{X_1,X_2} and f_{X_1,X_2} are the joint distribution function and density, respectively, of the process (X_1, X_2), F_{X_1}, F_{X_2} and f_{X_1}, f_{X_2} are the marginal distribution functions and densities of the same underlying process, respectively, $F_{X_1,X_2}^{[1]}\left[F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)\right] = \partial F_{X_1,X_2}\left[F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)\right] / \partial F_{X_1}^{-1}(u_1)$, and $F_{X_1,X_2}^{[2]}\left[F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)\right] = \partial F_{X_1,X_2}\left[F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)\right] / \partial F_{X_2}^{-1}(u_2)$. More details on the model and expressions for these quantities can be found in Huser and Wadsworth (2019). Again, we maximise likelihood (1.4.8) to obtain $\widehat{\boldsymbol{\psi}} = (\hat{\delta}, \hat{\theta})^\top$.

When accounting for uncertainty, either for maximum likelihood estimators (MLEs) $\widehat{\boldsymbol{\psi}}$ or empirical estimate of $\chi(u)$, the measures were obtained through the stationary bootstrap resampling method of Politis and Romano (1994) for both models. This approach is based on resampling blocks where the length of each block follows a geometric distribution, thus keeping any temporal dependence structure among observations.

1.4.3 Results and discussion

In this section, we shall report the results of the bivariate extreme value analysis we conducted for multiple pairs of sites in each of the two regions shown in Figure 1.2.1.

The data in each site present a significant amount of missing values due to quality concerns or magnetometers not being active for the whole span period. The percentage of missingness for each site as well as the cross-site average for each dataset are shown in Table 1.4.1. On average, both datasets are missing almost half of their data. To overcome the problem of incomplete data and be able to fit the bivariate Gaussian and HW copula models, we ignored all days for which there was a missing value at at least one location. Thus, for the Northern dataset (blue dots in Figure 1.2.1), the total final number of observations is 4340, which corresponds to daily events (not always consecutive) from 11 sites from 1996 to 2015. The missing data for this dataset accounts for 75.25% of the measurement period. For the Southern dataset (black dots in Figure 1.2.1), we have a final sample size of 5089 daily events (not always consecutive) from 9 sites from 1997 to 2012, with missing data accounting for 70.97%. Despite having to discard the majority of the observations, we still have a sufficient sample size to conduct the pairwise analysis.

In the bivariate Gaussian model we estimate the correlation parameter ρ between the response observed at two locations whilst in the bivariate HW model we estimate the extremal dependence through δ and θ. Before fitting the models, the data are transformed to uniform margins using (1.4.7) and a 95% site-specific threshold. For the copula model fitting, the selected threshold also corresponds to the 95^{th} quantile of the data in each site. The distance considered in this work was the geodesic distance, which is the distance between two sites across the curved surface of the world. Every distance unit in this work is equivalent to 100 kilometers.

By taking a first look at the empirical values of $\chi_h(u)$ for a range of thresholds u in Figure 1.4.2, we observe that the rate at which the dependence decreases with distance in the Northern region is faster for higher thresholds. The dependence for sites that are closer is stronger than for those sites that are far away, but it still decreases for extreme levels. For both sites in the auroral ring zone, the extremal dependence goes from strong

to moderate as we increase the threshold. Thus, if a large geomagnetic disturbance at a particular station in this zone is observed, it is very likely that other stations will also experience an extreme value. When we look in more details for $u = 0.90$, the sites (ABK, AND, SOD, TRO) in the auroral ring zone are clustered and very close to each other ($h \in \{1.20, 4.05\}$), suggesting that the strong to moderate dependence relates to the distance between sites. When both observatories are in the north pole zone, the dependence is stronger for sites that are closer and decreases for higher thresholds. For $u = 0.90$, we have $\chi \in \{0.37, 0.77\}$ and $h \in \{1.17, 16.09\}$, a wide range of dependence values and distance units. Now considering three pairs of sites that have the same distance approximately, we note three different levels of dependence. Both sites in Greenland (pair 31 - DMH and SOR) present $h = 7.09$ and $\chi = 0.50$, one site in Greenland and one site in Svalbard (pair 30 - DMH and NAL) present $h = 7.48$ and $\chi = 0.59$, one site in Svalbard and one site in Norway (pair 39 - HRN and SOR) present $h = 7.48$ and $\chi = 0.38$. It's worth noting that pairs 30 and 39 have the same distance between sites but very different dependence levels, the sites in pair 30 are somewhat close in latitude but are far apart in terms of longitude. In contrast, sites in pair 39 are closer in longitude and far apart in latitude. The fact that the level of dependence changes for pairs of sites with same distance in the same zone suggests that latitude plays an important role when accounting for dependence. Finally, when stations are in different zones, one in the north pole and the other in the auroral ring, the dependence structure is similar to cases where both sites are in the north pole zone.

Table 1.4.2 gives the MLEs, standard deviations (SDs), and 95% confidence intervals (CIs) of ρ, θ, and δ for each of the 55 pairs in the Northern dataset. The block size used in the bootstrap method to obtain the uncertainty measures was six days, chosen after inspection of autocorrelation functions for a variety of sites. The MLEs and corresponding bootstrap CIs in Table 1.4.2 are plotted against distance between sites in Figure 1.4.3. We note that the MLEs of ρ for the Gaussian model (red dots) decrease

as the distance units increase, but does not reach zero, suggesting that the dependence of extremes weakens for stations far apart but does not reach independence. For the HW model, the MLEs of δ suggest asymptotic independence ($\hat{\delta} < 0.5$), except for pair 10 ($\hat{\delta} = 0.52$). The MLEs of θ follow the trend of the MLEs of ρ, decreasing as the distance increases. Although pair 10 is AD as $\hat{\delta} > 0.5$, the 95% bootstrap CI for $\hat{\delta}$ for this pair includes both extremal dependence regimes, thus firm conclusions about the true dependence structure are difficult to draw. The CIs for $\hat{\delta}$ for pairs 1, 16, 19, 20 and 33 also include both types of asymptotic dependence, although $\hat{\delta} < 0.5$, which suggests asymptotic independence.

To assess the fit of the bivariate models across various thresholds, Figures 1.4.4, 1.4.5, and 1.4.6 show the model-based $\chi_h(u)$ estimates for $u \in \{0.90, 0.95, 0.99\}$, respectively, over a range of distances h. For higher thresholds, we observe that the dependence decays faster as the distance increases, and uncertainty in the estimates increases with threshold due to the small sample size at higher quantiles. Further, estimates from the Gaussian copula have wider CIs than estimates from the HW model. To check the fit of some pairs, Figure 1.4.7 shows both empirical and model-based estimates of $\chi(u)$ for a range of different thresholds. Both models were fitted using $u = 0.95$. For $u > 0.95$, they present similar fit to all pairs, and for $u < 0.95$ they are not expected to perform well as they were not tailored to this region.

Moving onto the analysis of the Southern dataset, the empirical pairwise extremal dependence structure is illustrated in Figure 1.4.8. The dependence for both sites in the subauroral zone decreases as we increase the threshold but it is unclear if it is the increased distance or the relative locations of the two sites that better accounts for the weaker dependence. For pairs where both sites are in lower latitudes, the dependence is strong across all quantiles for a wide range of distances, that is, even for sites that are far away the dependence is still strong, suggesting that the location of the sites accounts more for the strong dependence than the distance between them. Finally, for sites in

different zones we observe that the dependence decreases as we increase the threshold and distance between sites. Looking in more details for $u = 0.90$, we note that in all cases where sites are in different zones, at least one of the sites in the subauroral zone (LER and LRV) is present, with minimum and maximum distances of approximately 5 and 29 units, respectively. This might be an indicative that for any pair of sites that has a site in the subauroral zone, the weaker dependence as $u \to 1$ depends more strongly on the location than the distance between sites.

Table 1.4.3 presents the MLEs, SDs, and 95% CIs of ρ, θ, and δ for each of the 36 pairs in the Southern dataset. We again use a block of length six days to obtain the uncertainty measures. The MLEs and CIs in Table 1.4.3 are plotted against distance between sites in Figure 1.4.9. For this dataset, where the stations are located at lower latitudes but with a wide range of latitudes in relation to the Northern dataset, we observe spatial non-stationarity, that is, the dependence is not constant across the study region. The MLEs of ρ for the Gaussian model are close to one for pairs of stations that are separated by no more than ten geodesic distance units. Strong correlation is also seen for pairs of sites separated by more than ten geodesic distance units, with extremal dependence decreasing when we increase the threshold but not reaching asymptotic independence. Regarding the MLEs of δ, we have $\hat{\delta} \searrow 0$ with distance, meaning that the dependence structure of the W process is recovered for longer distances, in this case, the Gaussian copula. For short distances, some pairs display asymptotic dependence with MLEs of δ around or greater than 0.8, and wide bootstrap CIs for $\hat{\theta}$. According to Huser and Wadsworth (2019), $\hat{\delta} > 0.8$ may cause some numerical difficulties when fitting the model which might explain the large uncertainty on the parameter estimates.

Pairs 1, 10, 13, 15, 24, and 34 present $\hat{\delta} > 0.5$, indicating asymptotic dependence. Firm conclusions about pair 15 are difficult to draw as the 95% CI for this particular bivariate process includes both types of dependence. For pairs 3, 8, 14, 21, and 35 we have $\hat{\delta} < 0.5$ (asymptotic independence), however, their bootstrap CIs include the

case of asymptotic dependence, thus concrete conclusions about them cannot be made either.

Figures 1.4.10, 1.4.11, and 1.4.12 show the model-based $\chi_h(u)$ estimates and their corresponding 95% CIs for $u \in \{0.90, 0.95, 0.99\}$, respectively, across a range of h for the Southern dataset. For this dataset, we observe a decreasing linear structure as distance increases, compared to the exponential decay observed in Figures 1.4.4, 1.4.5, and 1.4.6. Also, the uncertainty increases as we increase the threshold, with again wider CIs for the estimates from the Gaussian copula than the HW copula. Figure 1.4.13 shows the estimated values of χ from the fitted bivariate models across a range of quantiles for some pairs. For short distances such as Figures 1.4.13(a) and 1.4.13(b), the dependence is constant across the range of thresholds. In this scenario, the HW model outperforms the Gaussian model for $u > 0.95$.

Overall, the pairwise extremal spatial dependence structure differs in the two regions. Pairs of sites in Northern Europe tend to be AI with extremal dependence decaying faster with distance and high thresholds, and pairs in Southern Europe present strong dependence across all thresholds, and persists for longer distances when both sites are in lower latitudes. This corroborates with Thomson et al. (2011), who point out that the most extreme geomagnetic field fluctuations are observed in the subauroral zone due to an enhanced auroral electrojet that travels south with the help of strong solar wind. This may also explain why observatories in Northern Europe display weaker dependence.

The statistical results from our analysis align with the expectations of the physicists. The combination of auroral and polar cap current systems in the Northern dataset is susceptible to moderate geomagnetic substorms, but less prone to experiencing the most extreme substorm current systems found further south, near Scotland. Additionally, the Northern dataset covers a wide range of longitudes, including the two Greenland sites. This geographic diversity suggests that surges of substorm currents or brief bursts of

ultralow frequency wave activity at one site may not extend across a broad longitudinal area. Consequently, this explains the observed lack of strong asymptotic dependence within this dataset.

On the other hand, the Southern dataset primarily includes sites located to the south of Scotland, indicating a subauroral region. In this context, extreme events are more likely to take the form of sudden commencements that can affect all latitudes. This explains the observed strong dependence within the Southern dataset.

These variations in asymptotic dependence provide valuable insights into the complex relationship between space weather phenomena and Earth's magnetic field. Ultimately, they enhance our understanding of how extreme events are distributed in space and their potential impacts on technological systems at different latitudes.

For the two bivariate extreme value models primarily discussed in this work, estimating the MLEs was somewhat computationally intensive. The complexity of the models, the size of the datasets and the convergence behaviour all influenced the computational effort required. For this, we implemented parallelization in the code and ran the experiments using high-performance computing resources. The HW model, in particular, posed some challenges related to convergence due to the latent variables and complex dependencies among extreme values. To address the issue of convergence, we applied multiple random starts with different initial values to ensure that the maximum likelihood estimation converged. For the Northern dataset (11 sites modelled in pairs), the MLEs of the Gaussian copula model were obtained in 58 minutes and for the HW model they were obtained in six hours. For the Southern dataset (9 sites modelled in pairs), the maximum likelihood estimation of the Gaussian copula model took 54 minutes and for the HW model it took 19 hours.

In terms of missingness, it can be particularly challenging because it might not always occur randomly. In some cases, the absence of data can be informative itself, indicating unusual conditions or instrument failures during specific time periods. This

Table 1.4.1: Percentage of missing values for each site and dataset.

Northern Europe		Southern Europe	
Site	Missing values (%)	Site	Missing values (%)
ABK	23.16	CLF	27.29
AND	58.56	DOU	60.15
BJN	41.48	ESK	29.21
DMH	56.08	HAD	29.26
HRN	50.10	LER	29.18
LYR	59.79	LRV	53.82
NAL	58.61	MAB	60.15
SCO	54.16	SPT	58.42
SOD	23.17	VAL	47.80
SOR	59.48	-	-
TRO	45.90	-	-
Average	48.23	Average	43.92

informative missingness can introduce bias in the analysis, as it may be related to the very extreme events we want to study. When some observations are missing, it can lead to underestimation or overestimation of rare events, resulting in inaccurate risk assessments and decision-making in space weather forecasting.

1.5 Concluding remarks

In this work, we focused on understanding the pairwise extremal dependence of geomagnetic field fluctuations in Europe, considering measurements from 20 observatories. We fitted two bivariate copula models to the data, the Gaussian copula and a model proposed by Huser and Wadsworth (2019). Our results showed that the pairwise extremal dependence structure differs for Northern and Southern Europe. For the Northern region, the extremal dependence between measurements of two sites decays faster as we increase the threshold and distance, whilst for the Southern region such dependence is strong for observatories up to ten geodesic distance units away, which corresponds to about 1000 kilometers in the great circle. Also, for pairs where both sites are in

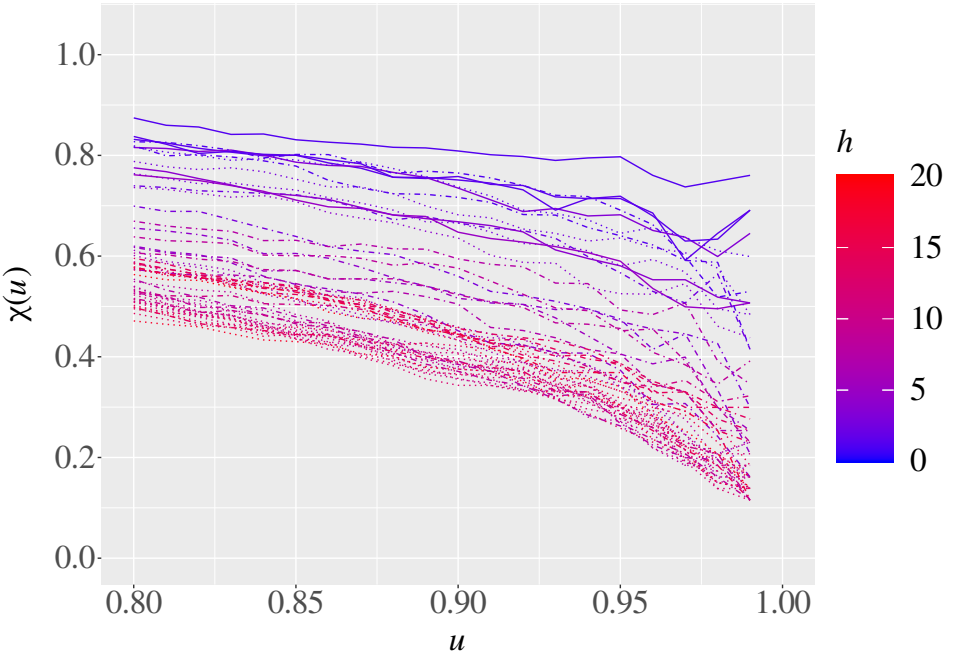


Figure 1.4.2: Empirical $\chi_h(u)$ colour-coded by geodesic distance units, h. The solid line means that both sites are in the auroral ring zone ($65°\text{N} - 70°\text{N}$), the dash–dot line both sites are in the north pole zone ($70°\text{N} - 90°\text{N}$), and the dotted line represents pairs where sites are in different zones. Results for Northern Europe.

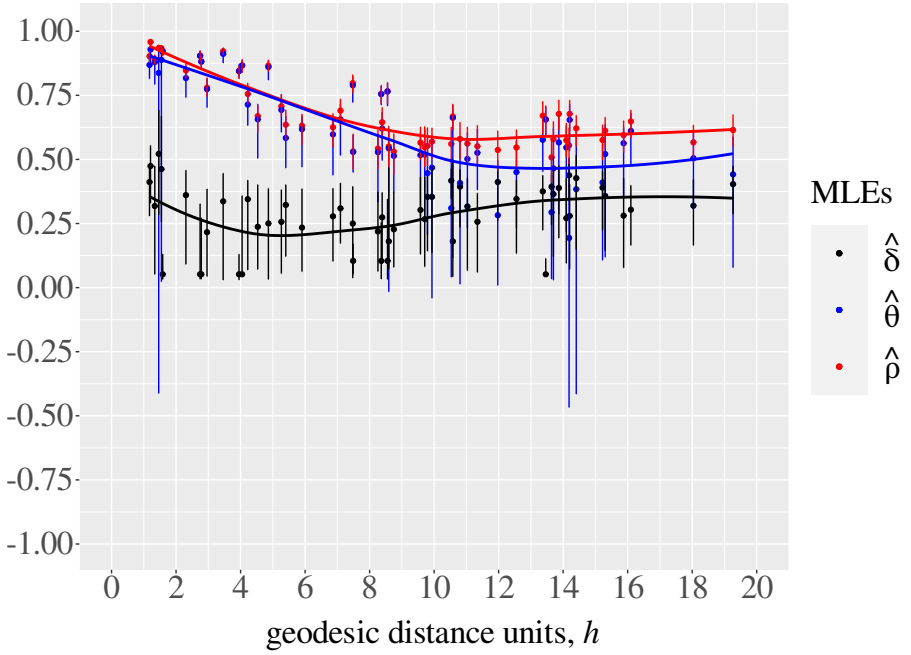


Figure 1.4.3: MLEs and corresponding 95% bootstrap CIs for the bivariate Gaussian and HW models fitted to the data in Northern Europe.

Table 1.4.2: MLEs, SDs of bootstrap resampling, and 95% bootstrap CIs for parameters of the bivariate Gaussian (ρ) and HW (δ, θ) models for each pair; Northern dataset.

Pair	$\hat{\rho}$	$\mathrm{SD}_{\hat{\rho}}$	95% CI	$\hat{\delta}$	$\mathrm{SD}_{\hat{\delta}}$	95% CI	$\hat{\theta}$	$\mathrm{SD}_{\hat{\theta}}$	95% CI
1	0.934	0.007	(0.919, 0.945)	0.462	0.088	(0.256, 0.669)	0.888	0.178	(0.023, 0.937)
2	0.625	0.033	(0.548, 0.679)	0.278	0.075	(0.103, 0.387)	0.598	0.059	(0.439, 0.666)
3	0.576	0.036	(0.498, 0.635)	0.390	0.061	(0.223, 0.456)	0.410	0.141	(0.107, 0.589)
4	0.548	0.041	(0.471, 0.629)	0.267	0.089	(0.082, 0.431)	0.519	0.148	(0.257, 0.619)
5	0.562	0.036	(0.483, 0.626)	0.317	0.094	(0.066, 0.428)	0.502	0.104	(0.233, 0.605)
6	0.537	0.040	(0.456, 0.612)	0.412	0.050	(0.262, 0.463)	0.282	0.138	(0.009, 0.497)
7	0.595	0.032	(0.526, 0.651)	0.281	0.087	(0.077, 0.426)	0.563	0.072	(0.347, 0.626)
8	0.922	0.008	(0.907, 0.936)	0.337	0.146	(0.030, 0.446)	0.912	0.116	(0.876, 0.932)
9	0.882	0.013	(0.854, 0.904)	0.052	0.054	(0.034, 0.272)	0.882	0.013	(0.852, 0.903)
10	0.935	0.008	(0.920, 0.947)	0.522	0.083	(0.284, 0.692)	0.837	0.284	(−0.412, 0.933)
11	0.632	0.032	(0.558, 0.677)	0.234	0.086	(0.063, 0.386)	0.618	0.060	(0.470, 0.669)
12	0.579	0.037	(0.512, 0.648)	0.365	0.085	(0.120, 0.459)	0.465	0.153	(0.029, 0.621)
13	0.550	0.043	(0.472, 0.630)	0.180	0.107	(0.072, 0.469)	0.543	0.146	(−0.016, 0.618)
14	0.569	0.034	(0.503, 0.623)	0.354	0.079	(0.136, 0.473)	0.468	0.144	(−0.041, 0.592)
15	0.580	0.038	(0.503, 0.645)	0.394	0.055	(0.236, 0.460)	0.408	0.150	(0.014, 0.590)
16	0.621	0.030	(0.551, 0.672)	0.427	0.081	(0.148, 0.517)	0.384	0.264	(−0.415, 0.626)
17	0.866	0.012	(0.843, 0.888)	0.250	0.096	(0.032, 0.387)	0.861	0.022	(0.810, 0.885)
18	0.904	0.010	(0.885, 0.924)	0.052	0.066	(0.033, 0.327)	0.904	0.012	(0.879, 0.924)
19	0.958	0.005	(0.947, 0.970)	0.474	0.052	(0.330, 0.554)	0.929	0.025	(0.869, 0.964)
20	0.667	0.026	(0.615, 0.716)	0.180	0.133	(0.043, 0.500)	0.663	0.161	(0.116, 0.712)
21	0.780	0.019	(0.745, 0.819)	0.216	0.097	(0.044, 0.385)	0.774	0.028	(0.702, 0.816)
22	0.755	0.022	(0.714, 0.798)	0.345	0.089	(0.069, 0.419)	0.714	0.039	(0.632, 0.785)
23	0.708	0.024	(0.667, 0.755)	0.257	0.094	(0.055, 0.382)	0.693	0.038	(0.606, 0.753)
24	0.678	0.031	(0.610, 0.731)	0.280	0.083	(0.072, 0.404)	0.654	0.047	(0.534, 0.716)
25	0.645	0.033	(0.581, 0.703)	0.274	0.068	(0.094, 0.371)	0.620	0.049	(0.496, 0.684)
26	0.669	0.029	(0.605, 0.714)	0.237	0.086	(0.071, 0.401)	0.656	0.055	(0.504, 0.716)
27	0.635	0.029	(0.575, 0.693)	0.323	0.068	(0.122, 0.401)	0.584	0.060	(0.467, 0.669)
28	0.766	0.020	(0.717, 0.799)	0.104	0.069	(0.033, 0.346)	0.765	0.023	(0.709, 0.799)
29	0.755	0.018	(0.715, 0.789)	0.104	0.053	(0.034, 0.255)	0.755	0.018	(0.712, 0.788)
30	0.798	0.019	(0.760, 0.831)	0.250	0.099	(0.038, 0.394)	0.789	0.031	(0.722, 0.825)
31	0.690	0.025	(0.634, 0.736)	0.309	0.071	(0.172, 0.408)	0.656	0.053	(0.516, 0.717)
32	0.567	0.036	(0.488, 0.634)	0.320	0.069	(0.165, 0.421)	0.505	0.088	(0.305, 0.610)
33	0.555	0.038	(0.483, 0.624)	0.438	0.047	(0.337, 0.507)	0.194	0.227	(−0.467, 0.503)
34	0.575	0.037	(0.508, 0.642)	0.271	0.084	(0.095, 0.426)	0.546	0.101	(0.216, 0.630)
35	0.891	0.010	(0.871, 0.908)	0.318	0.112	(0.052, 0.475)	0.880	0.034	(0.793, 0.903)
36	0.847	0.015	(0.817, 0.875)	0.361	0.090	(0.091, 0.458)	0.817	0.032	(0.742, 0.865)
37	0.656	0.029	(0.597, 0.709)	0.052	0.029	(0.050, 0.114)	0.656	0.030	(0.590, 0.705)
38	0.551	0.041	(0.475, 0.633)	0.257	0.077	(0.060, 0.369)	0.526	0.061	(0.381, 0.619)
39	0.531	0.040	(0.461, 0.600)	0.104	0.039	(0.053, 0.171)	0.529	0.043	(0.450, 0.596)
40	0.542	0.039	(0.463, 0.615)	0.219	0.077	(0.063, 0.373)	0.528	0.087	(0.333, 0.599)
41	0.902	0.012	(0.878, 0.924)	0.412	0.051	(0.280, 0.469)	0.869	0.024	(0.815, 0.909)
42	0.678	0.030	(0.615, 0.732)	0.389	0.068	(0.193, 0.476)	0.566	0.120	(0.287, 0.672)
43	0.547	0.039	(0.469, 0.616)	0.346	0.075	(0.134, 0.419)	0.451	0.089	(0.232, 0.572)
44	0.531	0.040	(0.441, 0.601)	0.227	0.080	(0.079, 0.374)	0.514	0.056	(0.372, 0.576)
45	0.566	0.033	(0.498, 0.627)	0.303	0.070	(0.130, 0.431)	0.517	0.084	(0.278, 0.604)
46	0.671	0.030	(0.599, 0.727)	0.375	0.053	(0.224, 0.438)	0.577	0.064	(0.452, 0.674)
47	0.508	0.045	(0.409, 0.586)	0.394	0.049	(0.244, 0.441)	0.294	0.119	(0.034, 0.513)
48	0.553	0.040	(0.475, 0.624)	0.354	0.074	(0.143, 0.435)	0.446	0.108	(0.206, 0.587)
49	0.561	0.038	(0.486, 0.626)	0.417	0.031	(0.346, 0.463)	0.310	0.140	(0.041, 0.504)
50	0.615	0.033	(0.548, 0.674)	0.404	0.048	(0.287, 0.475)	0.442	0.168	(0.078, 0.588)
51	0.648	0.028	(0.584, 0.693)	0.304	0.060	(0.165, 0.399)	0.611	0.055	(0.477, 0.680)
52	0.612	0.032	(0.545, 0.665)	0.357	0.082	(0.144, 0.478)	0.521	0.165	(0.119, 0.637)
53	0.845	0.013	(0.814, 0.865)	0.052	0.032	(0.030, 0.130)	0.845	0.014	(0.819, 0.870)
54	0.867	0.013	(0.838, 0.890)	0.052	0.089	(0.052, 0.358)	0.866	0.014	(0.840, 0.887)
55	0.921	0.008	(0.903, 0.936)	0.052	0.032	(0.032, 0.131)	0.921	0.008	(0.906, 0.938)

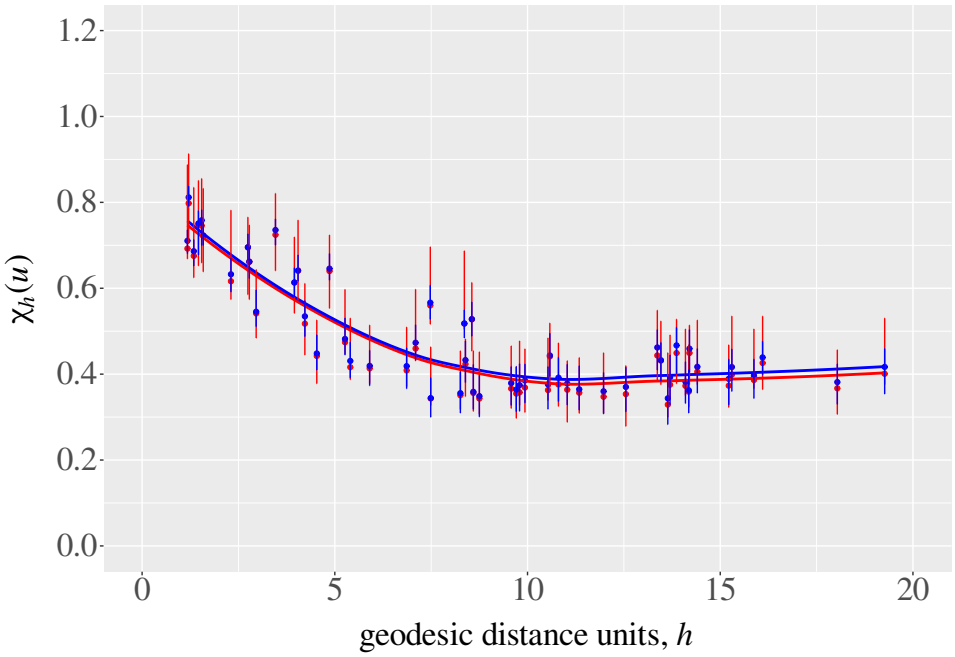


Figure 1.4.4: Model-based $\chi_h(u)$ estimates and corresponding 95% bootstrap CIs as a function of geodesic distance units for $u = 0.90$; Northern dataset. Red line: smooth curve of the estimates from the Gaussian model; blue line: smooth curve of the estimates from the HW model.

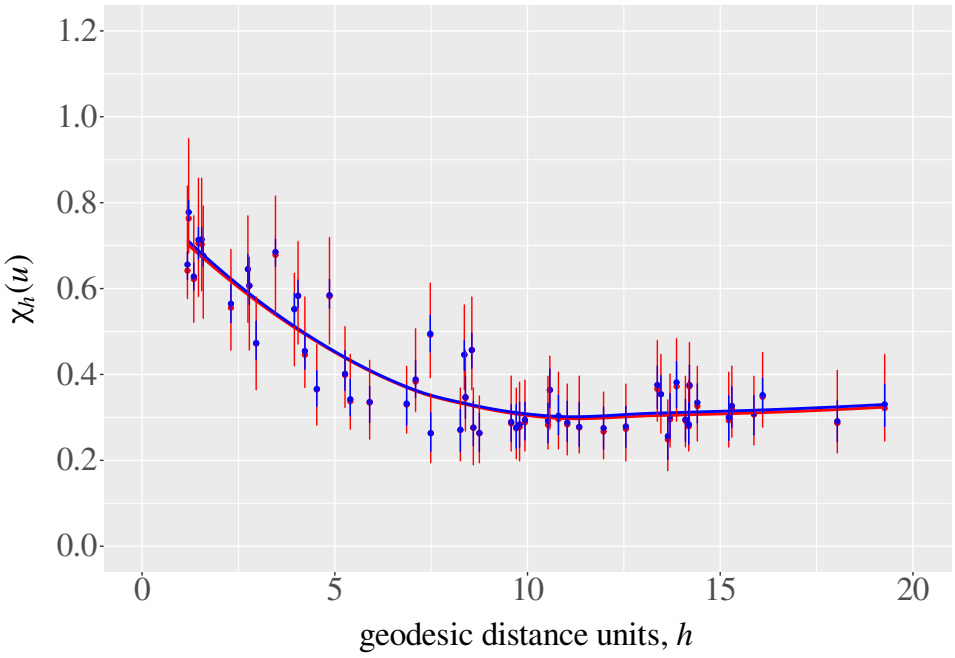


Figure 1.4.5: Model-based $\chi_h(u)$ estimates and corresponding 95% bootstrap CIs as a function of geodesic distance units for $u = 0.95$; Northern dataset. Red line: smooth curve of the estimates from the Gaussian model; blue line: smooth curve of the estimates from the HW model.

Table 1.4.3: MLEs, SDs of bootstrap resampling, and 95% bootstrap CIs for parameters of the bivariate Gaussian (ρ) and HW (δ, θ) models for each pair; Southern dataset.

Pair	$\hat{\rho}$	$\text{SD}_{\hat{\rho}}$	95% CI	$\hat{\delta}$	$\text{SD}_{\hat{\delta}}$	95% CI	$\hat{\theta}$	$\text{SD}_{\hat{\theta}}$	95% CI
1	0.986	0.002	(0.982, 0.990)	0.739	0.032	(0.698, 0.813)	0.677	0.181	(0.125, 0.797)
2	0.965	0.004	(0.958, 0.972)	0.395	0.036	(0.308, 0.445)	0.957	0.006	(0.947, 0.969)
3	0.986	0.002	(0.981, 0.990)	0.464	0.078	(0.258, 0.546)	0.977	0.008	(0.959, 0.988)
4	0.861	0.011	(0.840, 0.882)	0.104	0.069	(0.022, 0.350)	0.861	0.012	(0.833, 0.881)
5	0.710	0.022	(0.664, 0.749)	0.052	0.040	(0.034, 0.150)	0.710	0.029	(0.659, 0.750)
6	0.985	0.003	(0.979, 0.991)	0.785	0.035	(0.719, 0.849)	0.277	0.430	(−0.868, 0.690)
7	0.987	0.002	(0.983, 0.989)	0.314	0.118	(0.054, 0.459)	0.986	0.003	(0.978, 0.989)
8	0.971	0.004	(0.964, 0.978)	0.465	0.031	(0.394, 0.513)	0.954	0.010	(0.933, 0.970)
9	0.965	0.004	(0.957, 0.973)	0.431	0.036	(0.342, 0.479)	0.952	0.009	(0.932, 0.967)
10	0.980	0.002	(0.975, 0.984)	0.794	0.027	(0.720, 0.820)	−0.443	0.384	(−0.973, 0.546)
11	0.870	0.011	(0.850, 0.891)	0.052	0.022	(0.020, 0.103)	0.870	0.011	(0.849, 0.892)
12	0.732	0.021	(0.693, 0.773)	0.328	0.106	(0.087, 0.482)	0.693	0.117	(0.421, 0.754)
13	0.980	0.004	(0.972, 0.986)	0.768	0.036	(0.690, 0.828)	0.150	0.387	(−0.91, 0.683)
14	0.973	0.004	(0.965, 0.979)	0.447	0.075	(0.310, 0.569)	0.960	0.124	(0.891, 0.977)
15	0.969	0.004	(0.959, 0.975)	0.500	0.058	(0.412, 0.605)	0.938	0.171	(0.802, 0.967)
16	0.979	0.003	(0.974, 0.985)	0.438	0.041	(0.342, 0.495)	0.971	0.006	(0.957, 0.983)
17	0.913	0.007	(0.896, 0.925)	0.052	0.023	(0.022, 0.104)	0.913	0.008	(0.896, 0.925)
18	0.746	0.019	(0.710, 0.781)	0.314	0.081	(0.092, 0.427)	0.716	0.047	(0.580, 0.771)
19	0.965	0.004	(0.956, 0.974)	0.438	0.036	(0.357, 0.484)	0.950	0.009	(0.932, 0.969)
20	0.948	0.005	(0.939, 0.958)	0.377	0.053	(0.241, 0.436)	0.938	0.009	(0.920, 0.953)
21	0.982	0.002	(0.978, 0.987)	0.451	0.050	(0.315, 0.512)	0.974	0.007	(0.959, 0.984)
22	0.879	0.010	(0.857, 0.899)	0.104	0.056	(0.021, 0.271)	0.878	0.011	(0.856, 0.898)
23	0.730	0.021	(0.684, 0.764)	0.245	0.095	(0.057, 0.397)	0.718	0.044	(0.627, 0.754)
24	0.987	0.003	(0.981, 0.992)	0.791	0.034	(0.726, 0.855)	0.315	0.343	(−0.645, 0.715)
25	0.977	0.003	(0.971, 0.983)	0.340	0.112	(0.053, 0.484)	0.975	0.013	(0.953, 0.983)
26	0.984	0.002	(0.980, 0.987)	0.368	0.066	(0.192, 0.453)	0.981	0.003	(0.973, 0.986)
27	0.816	0.013	(0.792, 0.843)	0.324	0.091	(0.071, 0.420)	0.793	0.028	(0.728, 0.830)
28	0.859	0.011	(0.840, 0.878)	0.104	0.036	(0.021, 0.154)	0.859	0.011	(0.837, 0.878)
29	0.837	0.014	(0.808, 0.863)	0.052	0.027	(0.031, 0.136)	0.836	0.015	(0.808, 0.862)
30	0.906	0.007	(0.892, 0.919)	0.022	0.053	(0.021, 0.238)	0.906	0.007	(0.892, 0.919)
31	0.710	0.023	(0.659, 0.753)	0.052	0.026	(0.025, 0.125)	0.709	0.023	(0.656, 0.751)
32	0.686	0.025	(0.632, 0.720)	0.020	0.025	(0.020, 0.101)	0.686	0.025	(0.633, 0.721)
33	0.747	0.019	(0.704, 0.781)	0.356	0.071	(0.163, 0.450)	0.695	0.056	(0.542, 0.756)
34	0.981	0.003	(0.974, 0.987)	0.810	0.019	(0.763, 0.837)	−0.684	0.314	(−0.995, 0.109)
35	0.970	0.004	(0.963, 0.977)	0.438	0.074	(0.243, 0.535)	0.957	0.016	(0.915, 0.974)
36	0.958	0.006	(0.946, 0.968)	0.393	0.095	(0.122, 0.498)	0.948	0.018	(0.905, 0.967)

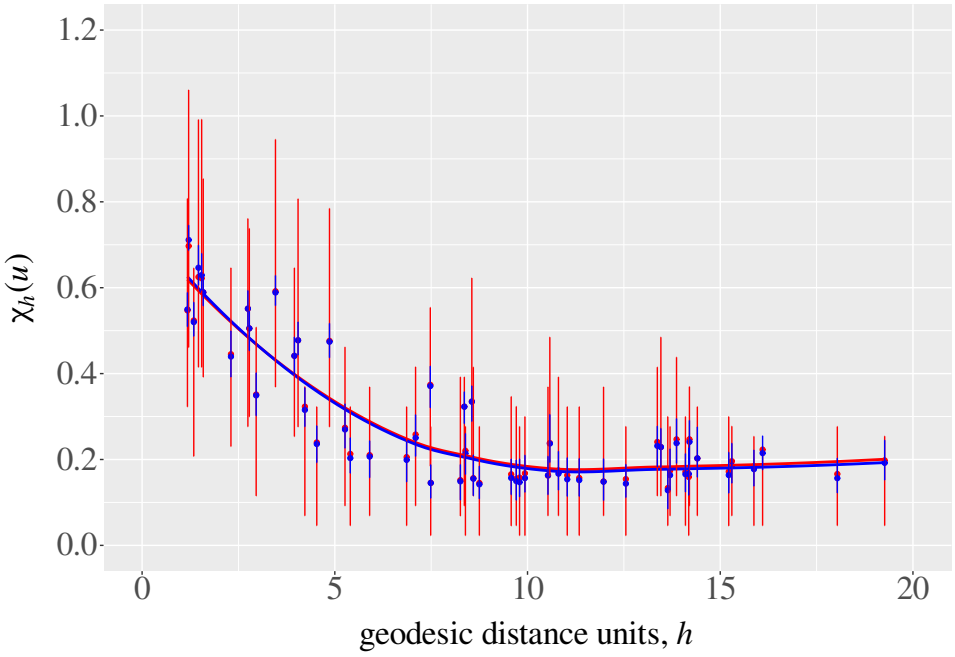
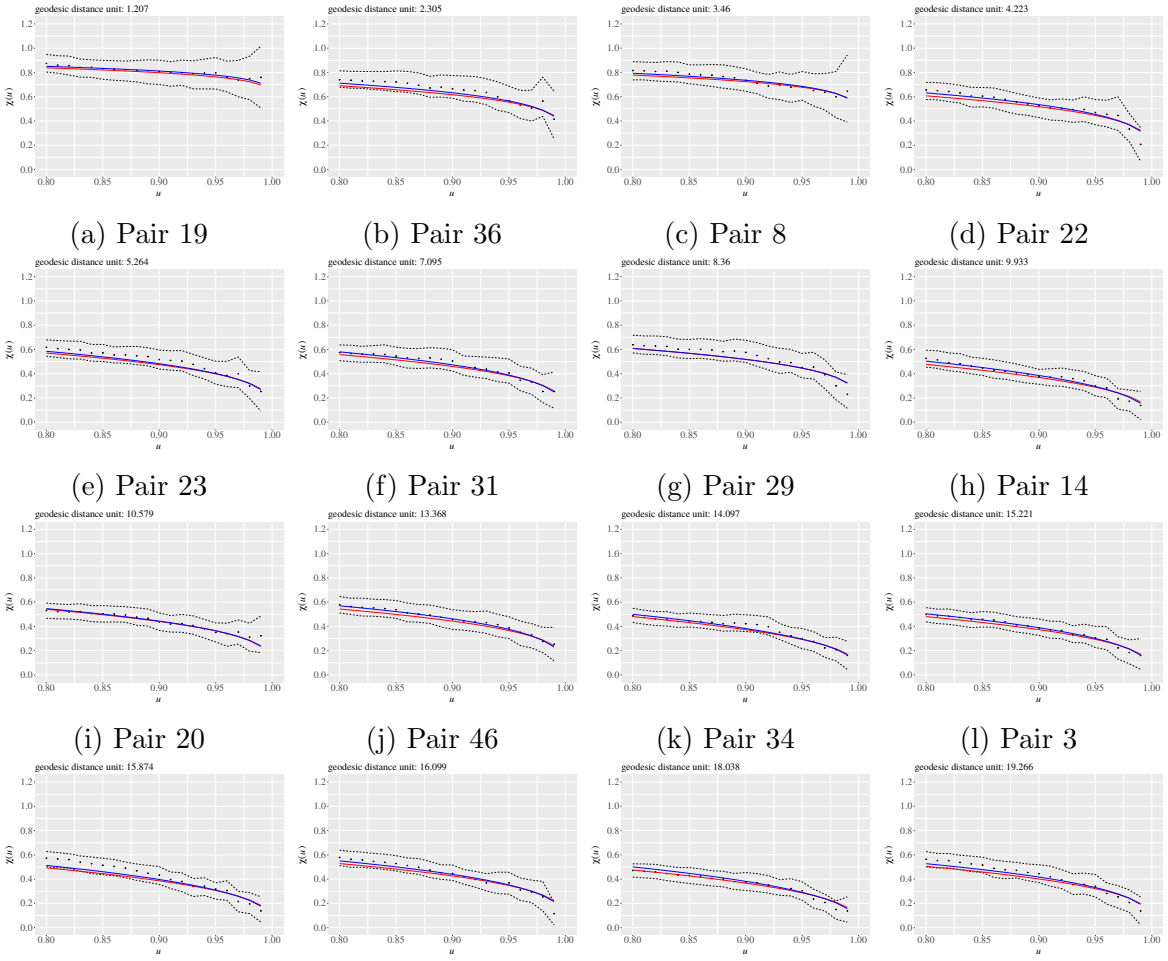


Figure 1.4.6: Model-based $\chi_h(u)$ estimates and corresponding 95% bootstrap CIs as a function of geodesic distance units for $u = 0.99$; Northern dataset. Red line: smooth curve of the estimates from the Gaussian model; blue line: smooth curve of the estimates from the HW model.

lower latitudes, the extremal dependence is strong for a range of distances and across all quantiles. Although the Gaussian copula was tailored to only the extreme events, it cannot capture both asymptotic dependence regimes for extreme quantiles. This is potentially problematic for pairs of sites in Southern Europe, where many dependence types exist. The results from our analysis showed that the HW model, which allows for both asymptotic dependence and asymptotic independence, better captures the varying dependence structures seen in Southern Europe.

The findings of this work have to be seen in light of some limitations. First, the pairwise analysis only is not sufficient to describe the global dependence structure and cannot provide spatial interpolation. Second, we have looked only at two small regions on one part of the globe, and higher dimensions might result in different conclusions as we incorporate other features of the process. Finally, it is unclear from the present analysis what drives the spatial non-stationarity in Southern Europe.



(a) Pair 19 (b) Pair 36 (c) Pair 8 (d) Pair 22

(e) Pair 23 (f) Pair 31 (g) Pair 29 (h) Pair 14

(i) Pair 20 (j) Pair 46 (k) Pair 34 (l) Pair 3

(m) Pair 7 (n) Pair 51 (o) Pair 32 (p) Pair 50

Figure 1.4.7: Estimates of $\chi(u)$ for the bivariate geomagnetic field fluctuation data in Northern Europe. Central black dots are the empirical estimates of $\chi(u)$, dashed lines are 95% bootstrap CIs based on block bootstrap resampling, solid red line is the fit from the bivariate Gaussian model and solid blue line is the fit from the bivariate HW model. Plots are ordered by ascending geodesic distance units between sites.

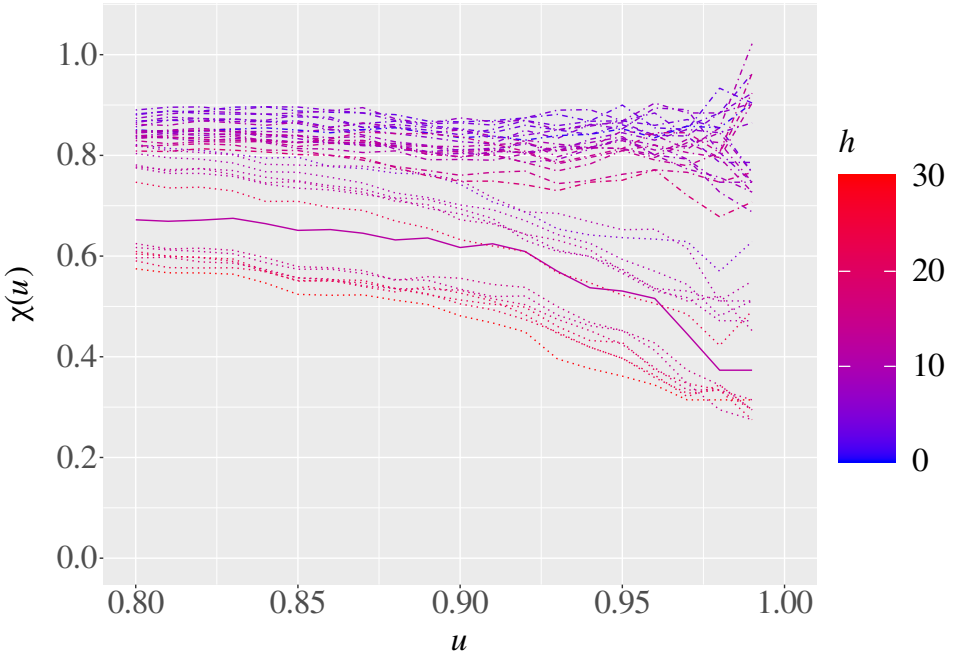


Figure 1.4.8: Empirical $\chi_h(u)$ as a function of geodesic distance units, h. The solid line means that both sites are in the subauroral zone ($60°\text{N} - 65°\text{N}$), the dash–dot line both sites are in the low latitudes ($39°\text{N} - 60°\text{N}$), and the dotted line represents pairs where sites are in different zones. Results for Southern Europe.

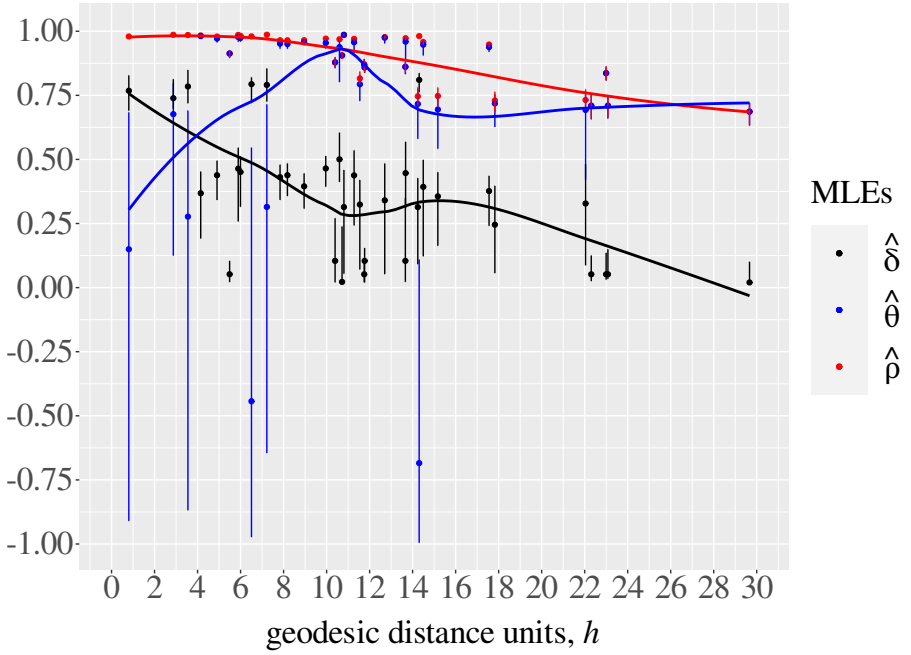


Figure 1.4.9: MLEs and corresponding 95% bootstrap CIs for the bivariate Gaussian and HW models fitted to the data in Southern Europe.

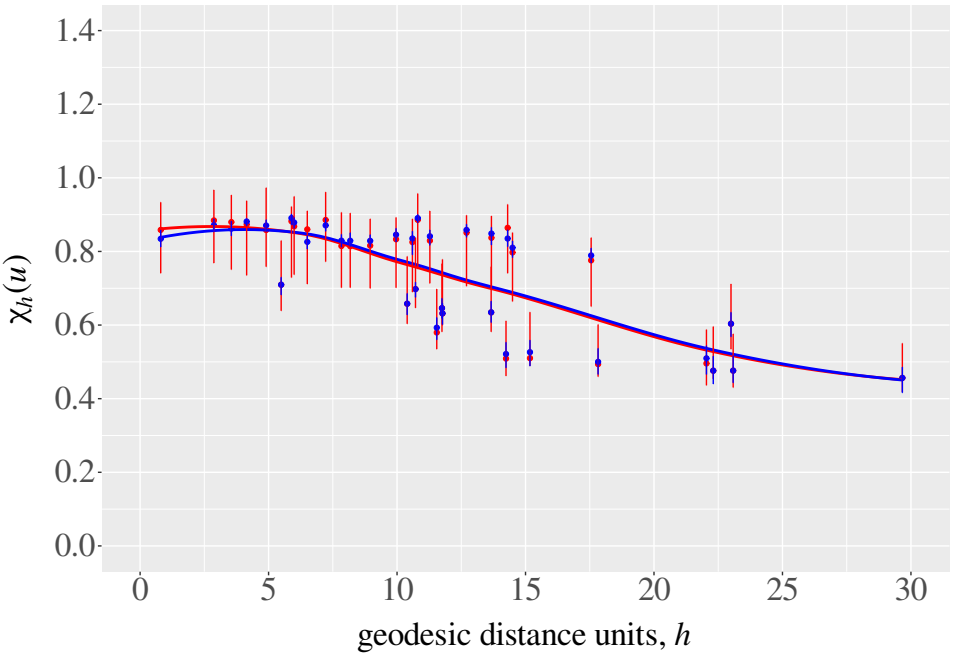


Figure 1.4.10: Model-based $\chi(u)$ estimates and corresponding 95% bootstrap CIs as a function of geodesic distance units for $u = 0.90$; Southern dataset. Red line: smooth curve of the estimates from the Gaussian model; blue line: smooth curve of the estimates from the HW model.

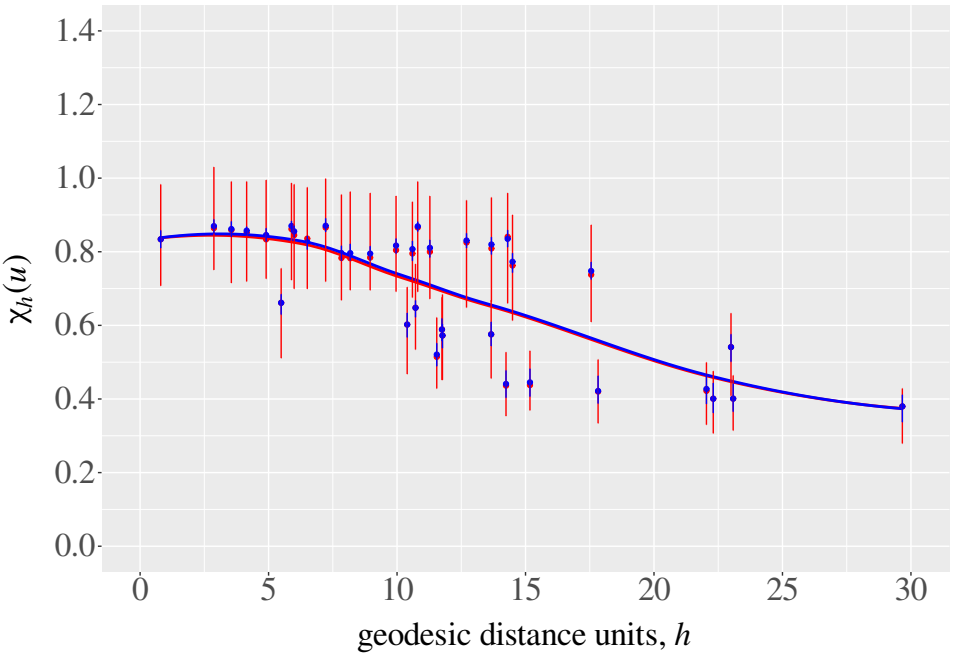


Figure 1.4.11: Model-based $\chi(u)$ estimates and corresponding 95% bootstrap CIs as a function of geodesic distance units for $u = 0.95$; Southern dataset. Red line: smooth curve of the estimates from the Gaussian model; blue line: smooth curve of the estimates from the HW model.

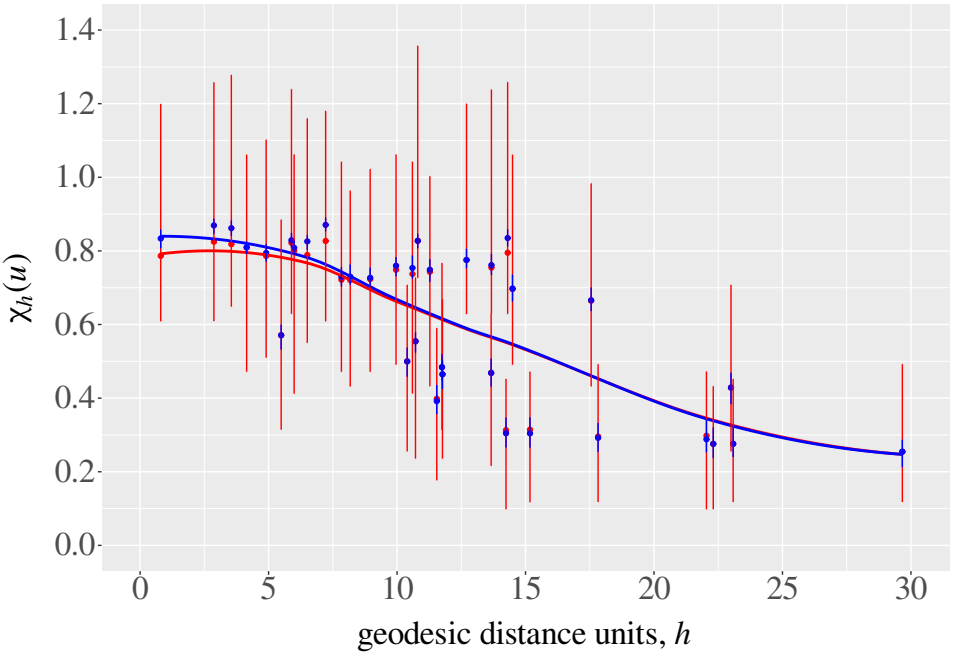
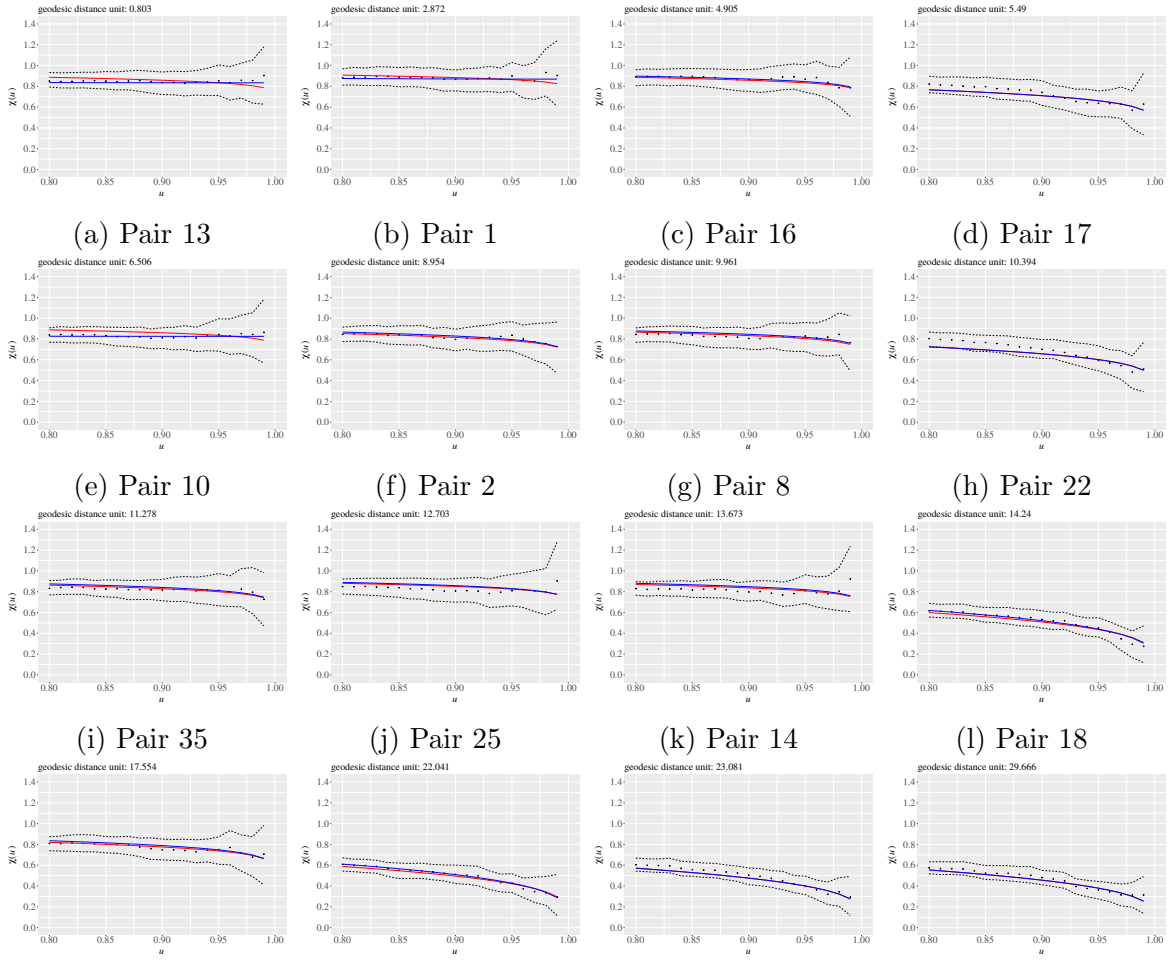


Figure 1.4.12: Model-based $\chi(u)$ estimates and corresponding 95% bootstrap CIs as a function of geodesic distance units for $u = 0.99$; Southern dataset. Red line: smooth curve of the estimates from the Gaussian model; blue line: smooth curve of the estimates from the HW model.

For future work, an alternative is to fit full spatial models to the two current sets of sites to consider all the sites together and obtain a picture of the global dependence structure. By expanding the study region, it might also be possible to investigate if it is the region or distance between sites that best describes the weaker dependence. Regions where different types of extremal dependence are observed should be better explored to incorporate drivers for this in the model, such as difference in latitude, which might be the case for Southern Europe.

Whilst we attempted to explore the above possibilites, the presence of missing data poses a challenge to extending the analysis into a full spatial framework, where we would need uniform data availability across all sites over the same time period. Furthermore, another possible limitation lies in the computational demands and time required for modelling the data. Given that the HW model already takes a significant amount of time in bivariate scenarios, modelling data from more than two sites simultaneously



(a) Pair 13 (b) Pair 1 (c) Pair 16 (d) Pair 17

(e) Pair 10 (f) Pair 2 (g) Pair 8 (h) Pair 22

(i) Pair 35 (j) Pair 25 (k) Pair 14 (l) Pair 18

(m) Pair 20 (n) Pair 12 (o) Pair 5 (p) Pair 32

Figure 1.4.13: Estimates of $\chi(u)$ for the bivariate geomagnetic field fluctuation data in Southern Europe. Central black dots are the empirical estimates of $\chi(u)$, dashed lines are 95% bootstrap CIs based on block bootstrap resampling, solid red line is the fit from the bivariate Gaussian model and solid blue line is the fit from the bivariate HW model. Plots are ordered by ascending geodesic distance units between sites.

would require a substantial computational effort and time investment. For instance, the estimation of the MLEs of the full spatial Gaussian copula model for the Northern dataset (11 sites) took 17 hours, and for the HW model it took over a day and a half.

Chapter 2

Residual-based CUSUM beta regression control chart for monitoring double-bounded processes

2.1 Introduction

Statistical process control (SPC) is a collection of techniques useful for monitoring and controlling a process (Fournier et al., 2006). Under natural variability, that is, a common variation that will always exist, the process is in statistical control. However, when the variability stems from external sources, the process is out of statistical control. Specialists desire to quickly detect shifts in the process monitoring, thus the control chart is the simplest and most used tool for this purpose (Montgomery, 2009).

The usual control chart proposed by Shewhart (1931) is widely used to monitor independent random variables and can detect large shifts in the mean of a process (Aslam et al., 2014). These charts are also known as memory-less control charts because

they consider observations at a given time. However, not observing and analyzing previous observations can lead to the poor performance of the control chart. Moreover, conventional control charts require some assumptions or approximations. Alternatively, more advanced statistical methods have been proposed in the literature, such as the cumulative sum (CUSUM) (Page, 1954) and the exponentially weighted moving average (EWMA) (Roberts, 1959) control charts. CUSUM and EWMA control charts have been studied by some authors over the years. Some past works can be found in Page (1961); Ewan (1963); Hawkins (1981); Crowder (1989); Lucas and Saccucci (1990); Gan (1991); Woodall and Adams (1993), while some recent developments are found in Park and Jun (2015); Haq (2017); Sanusi et al. (2018); Adegoke et al. (2019); Perry and Wang (2022); Xue and Qiu (2021); Aytaçoğlu et al. (2022). Such charts quickly detect small shifts in the mean of a process and use cumulative information from the observations, thus being called memory-type control charts.

CUSUM control charts can be built using different statistics. Recently, the discussion on residual-based CUSUM control charts has received considerable attention. For example, Asadzadeh et al. (2013) monitored the Cox-Snell residuals of accelerated failure time models based on two regression-adjusted control approaches. Chen and Huang (2014) used a residual-based CUSUM control chart to monitor syndromic data on the respiratory syndrome in Taiwan. Weiß and Testik (2015) discussed the problem of monitoring autocorrelated, count-type discrete data by investigating the CUSUM control chart based on three different residuals. Alencar et al. (2017) and Albarracin et al. (2018) evaluate the performance of the CUSUM control chart using the deviance residual, and other statistics, in the monitoring of negative binomial and negative binomial generalized autoregressive moving average processes (Benjamim et al., 2003), respectively. Kim and Lee (2021) introduced the residual-based CUSUM scheme in first-order Poisson integer-valued autoregressive models where the residuals are computed through squared difference estimates.

In practical situations, there is an interest in modeling and monitoring variables limited to the unit interval $(0, 1)$, such as fractions or proportions. Here, the traditional Shewhart-type control chart may not be adequate since fractions or proportions frequently follow a skew distribution, thus not holding the normality assumption. Alternatively, control charts for double bounded quality characteristics have been proposed in the literature. For example, Sant'Anna and ten Caten (2012) proposed the beta control chart (BCC), where the authors assume that the variable of interest is beta distributed and the control limits are estimated using the CDF of the beta distribution. The BCC is more advantageous than the Shewhart control chart as it naturally captures the asymmetry of the quality characteristic and its control limits range between $(0, 1)$. Lima-Filho et al. (2019) proposed a new control chart to model the mean of double bounded processes in the presence of zeros and ones. The control limits of this chart are based on the inflated beta probability distribution function. As the Kumaraswamy distribution (Jones, 2009) is a good alternative to the beta distribution, Lima-Filho and Bayer (2021) introduced a novel control chart based on the Kumaraswamy distribution to monitor environmental data limited to the unit interval. Nevertheless, production processes can also be related to external variables. For this purpose, Bayer et al. (2018) proposed the beta regression control chart (BRCC), where the control limits are defined using the quantile function of the beta distribution, and Lima-Filho et al. (2020) provided a general framework to a recent control chart based on the inflated beta regression model to monitor the mean of environmental processes containing zeros and ones. Recently, Hwang (2021) proposed a novel CUSUM control chart based on the deviance residual of a beta regression model to monitor the mean of a univariate process.

Although the BCC and BRCC are better alternatives than the conventional charts for double bounded data, they still are memory-less approaches and a CUSUM control chart alternative could detect more quickly a shift in the mean of a process. In this regard, the chief contribution of this paper is to propose and compare the performance of

the CUSUM beta regression control chart on the residuals of the beta regression model, named CUSUM-BRCC. The CUSUM-BRCC is useful for monitoring double bounded processes where the quality characteristic is affected by control variables in which the process output may represent individual measures (e.g. efficiency score) or a ratio between continuous numbers (e.g. relative humidity). As there are different residuals for the beta regression (Espinheira et al., 2008; Pereira, 2019), we explore the CUSUM-BRCC based on the standardized, two types of standardized weighted, and quantile residuals. Because (Hwang, 2021) proposed a similar approach using the deviance residual, we consider the deviance residual in our study in order to compare our proposal with this new control chart in the literature. It is noteworthy, however, that the deviance residual cannot be calculated for several observations in beta regression provided that the contribution of these observations to the deviance is negative. We conduct an extensive Monte Carlo simulation study to evaluate and compare the performance of the proposed control chart based on different residuals in terms of run length (RL) analysis. Our Monte Carlo simulation results show that the quantile residual is the most suitable residual to be considered when controlling and monitoring processes limited to the unit interval $(0, 1)$ and in the presence of external variables.

2.2 On the literature review of control charts

Control charts are powerful tools used to monitor a quality characteristic of interest. A typical control chart is shown in Figure 2.2.1. Usually, this graphical device consists of a centre line (CL), representing the mean of the quality variable corresponding to the in-control state, and two other horizontal lines representing the upper control limit (UCL) and the lower control limit (LCL). When all the sample points fall within the control limits, the process is in a state of control and no further action is needed. If an observation exceeds the desired limits (red dot shown in the plot), the process is

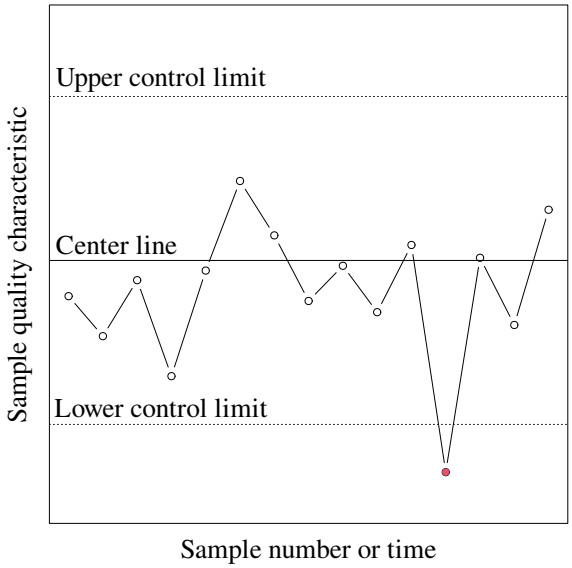


Figure 2.2.1: A typical control chart.

assumed to be out of control and corrections are required to understand the causes and improve the process.

2.2.1 Shewhart control charts

Amongst the several control charts used to monitor the parameters of a manufacturing process, the Shewhart control chart is the most well-known in the literature (Shewhart, 1931). The control chart consists of a CL, which is the mean of the quality variable, and UCL and LCL. For this situation, only one observation is available at a time, thus the Shewhart control chart is defined for independent random variables (Wardell et al., 1992).

Given a quality variable with mean μ and standard deviation σ, we obtain CL, LCL,

and UCL as follows:

$$\text{LCL} = \mu + k\sigma,$$

$$\text{CL} = \mu,$$

$$\text{UCL} = \mu - k\sigma,$$

where k is a constant, expressed in standard deviations, that determines the distance between the lower and upper control limits to the center line. Shewhart control charts are commonly used in normally distributed processes. Therefore, when the assumption of normality is not met, such charts can generate distorted results.

2.2.2 Cumulative Sum (CUSUM) control charts

The CUSUM control chart is a powerful tool able to detect small shifts in the mean of a process using the cumulative sum of the sample deviations, which is plotted in the control chart from a target value μ_0. Thus, the CUSUM control chart is obtained by plotting the following quantities (Montgomery, 2009):

$$C_i = \sum_{j=1}^{i} (\bar{x}_j - \mu_0),$$

where $\bar{x}_j$ is the average of the j-th sample, μ_0 is the target mean value and C_i is the cumulative sum up to and including the i-th sample.

The tabular CUSUM, on the other hand, is built by accumulating deviations from μ_0. Deviations above μ_0 are accumulated from the statistic C^+ and deviations below μ_0 from the statistic C^-. These statistics are called one-sided upper and lower CUSUMs

and are defined as (Montgomery, 2009):

$$C_t^+ = \max[0, x_i - (\mu_0 + K) + C_{t-1}^+], \tag{2.2.1}$$

$$C_t^- = \max[0, (\mu_0 - K) - x_i + C_{t-1}^-], \tag{2.2.2}$$

where the starting values are $C_0^+ = C_0^- = 0$. Here, K is usually a reference value representing the magnitude of the shift we want to detect and is expressed in terms of standard deviations. Note that the control chart accumulates deviations from μ_0 that are greater than K, so any negative CUSUM statistic resets to zero. The decision interval is defined as H, and if either C_t^+ or C_t^- exceed H, then the process is out of control.

2.2.3 Beta regression control charts

In this section, we present the beta distribution, the varying precision beta regression model, and two control charts in the literature based on the beta distribution and beta regression model.

Beta distribution

In the monitoring of processes limited to the unit interval $(0, 1)$, such as fractions or proportions, the beta distribution is frequently used. The beta law is very flexible and can assume a wide variety of shapes, such as symmetric, asymmetric, J-shaped, inverted J-shaped, U-shaped, and uniform. Using a parametrization that considers a location (μ) and a precision (ϕ) parameter, Ferrari and Cribari-Neto (2004) introduce the beta density as

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1, \tag{2.2.3}$$

where $0 < \mu < 1$, $\phi > 0$, and $\Gamma(\cdot)$ is the gamma function. The mean and variance of y are $\mathbb{E}[y] = \mu$ and $\text{Var}(y) = \mu(1-\mu)/(1+\phi)$, respectively.

The CDF of y is given by

$$F(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} \int_0^y y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}\mathrm{d}y. \tag{2.2.4}$$

Therefore, the quantile function of y is expressed as $\psi(u,\mu,\phi) = F^{-1}(u,\mu,\phi)$, where u is the desired quantile.

Varying precision beta regression model

Let $y_1, y_2, \ldots, y_n$ be a set of independent random variables where each y_t, with $t = 1, 2, \ldots, n$, has probability density function (PDF) and CDF given in (2.2.3) and (2.2.4), respectively, with mean μ_t and precision ϕ_t. The varying precision beta regression model is defined as (Simas et al., 2010; Smithson and Verkuilen, 2006):

$$g(\mu_t) = \sum_{i=1}^{r} x_{ti}\omega_i = \eta_{1t},$$
$$h(\phi_t) = \sum_{j=1}^{s} z_{tj}\gamma_j = \eta_{2t},$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_r)^\top \in \mathbb{R}^r$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_s)^\top \in \mathbb{R}^s$ are unknown parameters vectors, $x_{1t}, \ldots, x_{rt}$ and $z_{1t}, \ldots, z_{st}$ are the known and fixed covariates of the mean and precision submodels, respectively, $\boldsymbol{\eta_1} = (\eta_{11}, \ldots, \eta_{1n})^\top$ and $\boldsymbol{\eta_2} = (\eta_{21}, \ldots, \eta_{2n})^\top$ being the mean and precision linear predictor vectors, respectively. Here, $g(\cdot)$ and $h(\cdot)$ denote the link functions that are strictly monotonic and twice differentiable such that $g : (0,1) \mapsto \mathbb{R}$ and $h : (0,\infty) \mapsto \mathbb{R}$. Thus, the mean and precision of each y_t are given,

respectively, by

$$\mu_t = g^{-1}(\eta_{1t}), \tag{2.2.5}$$

$$\phi_t = h^{-1}(\eta_{2t}). \tag{2.2.6}$$

One can choose the logit, probit, loglog, and cloglog for the mean link function, while for the precision link function we have the log as a common choice.

Let $\boldsymbol{\theta} = (\boldsymbol{\omega}^\top, \boldsymbol{\gamma}^\top)^\top$ be the beta regression parameter vector. The log-likelihood function of $\boldsymbol{\theta}$ is given by

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n} \ell_t(\mu_t, \phi_t), \tag{2.2.7}$$

where

$$\begin{aligned}\ell_t(\mu_t, \phi_t) = &\log\Gamma(\phi_t) - \log\Gamma(\mu_t\phi_t) - \log\Gamma((1-\mu_t)\phi_t) + (\mu_t\phi_t - 1)\log y_t \\ &+ [(1-\mu_t)\phi_t - 1]\log(1-y_t),\end{aligned}$$

with μ_t and ϕ_t defined in (2.2.5) and (2.2.6), respectively.

By taking first-order derivatives of (2.2.7) with respect to each element of $\boldsymbol{\theta}$, we obtain a system of equations called score vector. By setting the score vector equals zero, we obtain the MLEs of the model parameters. Since the MLEs do not have closed-form expressions, we maximize the log-likelihood function in (2.2.7) using numerical methods. It is noteworthy that using the MLE of $\boldsymbol{\theta}$ gives us $\hat{\eta}_{1t}$ and $\hat{\eta}_{2t}$, which are used to obtain $\hat{\mu}_t$ and $\hat{\phi}_t$. These quantities are essential in the construction of the residuals considered in the present work. More details on inferences in beta regression models can be found in Ferrari and Cribari-Neto (2004) and Cribari-Neto and Zeileis (2010).

Beta control chart

The beta control chart was proposed by Sant'Anna and ten Caten (2012) and is based on the beta distribution in (2.2.3) and (2.2.4). The LCL and UCL of the beta control chart are given by:

$$\text{LCL} = \bar{p} + w_1\sqrt{s^2(\bar{p})}, \tag{2.2.8}$$

$$\text{UCL} = \bar{p} - w_2\sqrt{s^2(\bar{p})}, \tag{2.2.9}$$

where $\bar{p}$ and $s^2(\bar{p})$ are the mean and variance of the proportion variable, and w_1 and w_2 are constants that specify the width of the control limits. These constants are given by:

$$w_1 = \bar{p} - \frac{\psi([\alpha/2], \mu, \phi)}{\sqrt{s^2(\bar{p})}}, \tag{2.2.10}$$

$$w_2 = \frac{\psi([1-\alpha/2], \mu, \phi)}{\sqrt{s^2(\bar{p})}} - \bar{p}, \tag{2.2.11}$$

for a control region $1-\alpha$. By replacing (2.2.10) in (2.2.8) and (2.2.11) in (2.2.9), we obtain

$$\text{LCL} = \psi([\alpha/2], \mu, \phi),$$

$$\text{UCL} = \psi([1-\alpha/2], \mu, \phi).$$

Although it seems easy and simple, the beta control chart cannot accomodate situations where a regression structure is imposed.

Beta regression control chart

For monitoring double bounded processes with covariates, based on the beta regression model, Bayer et al. (2018) proposed the beta regression control chart, where the control

limits are based on the quantile function and evaluated under the estimated parameters in a Shewhart fashion. The limits of the BRCC are defined as:

$$\text{LCL} = \psi([\alpha/2], \mu_t, \phi_t),$$
$$\text{UCL} = \psi([1-\alpha/2], \mu_t, \phi_t),$$

where μ_t and ϕ_t are given by (2.2.5) and (2.2.6), respectively, and α is the fixed false alarm probability. To obtain the MLEs of μ and ϕ, the log-likelihood function in (2.2.7) is maximised. The authors show that the BRCC works well for monitoring fractions and proportions and easily detects changes in the mean of the process.

2.3 Residual-based CUSUM beta regression control chart

When the interest response variable depends on control variables, one can apply a conventional control chart to the residuals of the model as they approximately follow a normal distribution and are independently distributed if the fitted model is well specified (Montgomery, 2009). However, it is well known that CUSUM control charts are more sensitive to detect shifts than Shewhart-type charts (Lucas, 1976). In this way, this section proposes the residual-based CUSUM beta regression control chart to model the process mean.

As discussed in Espinheira et al. (2008), Pereira (2019) and Ferrari and Cribari-Neto (2004), there are different residuals for the beta regression model. The residuals we consider are the standardized and deviance residuals defined by Ferrari and Cribari-Neto (2004), the standardized weighted 1 and 2 residuals introduced by Espinheira et al. (2008), and the quantile residual (Dunn and Smyth, 1996) considered in the beta regression model by Pereira (2019). Although Anholeto et al. (2014) proposed adjusted

Pearson residuals for beta regressions, we do not consider such residuals because we restricted the work to residuals which are already implemented within R.

Residuals are usually used to check how well a model can fit the observed data, meaning that if a residual is close to zero, we can say that the estimated value is significantly close to the observed value. The standardized residual, for example, is the simplest residual and measures the strength of the difference between observed and expected values. The standardized residual is given by

$$r_t^s = \frac{y_t - \hat{\mu}_t}{\sqrt{\widehat{\text{Var}}(y_t)}}, \tag{2.3.1}$$

where $\widehat{\text{Var}}(y_t) = \hat{\mu}_t(1-\hat{\mu}_t)/(1+\hat{\phi}_t)$ and $\hat{\mu}_t = g^{-1}(\hat{\eta}_{1t})$.

The two types of standardized weighted residuals proposed by Espinheira et al. (2008) better approximate to the standard normal distribution compared to the standardized residual, thus being more reliable and adequate. The standardized weighted 1 residual is given by:

$$r_t^w = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{\hat{v}_t^*}}, \tag{2.3.2}$$

where $y_t^* = \log(y_t/(1-y_t))$, $\hat{\mu}_t^* = \hat{\text{E}}(y_t^*) = \psi(\hat{\mu}_t\hat{\phi}_t) - \psi((1-\hat{\mu}_t)\hat{\phi}_t)$, $\hat{v}_t^* = \widehat{\text{Var}}(y_t^*) = \psi'(\hat{\mu}_t\hat{\phi}_t) + \psi'((1-\hat{\mu}_t)\hat{\phi}_t)$, ψ and ψ' denoting the digamma and trigamma functions, respectively. The standardized weighted 2 residual, which also has a distribution that better approximates to the standard normal distribution than that of the standardized residual, is the most efficient in identifying large influence on the estimates of the parameters of the model. This residual is computed by:

$$r_t^{ww} = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{\hat{v}_t^*(1-v_{tt})}}, \tag{2.3.3}$$

where v_{tt} is the t-th element of $V = \widehat{W}^{1/2}X(X^\top\widehat{W}X)^{-1}X^\top\widehat{W}^{1/2}$, $X = (x_1, \ldots, x_n)^\top$ is the regressor matrix and $\widehat{W} = \text{diag}(\hat{w}_1, \ldots, \hat{w}_n)$, with $\hat{w}_t = \hat{\phi}_t^2\hat{v}_t^*\left[1/\left\{g'(\hat{\mu}_t)\right\}^2\right]$.

The quantile residual is a simple and generic residual and is mostly used in complex regression models, such as the generalised additive models for location, scale, and shape (Rigby and Stasinopoulos, 2005). Its distribution better approximates to the standard normal distribution when compared to the standardized weigthed 1 and 2 residuals. In Pereira (2019), the results suggest that the quantile residual performs better for large to moderate sample sizes in diagnostic analysis in beta regression as well as it can be used for any link function. This residual is defined as

$$r_t^q = \Phi^{-1}\{F(y_t; \hat{\mu}_t, \hat{\phi}_t)\}, \tag{2.3.4}$$

where $\Phi(\cdot)$ denotes the CDF of the standard normal distribution and $F(\cdot)$ is the CDF of the beta distribution in (2.2.4).

Finally, the deviance residual used in the CUSUM-BRCC$_{\text{Hwang}}$ and in our study it is defined as

$$r_t^d = \text{sign}(y_t - \hat{\mu}_t)\{2|(\ell_t(\tilde{\mu}_t, \hat{\phi}_t) - \ell_t(\hat{\mu}_t, \hat{\phi}_t))|\}^{1/2}.$$

Here, sign$(\cdot)$ is the signal function, $\ell_t(\mu_t, \phi_t)$ is the contribution of the t-th observation of the log-likelihood function given in (2.2.7), $\hat{\mu}_t$ and $\hat{\phi}_t$ are the MLEs of μ_t and ϕ_t, respectively, and $\tilde{\mu}_t$ being the estimate of μ_t in the saturated model (maximum log-likelihood achievable). It is noted by Ferrari and Cribari-Neto (2004) that for large values of ϕ, $\tilde{\mu}_t \approx y_t$, so we replace $\ell_t(\tilde{\mu}_t, \hat{\phi}_t)$ by $\ell_t(y_t, \hat{\phi}_t)$. Note that this residual considers the absolute value of the contribution of each observation to the deviance, which is not reasonable for defining a residual. Espinheira et al. (2008) also discussed the drawbacks of the deviance residual and recommended that it is not used in the beta regression.

Based on the regression structures defined in (2.2.5) and (2.2.6) and on the residuals given in (2.3.1), (2.3.2), (2.3.3), and (2.3.4), we propose the CUSUM-BRCC using each of the residuals mentioned, thus resulting in four new control charts.

The CUSUM-BRCC statistics are similar to the ones in (2.2.1) and (2.2.2). They are given by:

$$C_t^+ = \max[0, r_t - (m_0 + K) + C_{t-1}^+],$$
$$C_t^- = \max[0, (m_0 - K) - r_t + C_{t-1}^-],$$

where r_t is the t-th observation of each residual considered, m_0 is the target mean value, that is, the residual mean, K is a reference value, and $C_0^+ = C_0^- = 0$ are the starting values. The reference value is given by $K = k \times \sigma$, where σ is the standard deviation of the residual used to build the CUSUM chart, and the choice of k is related to the magnitude of the change that we want to identify, that is, $k = \frac{1}{2} \times \Delta$, where Δ is the size of the shift in standard deviation units. Here, the decision interval of the CUSUM control chart is expressed as $H = h \times \sigma$ where h is estimated for each residual. Observations of the CUSUM statistics that exceed H show that the process is out of control.

2.4 Simulation study

In what follows, we shall present the results of a Monte Carlo simulation study we performed to evaluate and compare the CUSUM-BRCC with the BRCC proposed by Bayer et al. (2018) and the CUSUM-BRCC proposed by Hwang (2021). We considered the residuals in (2.3.1), (2.3.2), (2.3.3), and (2.3.4) to be used in the CUSUM-BRCC proposed in this paper, and $n \in \{100, 200, 500\}$. For brevity and similarity of results, we only present numerical evidence for $n = 500$ based on 5,000 Monte Carlo replications. All simulations were performed using the `R` programming language (R Core Team,

2022).

The performance of control charts is usually measured in terms of RL analysis. The average run length (ARL) is the average number of observations that must be plotted until the control chart signals. Thus, for an in-control process, this measure is known as ARL_0, whereas for an out-of-control process it is named ARL_1. The latter case means that the mean has shifted, then a smaller number of samples would allow the detection of the shift more quickly. In this work, we considered the ARL, median run length (MRL), and standard deviation run length (SDRL) for a process in control, where $\text{ARL} = 1/\alpha$, $\text{MRL} = \ln(0.5)/\ln(1-\alpha)$, and $\text{SDRL} = \sqrt{(1-\alpha)/\alpha^2}$ (Lee Ho et al., 2019; Lima-Filho et al., 2019). Here, α is the false alarm probability, that is, the probability of a single observation falling outside the control limits when the process is in control. Therefore, assuming that the process is in control and $\alpha = 0.005$, we obtain $\text{ARL}_0 = 200$, i.e. we expect an out-of-control signal every 200 samples, on average, even when the process remains in control. The nominal values of MRL and SDRL for a process with the same characteristics are 138.3, and 199.5, respectively. In order for the proposed control charts to present the same target in-control ARL, we first calibrated them using Algorithm 1 to find the optimal h for each residual. To the best of our knowledge, this calibration is not needed for the CUSUM-$\text{BRCC}_{\text{Hwang}}$, therefore we used the approximation given by Siegmund (1985) to obtain the optimal h for a target $\text{ARL}_0 = 200$.

To evaluate the RL when the process is out of control, we introduced a δ change in the mean regression structure of the process. In the true data generation process, we used the following beta regression model:

$$\text{logit}(\mu_t) = \delta + \omega_0 + \omega_1 x_t,$$

$$\log(\phi_t) = \gamma_0 + \gamma_1 z_t,$$

Table 2.4.1: True parameter values for the scenarios considered in the simulation study.

Scenario	ω_0	ω_1	γ_0	γ_1
1	-3.2	2.0	3.0	1.0
2	-3.2	2.0	4.0	0.5
3	-1.0	2.0	3.0	1.5
4	-1.0	2.0	2.0	2.0
5	1.0	2.4	2.0	3.0
6	1.0	2.4	4.0	2.5

where $t = 1, \ldots, n$, δ ranging from -0.5 to 0.5 by steps of 0.1, ω_0, ω_1, γ_0, and γ_1 being the regression coefficients. Since δ is the induced change in the mean, the process is in control when $\delta = 0$. The values of x_t and z_t were obtained from a uniform distribution in the interval $(0, 1)$ and considered constant through all Monte Carlo replications.

Table 2.4.1 shows six scenarios of the true parameter values with different characteristics considered in the numerical evaluation. In Scenarios 1 and 2, the mean is close to 0.1, with $\phi \in [20, 54]$ in Scenario 1 and $\phi \in [55, 90]$ in Scenario 2. In Scenarios 3 and 4, we have $\phi \in [20, 90]$ and $\phi \in [7, 54]$, respectively, and the mean is centered on the standard unit interval. Finally, in Scenarios 5 and 6, the mean is close to 0.9 with $\phi \in [7, 147]$ in Scenario 5 and $\phi \in [55, 659]$ in Scenario 6. Note that we covered a wide range of scenarios for the mean and precision of the process. The Monte Carlo simulation study is divided into two procedures and summarized by Algorithms 1 and 2.

Tables 2.4.2 and 2.4.3 present results for $\widehat{\text{ARL}}$, $\widehat{\text{MRL}}$, and $\widehat{\text{SDRL}}$ for all scenarios. Notice that the control charts obtained similar performance when the process was in control ($\delta = 0$) and presented estimates close to their nominal values, except for CUSUM-BRCC$_{\text{Hwang}}$, which presented the worst performance among the control charts studied. We emphasize that the proposed control chart was calibrated (Algorithm 1) for each residual in order to have a target ARL_0 of 200. We observe in Scenario 1 that, when considering CUSUM-BRCC$_{r_t^w}$ (Equation (2.3.2)), we obtained $\widehat{\text{ARL}}_0 = 222.52$

Algorithm 1: Algorithm for estimating h in forming the proposed control charts.

1. Define the desired probability of false alarm α (herein we used $\alpha = 0.005$);
2. Generate n observations from a standard uniform distribution $(0, 1)$;
3. Using covariates and parameter values, compute μ_t and ϕ_t;
4. Generate n beta-distributed observations with parameters μ_t and ϕ_t;
5. Fit the beta regression model with varying precision and obtain the residuals in (2.3.1), (2.3.2), (2.3.3), and (2.3.4);
6. Obtain the CUSUM-BRCC for each residual in step 5 using H as decision interval;
7. Plot each data point r_t together with the control limits, for $t = 1, \ldots, n$. The observation r_t that is out of the control limits interval is an out-of-control observation;
8. Repeat steps 4 to 7 a large number of times, say 5,000;
9. At the end of the replications, the in-control average run length is calculated, obtaining $\widehat{\text{ARL}}_0$;
10. Repeat steps 4 to 9 for different values of H;
11. Fit a linear regression where H is the response variable and the logarithm of the estimated ARL_0 is the control variable to obtain the exact value of H that yields the desired ARL_0 (herein we chose $\text{ARL}_0 = 200$).

Algorithm 2: Algorithm for the performance evaluation of the proposed control charts.

1. Use the same probability of false alarm α defined in Algorithm 1;
2. Generate n observations from a standard uniform distribution $(0, 1)$;
3. Using covariates and parameter values, compute μ_t and ϕ_t;
4. Generate n beta-distributed observations with parameters μ_t and ϕ_t;
5. Fit the beta regression model with varying precision and obtain the residuals in (2.3.1), (2.3.2), (2.3.3), and (2.3.4);
6. Obtain the CUSUM-BRCC for each residual in step 5 using the estimated value of h obtained from Algorithm 1 as decision interval;
7. Plot each data point r_t together with the control limits, for $t = 1, \ldots, n$. The observation r_t that is out of the control limits interval is an out-of-control observation;
8. Repeat steps 4 to 7 a large number of times, say 5,000;
9. At the end of the replications, the average of each measure is calculated, obtaining the following Monte Carlo estimates: $\widehat{\text{ARL}}$, $\widehat{\text{MRL}}$, and $\widehat{\text{SDRL}}$;
10. Include a shift (δ) in the linear predictor of the mean after fitting the beta model in step 5 and repeat steps 6 to 9 for different values of δ.

when $\text{ARL}_0 = 200$ was expected, that is, in the worst scenario, there was a distortion of 11%.

When the process was out of control, the CUSUM-BRCC presented smaller values of $\widehat{\text{ARL}}_1$ than the BRCC, evidencing that the proposed control chart is more sensitive to detect changes in the mean of the variable of interest. For example, in Table 2.4.2 for Scenario 2 and $\delta = 0.1$, the CUSUM-BRCC$_{r_t^q}$ signaled at sample 16, while the BRCC took on average 163 samples to detect an out-of-control observation. The CUSUM-BRCC$_{\text{Hwang}}$ signaled at sample 8, however, this is not an accurate estimate since the control chart presented a distorted ARL for an in-control process. The exception is for the CUSUM-BRCC$_{r_t^s}$ in Scenarios 1 and 5 for $\delta = -0.1$ and $\delta = 0.1$, respectively, which yielded an $\widehat{\text{ARL}}_1$ greater than 200. This control chart tend to be ARL-biased in the sense that some out-of-control ARL values are larger than the in-control ARL (Paulino et al., 2016).

According to the results obtained, it is important to highlight that all proposed charts and the BRCC presented a similar performance when the process was in control. However, the performance of the proposed control chart was far superior when the process was out of control. For example, in Scenario 3 and $\delta = -0.1$, the BRCC took on average 131 samples to detect a change in the process while the proposed CUSUM-BRCC, considering all residuals, took on average 8 samples to detect a change (approximately 16 times faster). Comparing the CUSUM-BRCC using the quantile residual with the CUSUM-BRCC using the other residuals when the process shifted, we note from Tables 2.4.2 and 2.4.3 that the CUSUM-BRCC$_{r_t^q}$ outperformed in Scenarios 1, 2 and 4 for $\delta = 0.1$ while in the other scenarios its performance was quite similar to that of using the other residuals. For a negative shift, the CUSUM-BRCC$_{r_t^q}$ presented better performance in Scenario 5. Although the CUSUM-BRCC using the standardized residual performed better in some scenarios, the distribution of such residual is not well approximated by the standard normal distribution Espinheira et al. (2008) compared

Table 2.4.2: Performance of the BRCC and CUSUM-BRCC considering different residuals with $\alpha = 0.005$ for Scenarios 1, 2, and 3.

Scenario	δ	BRCC	CUSUM-BRCC						BRCC	CUSUM-BRCC						BRCC	CUSUM-BRCC					
			Hwang	r_t^{ww}	r_t^{w}	r_t^{s}	r_t^{q}	r_t^{d}		Hwang	r_t^{ww}	r_t^{w}	r_t^{s}	r_t^{q}	r_t^{d}		Hwang	r_t^{ww}	r_t^{w}	r_t^{s}	r_t^{q}	r_t^{d}
					$\widehat{\text{ARL}}$							$\widehat{\text{MLR}}$								$\widehat{\text{SDRL}}$		
1	−0.5	19.86	1.02	1.02	1.02	1.05	1.03	1.03	13.41	0.18	0.18	0.18	0.23	0.19	0.19	19.35	0.14	0.16	0.16	0.23	0.16	0.17
	−0.4	31.48	1.04	1.04	1.04	1.12	1.05	1.06	21.47	0.21	0.22	0.22	0.31	0.23	0.24	30.97	0.20	0.21	0.21	0.36	0.23	0.24
	−0.3	51.71	1.22	1.17	1.17	1.92	1.23	1.31	35.50	0.41	0.36	0.36	0.94	0.41	0.48	51.21	0.52	0.45	0.45	1.33	0.53	0.64
	−0.2	89.03	3.40	3.12	3.12	14.30	3.86	4.96	61.37	1.99	1.79	1.80	9.56	2.31	3.08	88.53	2.86	2.57	2.57	13.79	3.32	4.43
	−0.1	148.92	18.79	**24.88**	24.89	244.67	34.21	50.97	102.88	12.67	16.90	16.90	169.24	23.37	34.98	148.42	18.28	24.37	24.38	244.17	33.71	50.46
	0.0	199.97	47.65	222.44	222.52	195.47	**200.40**	207.56	138.26	32.68	153.84	153.89	135.14	138.56	143.52	199.47	47.15	221.94	222.02	194.96	199.90	207.05
	0.1	186.73	14.10	99.82	99.89	**28.86**	39.40	32.94	129.09	9.42	68.84	68.89	19.65	26.96	22.48	186.23	13.59	99.32	99.39	28.35	38.90	32.44
	0.2	130.50	2.71	6.13	6.14	3.48	3.93	3.68	90.11	1.50	3.89	3.90	2.04	2.36	2.19	130.00	2.15	5.61	5.62	2.94	3.40	3.14
	0.3	77.57	1.12	1.24	1.24	1.17	1.17	1.17	53.42	0.31	0.43	0.43	0.36	0.21	0.35	77.06	0.36	0.55	0.55	0.44	0.44	0.44
	0.4	45.11	1.03	1.05	1.05	1.04	1.04	1.04	30.92	0.19	0.22	0.22	0.21	0.18	0.21	44.60	0.16	0.22	0.22	0.20	0.20	0.19
	0.5	26.45	1.01	1.03	1.03	1.02	1.02	1.02	17.99	0.16	0.19	0.19	0.18	0.19	0.18	25.95	0.12	0.16	0.16	0.15	0.15	0.14
					$\widehat{\text{ARL}}$							$\widehat{\text{MRL}}$								$\widehat{\text{SDRL}}$		
2	−0.5	11.21	1.01	1.01	1.01	1.02	1.01	1.01	7.42	0.15	0.16	0.16	0.17	0.16	0.16	10.70	0.10	0.11	0.11	0.14	0.12	0.12
	−0.4	18.62	1.02	1.02	1.02	1.03	1.02	1.02	12.55	0.17	0.18	0.18	0.20	0.18	0.18	18.11	0.13	0.14	0.14	0.17	0.15	0.15
	−0.3	34.04	1.03	1.04	1.04	1.07	1.04	1.04	23.24	0.20	0.21	0.21	0.25	0.22	0.22	33.53	0.18	0.21	0.21	0.27	0.21	0.22
	−0.2	67.04	1.33	1.36	1.36	1.98	1.44	1.46	46.12	0.50	0.52	0.52	0.99	0.59	0.60	66.54	0.67	0.70	0.70	1.40	0.80	0.82
	−0.1	133.69	8.50	**12.38**	12.39	42.52	15.20	16.54	92.32	5.54	8.23	8.24	29.12	10.19	11.12	133.19	7.99	11.87	11.88	42.01	14.69	16.04
	0.0	200.53	48.41	206.27	206.27	**202.40**	194.52	205.80	138.65	33.21	142.63	142.63	139.94	134.49	142.30	200.03	47.91	205.77	205.77	201.90	194.02	205.29
	0.1	163.43	8.09	24.50	24.52	**12.34**	15.96	15.37	112.93	5.25	16.64	16.65	8.20	10.71	10.30	162.93	7.58	24.00	24.02	11.83	15.45	14.86
	0.2	88.52	1.24	1.44	1.44	1.30	1.34	1.33	61.01	0.42	0.59	0.59	0.47	0.51	0.50	88.02	0.54	0.80	0.80	0.62	0.68	0.66
	0.3	42.96	1.03	1.04	1.04	1.03	1.04	1.04	29.43	0.19	0.21	0.21	0.20	0.21	0.21	42.45	0.16	0.21	0.21	0.19	0.19	0.19
	0.4	21.49	1.01	1.02	1.02	1.02	1.02	1.02	14.55	0.16	0.18	0.18	0.17	0.17	0.17	20.99	0.11	0.14	0.14	0.13	0.13	0.13
	0.5	11.61	1.01	1.01	1.01	1.01	1.01	1.01	7.69	0.14	0.16	0.16	0.15	0.15	0.15	11.10	0.09	0.11	0.11	0.10	0.11	0.11
					$\widehat{\text{ARL}}$							$\widehat{\text{MRL}}$								$\widehat{\text{SDRL}}$		
3	−0.5	7.27	1.01	1.01	1.01	1.01	1.01	1.01	4.68	0.14	0.15	0.15	0.15	0.15	0.15	6.75	0.08	0.10	0.10	0.10	0.10	0.10
	−0.4	13.15	1.01	1.01	1.01	1.01	1.01	1.01	8.76	0.15	0.16	0.16	0.16	0.16	0.16	12.64	0.10	0.12	0.12	0.12	0.12	0.12
	−0.3	26.50	1.02	1.02	1.02	1.02	1.02	1.02	18.02	0.17	0.19	0.19	0.18	0.18	0.18	25.99	0.13	0.16	0.16	0.16	0.16	0.16
	−0.2	58.56	1.07	1.09	1.09	1.09	1.09	1.09	40.24	0.25	0.28	0.28	0.28	0.28	0.28	58.05	0.27	0.32	0.32	0.32	0.32	0.32
	−0.1	130.71	4.96	8.01	8.01	8.10	**7.89**	7.97	90.25	3.08	5.20	5.20	5.26	5.11	5.17	130.20	4.43	7.49	7.49	7.59	7.37	7.45
	0.0	202.12	48.79	194.24	194.16	193.74	185.39	**194.66**	139.75	33.47	134.29	134.23	133.94	128.16	134.58	201.62	48.28	193.73	193.66	193.24	184.89	194.16
	0.1	132.96	5.00	8.10	8.10	8.15	**7.94**	8.07	91.82	3.10	5.26	5.26	5.30	5.15	5.24	132.46	4.47	7.58	7.58	7.64	7.43	7.55
	0.2	58.89	1.07	1.09	1.09	1.10	1.09	1.10	40.47	0.25	0.28	0.28	0.29	0.28	0.28	58.39	0.27	0.32	0.32	0.33	0.32	0.32
	0.3	26.75	1.02	1.02	1.02	1.02	1.02	1.02	18.20	0.17	0.18	0.18	0.19	0.18	0.19	26.25	0.13	0.16	0.16	0.16	0.16	0.16
	0.4	13.56	1.01	1.01	1.01	1.01	1.01	1.01	9.05	0.15	0.16	0.16	0.16	0.16	0.16	13.05	0.10	0.12	0.12	0.12	0.12	0.12
	0.5	7.69	1.01	1.01	1.01	1.01	1.01	1.01	4.98	0.14	0.15	0.15	0.15	0.15	0.15	7.18	0.08	0.10	0.10	0.10	0.10	0.10

Table 2.4.3: Performance of the BRCC and CUSUM-BRCC considering different residuals with $\alpha = 0.005$ for Scenarios 4, 5, and 6.

Scenario	δ	BRCC	CUSUM-BRCC						BRCC	CUSUM-BRCC						BRCC	CUSUM-BRCC					
			Hwang	r_t^{ww}	r_t^{w}	r_t^{s}	r_t^{q}	r_d^{ww}		Hwang	r_t^{ww}	r_t^{w}	r_t^{s}	r_t^{q}	r_d^{ww}		Hwang	r_t^{ww}	r_t^{w}	r_t^{s}	r_t^{q}	r_d^{ww}
		$\widehat{\text{ARL}}$							$\widehat{\text{MRL}}$							$\widehat{\text{SDRL}}$						
	−0.5	16.28	1.01	1.02	1.02	1.02	1.02	1.02	10.93	0.15	0.17	0.17	0.17	0.17	0.17	15.77	0.11	0.13	0.13	0.13	0.13	0.13
	−0.4	27.90	1.02	1.03	1.03	1.03	1.03	1.03	18.99	0.17	0.19	0.19	0.19	0.19	0.19	27.39	0.14	0.16	0.16	0.17	0.16	0.16
	−0.3	49.82	1.05	1.06	1.06	1.07	1.06	1.06	34.19	0.22	0.24	0.24	0.25	0.25	0.24	49.32	0.22	0.26	0.26	0.27	0.26	0.26
	−0.2	91.15	1.59	1.80	1.80	1.95	1.83	1.83	62.84	0.70	0.86	0.86	0.97	0.88	0.87	90.65	0.96	1.20	1.20	1.36	1.23	1.23
	−0.1	158.18	8.96	**17.06**	17.07	19.25	17.35	17.71	109.29	5.86	11.48	11.48	13.00	11.68	11.93	157.68	8.45	16.56	16.56	18.75	16.85	17.21
4	0.0	199.63	48.00	195.60	195.63	203.04	188.62	**197.30**	138.03	32.92	135.23	135.26	140.39	130.40	136.41	199.13	47.50	195.10	195.13	202.54	188.12	196.80
	0.1	156.68	14.54	34.15	34.17	32.46	**32.10**	32.78	108.26	9.72	23.32	23.33	22.15	21.90	22.37	156.18	14.03	33.64	33.66	31.96	31.59	32.28
	0.2	90.72	1.91	2.31	2.31	2.38	2.30	2.32	62.53	0.94	1.23	1.23	1.27	1.22	1.23	90.22	1.32	1.74	1.74	1.81	1.73	1.75
	0.3	49.71	1.05	1.07	1.07	1.07	1.07	1.07	34.11	0.23	0.25	0.25	0.26	0.25	0.26	49.21	0.23	0.27	0.27	0.28	0.27	0.28
	0.4	28.36	1.02	1.03	1.03	1.03	1.03	1.03	19.31	0.18	0.19	0.19	0.19	0.19	0.19	27.85	0.14	0.17	0.17	0.17	0.17	0.17
	0.5	17.01	1.01	1.02	1.02	1.02	1.02	1.02	11.44	0.16	0.17	0.17	0.17	0.17	0.17	16.50	0.11	0.13	0.13	0.13	0.13	0.13
		$\widehat{\text{ARL}}$							$\widehat{\text{MRL}}$							$\widehat{\text{SDRL}}$						
	−0.5	36.39	1.02	1.04	1.04	1.02	1.03	1.03	24.87	0.18	0.21	0.21	0.18	0.20	0.19	35.88	0.15	0.21	0.21	0.16	0.18	0.17
	−0.4	58.18	1.04	1.11	1.11	1.05	1.06	1.06	39.98	0.22	0.30	0.30	0.22	0.25	0.24	57.68	0.21	0.34	0.34	0.22	0.26	0.25
	−0.3	93.73	1.26	1.82	1.81	1.21	1.41	1.35	64.62	0.44	0.87	0.87	0.40	0.53	0.52	93.23	0.57	1.22	1.22	0.51	0.76	0.69
	−0.2	144.13	3.21	9.28	9.27	3.28	5.09	4.58	99.56	1.86	6.08	6.07	1.91	3.17	2.82	143.63	2.67	8.76	8.75	2.73	4.56	4.05
	−0.1	196.62	13.96	84.51	84.34	**22.85**	35.29	32.91	135.94	9.32	58.23	58.11	15.49	25.50	22.46	196.12	13.45	84.01	83.84	22.35	36.78	32.40
5	0.0	200.88	47.61	203.65	203.22	196.43	191.20	**200.22**	138.90	32.65	140.81	140.51	135.81	132.19	138.44	200.38	47.11	203.15	202.72	195.93	190.70	199.72
	0.1	152.88	27.87	49.52	**49.43**	377.42	65.30	83.34	105.62	18.97	33.98	33.91	261.26	44.92	57.42	152.38	27.37	49.02	48.93	376.92	64.80	82.84
	0.2	97.19	6.65	8.81	8.80	31.61	10.06	11.82	67.02	4.25	5.75	5.75	21.56	6.62	7.84	96.69	6.13	8.30	8.29	31.10	9.55	11.31
	0.3	59.42	1.78	1.96	1.95	3.03	1.98	2.13	40.84	0.84	0.97	0.97	1.73	0.99	1.09	58.92	1.18	1.37	1.37	2.48	1.39	1.55
	0.4	36.88	1.09	1.12	1.12	1.18	1.11	1.12	25.22	0.28	0.31	0.31	0.37	0.30	0.31	36.38	0.31	0.36	0.36	0.47	0.35	0.37
	0.5	24.23	1.03	1.04	1.04	1.06	1.04	1.05	16.45	0.20	0.22	0.22	0.25	0.22	0.22	23.72	0.19	0.21	0.21	0.26	0.21	0.22
		$\widehat{\text{ARL}}$							$\widehat{\text{MRL}}$							$\widehat{\text{SDRL}}$						
	−0.5	3.78	1.00	1.01	1.01	1.01	1.01	1.01	2.25	0.13	0.14	0.14	0.13	0.14	0.14	3.24	0.07	0.08	0.08	0.07	0.08	0.08
	−0.4	7.38	1.01	1.01	1.01	1.01	1.01	1.01	4.76	0.14	0.15	0.15	0.14	0.15	0.15	6.86	0.08	0.10	0.10	0.09	0.09	0.09
	−0.3	16.85	1.01	1.02	1.02	1.01	1.02	1.02	11.33	0.15	0.17	0.17	0.16	0.16	0.16	16.34	0.11	0.13	0.13	0.12	0.12	0.12
	−0.2	43.71	1.03	1.04	1.04	1.03	1.04	1.04	29.95	0.19	0.21	0.21	0.20	0.21	0.21	43.21	0.17	0.21	0.21	0.19	0.20	0.20
	−0.1	118.18	2.32	3.46	3.46	**2.51**	2.98	2.97	81.57	1.23	2.04	2.03	1.36	1.70	1.69	117.67	1.75	2.92	2.92	1.95	2.43	2.42
6	0.0	202.18	48.54	189.16	189.22	**203.73**	184.03	192.71	139.80	33.30	130.77	130.81	140.87	127.21	133.23	201.68	48.04	188.66	188.72	203.23	183.53	192.21
	0.1	101.56	2.80	**3.43**	**3.43**	4.91	3.69	3.78	70.05	1.57	2.01	2.01	3.05	2.19	2.25	101.05	2.25	2.89	2.89	4.38	3.15	3.24
	0.2	37.93	1.03	1.04	1.04	1.05	1.04	1.04	25.94	0.20	0.22	0.22	0.23	0.22	0.22	37.43	0.16	0.21	0.21	0.23	0.21	0.21
	0.3	16.02	1.01	1.02	1.02	1.02	1.02	1.02	10.75	0.16	0.17	0.17	0.18	0.17	0.17	15.51	0.11	0.13	0.13	0.14	0.13	0.13
	0.4	8.02	1.01	1.01	1.01	1.01	1.01	1.01	5.20	0.14	0.15	0.15	0.16	0.15	0.15	7.50	0.09	0.10	0.10	0.11	0.10	0.10
	0.5	4.68	1.01	1.01	1.01	1.01	1.01	1.01	2.88	0.13	0.14	0.14	0.15	0.14	0.14	4.15	0.07	0.08	0.08	0.10	0.09	0.09

to its weighted competitors, also this control chart can be ARL-biased as mentioned before.

Note that the variability of the model is directly related to the ability of the control chart to detect changes in the process. For example, in Scenario 4, where we considered a low precision ($\phi \in [7, 54]$) and $\delta = -0.1$, the CUSUM-BRCC$_{r_t^q}$ signaled every 17 samples, on average. On the other hand, when the precision in the process was high ($\phi \in [55, 659]$, Scenario 6), the same control chart took on average 3 samples to detect a change of the same magnitude in the mean.

Finally, the numerical evidence showed that regardless of the value of δ and the scenario studied, the proposed CUSUM-BRCC presented a better behavior compared to the other control charts in the literature. As the quantile residual has proved itself to be a good residual for beta regressions (Pereira, 2019) and our simulation results suggest it performs better in some scenarios and equally in others, we recommend using such residual in the proposed CUSUM-BRCC.

2.5 Applications

In this section, we shall present and discuss two applications to show the applicability of the proposed control chart. As the CUSUM-BRCC$_{\text{Hwang}}$ presented poor perfomance under a controlled scenario, we do not consider this control chart in our applications. We performed the applications using the quantile residual for the CUSUM-BRCC and compare it with the BRCC. The construction of the control charts follows Algorithms 1 and 2 and ARL$_0 = 200$ ($\alpha = 0.005$) for both applications.

Table 2.5.1: Descriptive statistics of the quantitative variables; simulated data.

Variables	min	$Q_{1/4}$	median	mean	$Q_{3/4}$	max	SD
y	0.001	0.073	0.153	0.216	0.306	0.755	0.187
x_1	0.001	0.264	0.511	0.509	0.758	0.999	0.288

2.5.1 Application to simulated data

In the first application, we considered simulated data to better illustrate the studied methodology. We used the following structures for the data generation process:

$$\text{logit}(\mu_t) = -3.2 + 2x_{1t} + 2x_{2t}$$
$$\log(\phi_t) = 3.0 + z_{1t} + z_{2t},$$

where $(x_1, x_2) = (z_1, z_2)$. The values of x_1 (x_2) were obtained from a uniform distribution in the unit interval (Bernoulli distribution with $p = 0.3$). We generated 1000 observations considering the process in a state of control.

Some descriptive statistics about y and x_1 are shown in Table 2.5.1, namely: minimum (min), first quartile ($Q_{1/4}$), median, mean, third quartile ($Q_{3/4}$), maximum (max), and SD. Descriptive statistics for x_2 are not presented because it is a binary covariate. Note that 25% of the response variable (y) does not exceed the value of 0.306, and the largest value is 0.755. The mean and median are 0.216 and 0.153, respectively. For x_1, we have a minimum of 0.001 and a maximum of 0.999.

The simulated data were split into two groups, Phase I (first 500 observations) and Phase II (last 500 observations). In Phase I, we estimated the submodels parameters, and in Phase II we monitored the process. Table 2.5.2 presents the parameter estimates, standard errors (SEs), and p-values for the models adjusted in Phase I. We note that all covariates were significant in both models at 5%, as expected. Before determining the control limits, we performed a diagnostic analysis of the residuals of the model fitted

Table 2.5.2: MLEs, SEs, and p-values for the fitted beta regression model with varying precision; simulated data.

Submodel for μ			
	Estimate	SE	p-value
Intercept	-3.2190	0.0480	< 0.0001
x_1	2.0835	0.0591	< 0.0001
x_2	2.0026	0.0341	< 0.0001
Submodel for ϕ			
	Estimate	SE	p-value
Intercept	3.2316	0.1306	< 0.0001
z_1	0.6345	0.2123	0.0028
z_2	0.8498	0.1412	< 0.0001

in Phase I. Figures 2.5.1 and 2.5.2 show the quantile residuals and quantile-quantile (QQ) plot for the fitted model in Table 2.5.2. Figure 2.5.1 shows that the residuals are randomly distributed around zero and within three deviations from the mean. Similarly, Figure 2.5.2 suggests that there is no violation of the model assumptions as the residuals are in agreement with the 45 degree line. The next step was the calibration of both control charts to have a target ARL_0, then in Phase II (last 500 observations) we introduced a perturbation in the structure of the mean submodel of magnitude $\delta = 0.3$ to assess the power of the chart in detecting changes within a controlled scenario.

Figures 2.5.3 and 2.5.4 present the BRCC and CUSUM-BRCC$_{r_t^q}$ with Phase II data, respectively. The BRCC indicated only one out-of-control point below the lower limit and one out-of-control point above the upper limit while the CUSUM-BRCC$_{r_t^q}$ indicated 494 out-of-control observations (almost all the data in Phase II). It is noteworthy that in this phase a disturbance of $\delta = 0.3$ was introduced in the process mean, that is, the process was out of control.

Figure 2.5.5 presents the performance of the BRCC and CUSUM-BRCC$_{r_t^q}$ with Phase II data. The values were obtained considering the steps of Algorithms 1 and

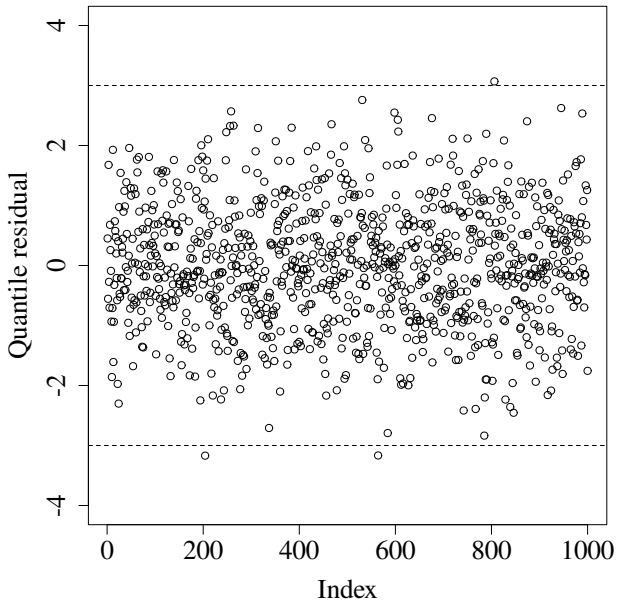


Figure 2.5.1: Quantile residuals; simulated data.

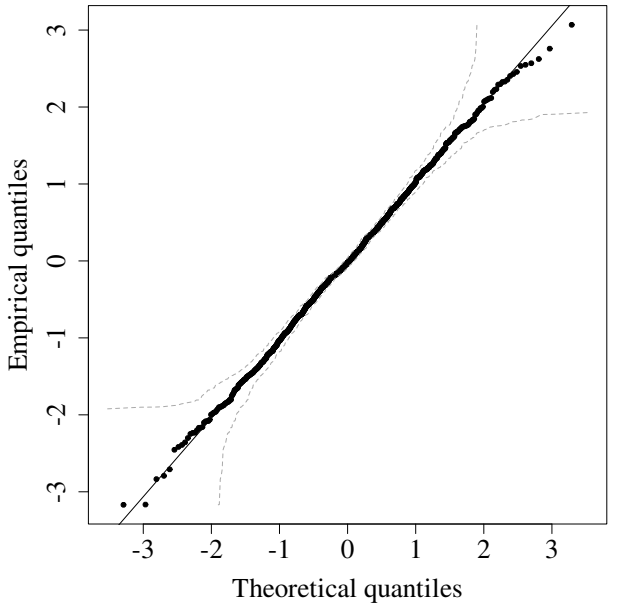


Figure 2.5.2: Quantile-Quantile plot; simulated data.

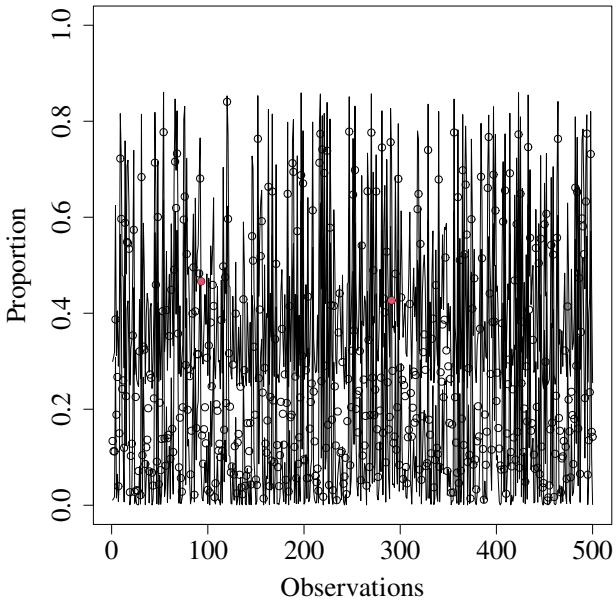


Figure 2.5.3: BRCC for simulated data with out-of-control observations highlighted in red.

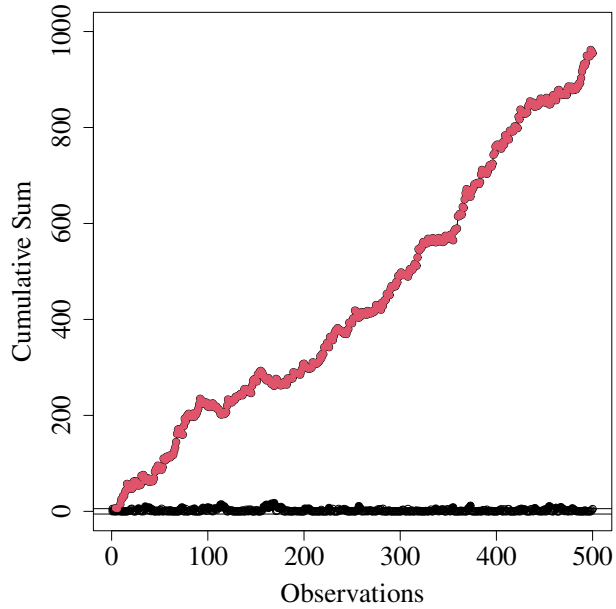


Figure 2.5.4: CUSUM-BRCC$_{r_t^q}$ for simulated data with out-of-control observations highlighted in red.

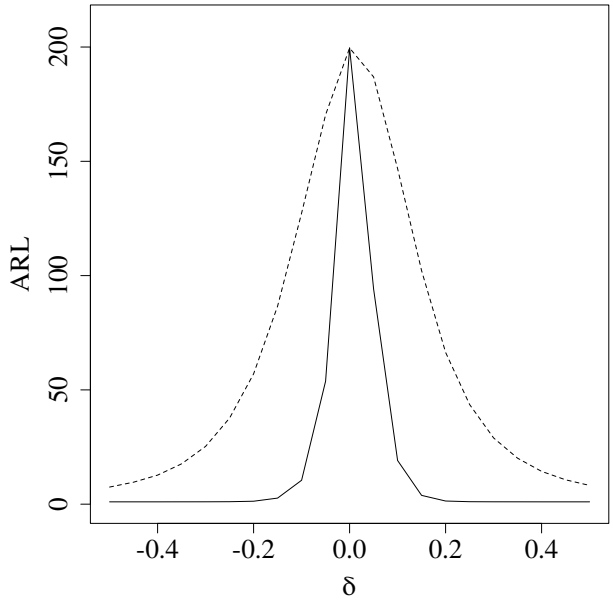


Figure 2.5.5: Performance of the BRCC (dashed line) and CUSUM-BRCC$_{r_t^q}$ (solid line) considering $\text{ARL}_0 = 200$; simulated data.

2. Lastly, the covariates and parameters described in Table 2.5.2 were used in this evaluation. We notice that, as evidenced in the simulation study, the CUSUM-BRCC$_{r_t^q}$ presented better results than the BRCC (smaller ARL_1), proving that the proposed control chart is more sensitive to detect minor changes in the manufacturing process.

2.5.2 Empirical application

In the second application, the dataset refers to the relative humidity (RH) in Australia and highlights the relevance of the proposed chart in monitoring double bounded environmental data. The RH is a ratio between continuous numbers, being the ratio of the partial pressure of water to the equilibrium vapor pressure of water, assuming values in $(0, 1)$. Due to the genesis of the beta regression model, rates and proportions usually can be well fitted by this model. Additionally, the monitoring of RH is relevant because it exerts influence on temperature, rain, and thermal sensation (Lima-Filho and Bayer, 2021).

Table 2.5.3: Description of the variables; relative humidity data.

Variable	Description
RelHumid3pm	Relative humidity (%) at 3pm
Cloud3pm	Fraction of sky obscured by cloud at 3pm.
Evaporation	The so-called Class A evaporation pan (mm) in the 24 hours to 9am
MaxTemp	The maximum temperature in degrees celsius
MinTemp	The minimum temperature in degrees celsius
Pressure3pm	Atmospheric pressure reduced to mean sea level at 3pm
Rainfall	The amount of rainfall recorded for the day in mm
Sunshine	The number of hours of bright sunshine in the day

Table 2.5.4: Descriptive statistics of the quantitative variables; relative humidity data.

Variables	min	$Q_{1/4}$	median	mean	$Q_{3/4}$	max	SD
RelHumid3pm	10.00	43.00	54.00	53.02	63.00	95.00	15.99
Cloud3pm	0.00	1.00	4.00	4.15	7.00	8.00	2.61
Evaporation	0.00	3.20	5.00	5.39	7.20	18.40	2.85
MaxTemp	11.70	20.20	23.30	23.45	26.40	45.80	4.48
MinTemp	5.00	11.30	15.05	15.03	18.90	27.10	4.52
Pressure3pm	994.00	1012.00	1016.00	1016.00	1021.00	1036.00	7.06
Rainfall	0.00	0.00	0.00	2.83	1.00	94.40	8.23
Sunshine	0.00	4.60	8.40	7.40	10.30	13.60	3.75

Table 2.5.3 contains a brief description of the variables used in this analysis. The quality characteristic monitored was measured daily at 3:00pm, and the other variables were used to adjust the beta regression model for the μ and/or ϕ structures. This dataset is available in the `R rattle` package (Graham, 2011) from October 2010 to June 2017 for the Sydney Station in Australia.

Table 2.5.4 includes descriptive statistics of the considered variables. We observe that 25% of the RH does not exceed 43%, the largest prevalence of RH is 95%, and the mean and median values for RH are 54% and 53%, respectively. In the analyzed period, the lowest temperature was 5°C and the highest was 45.8°C.

The dataset has a total of 1690 observations. We used the 845 (50%) first observa-

Table 2.5.5: MLEs, SEs, and p-values for the fitted beta regression model with varying precision; relative humidity data.

Submodel for μ			
	Estimates	SE	p-value
Intercept	-20.6930	2.6299	< 0.0001
Cloud3pm	0.0657	0.0062	< 0.0001
Evaporation	-0.0509	0.0066	< 0.0001
MaxTemp	-0.0729	0.0062	< 0.0001
MinTemp	0.1161	0.0060	< 0.0001
Pressure3pm	0.0204	0.0025	< 0.0001
Rainfall	0.0121	0.0026	< 0.0001
Submodel for ϕ			
	Estimates	SE	p-value
Intercept	-60.4090	7.6854	< 0.0001
MinTemp	0.0284	0.0121	0.0190
Sunshine	0.0765	0.0129	< 0.0001
Pressure3pm	0.0615	0.0075	< 0.0001

tions (October 2010 to December 2014) to estimate the submodels parameters (Phase I). In Phase II, we used the observations from January 2015 to June 2017 to monitor the relative humidity.

Table 2.5.5 shows the parameter estimates, SEs, and p-values for the fitted beta regression model with varying precision. We considered the logit and log link functions in the mean and precision submodels, respectively. Considering the covariates that were statistically significant at the significance level of 5%, two of them were significant in both submodels, namely: minimum temperature (MinTemp) and atmospheric pressure reduced to mean sea level (Pressure3pm).

As in the previous application, we performed a diagnostic analysis of the residuals to check if the beta model is a good fit to the data. Figures 2.5.6 and 2.5.7 show the residuals and QQ plot for the model fitted to RH data. In Figure 2.5.6, the residuals are randomly distributed around zero and within three deviations from the mean. Fig-

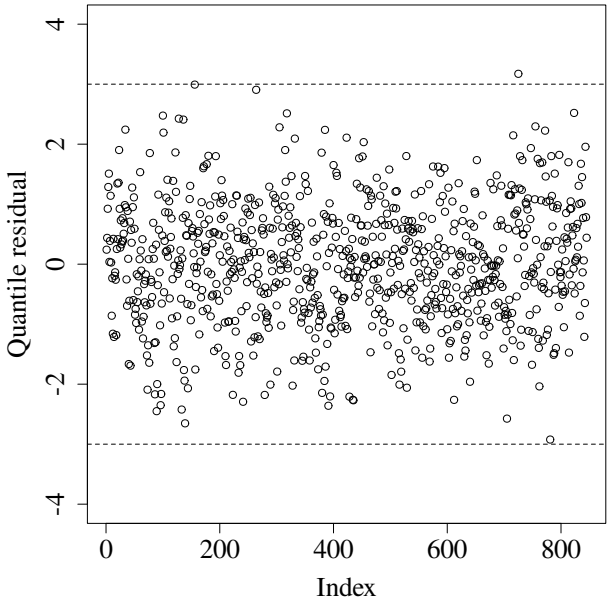


Figure 2.5.6: Quantile residuals; relative humidity data.

ure 2.5.7 displays theoretical quantiles against the empirical quantiles of the residuals. There is no evidence of violation of the model assumptions as the residuals are mostly on the 45 degree line, indicating that this model is a good fit to the data.

Figures 2.5.8 and 2.5.9 show the BRCC and CUSUM-BRCC$_{r_t^q}$ with Phase II data, respectively. Considering the BRCC for monitoring relative humidity, the control chart indicated no more than ten out-of-control points below the lower limit and three points exceeded the upper limit. Differently, the CUSUM-BRCC$_{r_t^q}$ triggered 62 out-of-control points. These results reinforce the characteristic of the CUSUM control chart's power to detect changes in the quality characteristic of interest.

Figure 2.5.10 shows the performance of the BRCC and CUSUM-BRCC$_{r_t^q}$. The construction of the control charts followed the same steps of the application to simulated data. The control charts obtained similar performance when the process was in control ($\delta = 0$); however their performance differed when the process was out of control. The CUSUM-BRCC$_{r_t^q}$ presented smaller values of $\widehat{\text{ARL}}_1$ than its counterpart, evidencing that the proposed control chart is more sensitive to trigger a signal in the quality

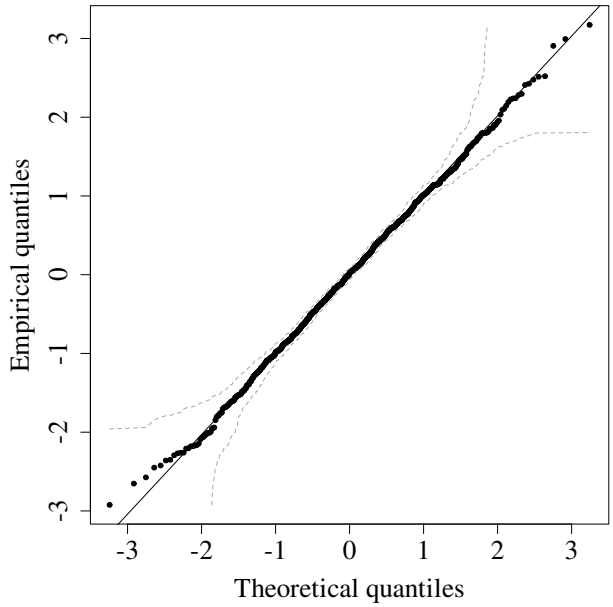


Figure 2.5.7: Quantile-Quantile plot; relative humidity data.

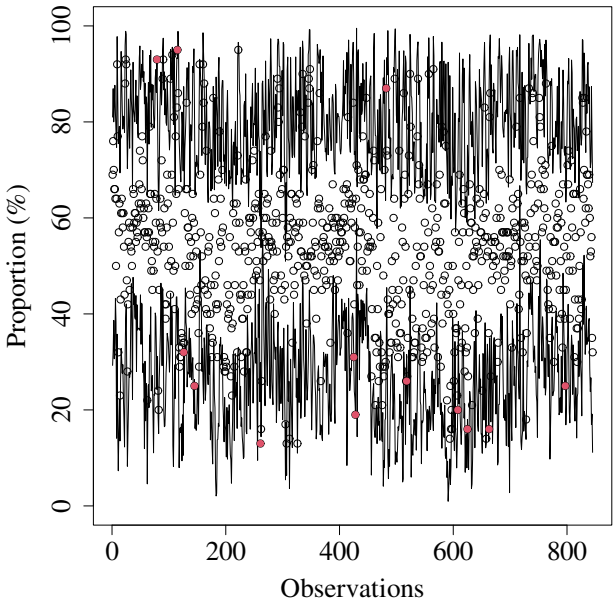


Figure 2.5.8: BRCC for the monitoring of relative humidity in Australia with out-of-control observations highlighted in red.

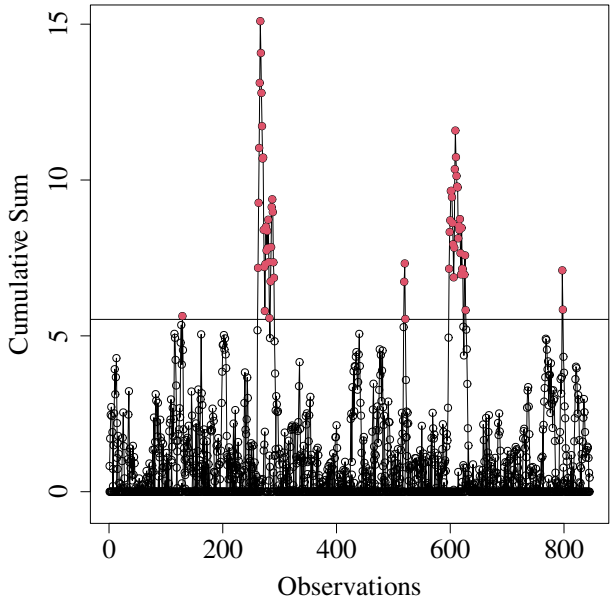


Figure 2.5.9: CUSUM-BRCC$_{r_t^q}$ for the monitoring of relative humidity in Australia with out-of-control observations highlighted in red.

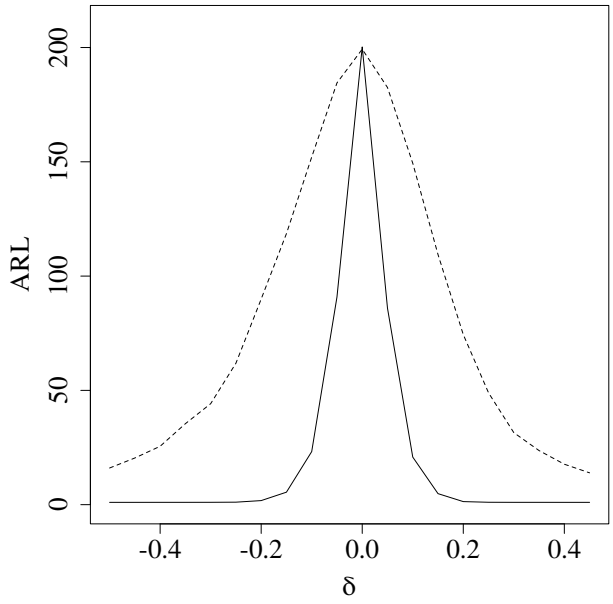


Figure 2.5.10: Performance of the BRCC (dashed line) and CUSUM-BRCC$_{r_t^q}$ (solid line) considering $\text{ARL}_0 = 200$; relative humidity data.

characteristic. For example, considering $\delta = 0.1$, the BRCC took on average 149 samples to detect a change in the process while the CUSUM-BRCC$_{r_t^q}$ took on average 20 samples to detect a change of the same magnitude. In a nutshell, in the presence of control variables, this analysis shows that the proposed CUSUM-BRCC is useful to monitor quality characteristics in the interval $(0, 1)$ in practical situations.

2.6 Concluding remarks

In this paper, we developed a new control chart for monitoring double bounded quality characteristics in the presence of control variables (covariates). For this purpose, we proposed the residual-based CUSUM beta regression control chart considering different residuals of the beta distribution. This control chart has the advantage of accumulating information from the past as well as being more sensitive to detect changes in the mean of a process. We conducted a Monte Carlo simulation study to evaluate and compare the performance of the proposed control chart with two competing control charts in the literature. The numerical results evidenced the superiority of the proposed control chart, presenting values of $\widehat{\text{ARL}}_0$, $\widehat{\text{MRL}}_0$, and $\widehat{\text{SDRL}}_0$ close to their nominal values when the process was in control, and smaller $\widehat{\text{ARL}}_1$ for an out-of-control process. We also presented and discussed applications to real and simulated data that showed the practical importance of our proposal. Finally, we suggest the use of the CUSUM beta regression control chart with the quantile residual when the objective is to monitor double bounded quality characteristics in the presence of control variables and detect small shifts in the mean of the process.

Appendix A

Computational implementation

In this appendix, we present the computer code used to obtain the MLEs and negative log-likelihood estimates of the Gaussian copula model in Section 1.4.2 considering the Southern dataset in Figure 1.2.1, and also the code used in the empirical application in Section 2.5.2. Note that the data used in Section 2.5.2 has been updated so the results might differ from that of the original paper. The files and datasets are available at https://github.com/rauberc/thesis-lu.

```
######################################################################
# PROGRAM: mle-gaussian-copula.R
# USAGE: Computation of the maximum likelihood estimators and
#         negative log-likelihood estimates for the bivariate
#         Gaussian copula
# AUTHOR: Cristine Rauber Oliveira
######################################################################

# loading the packages
library(extRemes)
```

```
library(tidyverse)
library(geodist)
library(foreach)
library(mvnfast)
library(mvtnorm)

#### coordinates and geodesic distance ####
geocoord <- read.csv("coordinates.csv", header = TRUE)

ind_sites <- which(between(geocoord$geolat, 35, 67) &
                   between(geocoord$geolon, -25, 8))
geocoordinates <- data.frame(geocoord$geolon[ind_sites],
                             geocoord$geolat[ind_sites])
names(geocoordinates) <- c("lon", "lat")

dist  <- geodist(geocoordinates, measure = "geodesic")
dist <- dist/100000
ind_dist <- which(upper.tri(dist, diag = FALSE), arr.ind = TRUE)
geodist <-  dist[ind_dist[order(ind_dist[,1]),]]

#### data ####
data <- read.csv("data.csv", header = TRUE)

# obtaining only the nine sites of interest
df_sites <- data[, ind_sites]

# removing rows where there is at least one missing value
```

```
df_final <- df_sites[complete.cases(df_sites),]


# preparing the pairwise datasets
pairlist <- as.list(df_final)
pairwise <- as.list(combn(pairlist, 2))
ind_col_1 <- seq(1, length(pairwise) - 1, by = 2)
ind_col_2 <- ind_col_1 + 1
col_1 <- purrr::map(ind_col_1, ~pairwise[[.x]])
col_2 <- purrr::map(ind_col_2, ~pairwise[[.x]])
length(col_1) == length(col_2)


# function to obtain the dataframes in a pairwise fashion
make_df <- function(col1, col2){
  new_data <- data.frame(col1, col2)
  return(new_data)
}


# complete dataframes in a list
complete_dfs <- purrr::map2(.x = col_1, .y = col_2, make_df)



#### transforming to uniform margins ####
# function to transform the margins of each dataframe to uniform
unif_margins <- function(data){
  data1 <- data[,1]
  data2 <- data[,2]
```

```
vec_unif1 <- numeric(length = length(data1))
vec_unif2 <- numeric(length = length(data2))

q <- 0.95
th1 <- quantile(data1, q)
th2 <- quantile(data2, q)

overth1 <- data1 >= th1
underth1 <- data1 < th1
overth2 <- data2 >= th2
underth2 <- data2 < th2

fitgpd1 <- fevd(data1,  method = "MLE", type = "GP", threshold = th1)
par1 <- fitgpd1$results$par
fitgpd2 <- fevd(data2,  method = "MLE", type = "GP", threshold = th2)
par2 <- fitgpd2$results$par

ranks1 <- rank(data1)/(length(data1)+1)
ranks2 <- rank(data2)/(length(data2)+1)

pgpd_new <- function(data, scale, shape, lambda, threshold){
  p <- pmax(1 + (shape*(data - threshold))/scale, 0)
  p <- 1 - lambda*p^(-1/shape)
  return(p)
}

unifgpd1 <-  pgpd_new(data1, scale = par1[1], shape = par1[2],
```

```
                        lambda = 1 - q, threshold = th1)
  unifgpd2 <-  pgpd_new(data2, scale = par2[1], shape = par2[2],
                        lambda = 1 - q, threshold = th2)
  vec_unif1[underth1] <- ranks1[underth1]
  vec_unif1[overth1] <- unifgpd1[overth1]
  vec_unif2[underth2] <- ranks2[underth2]
  vec_unif2[overth2] <- unifgpd2[overth2]
  unif_mar <- data.frame(vec_unif1, vec_unif2)
  return(unif_mar)
}

# applying the function above to the list of dataframes
unif_df <- purrr::map(complete_dfs, unif_margins)

# censored Gaussian negative log-likelihood
nll_Gaussian <- function(rho, datU, thresh){

  z <- matrix(sapply(datU, qnorm), ncol = ncol(datU), nrow = nrow(datU))
  uz <- qnorm(thresh)
  if (length(uz) == 1) {
    uz <- rep(uz, dim(z)[2])
  }
  else if (length(uz) < dim(z)[2]) {
    stop("Invalid censoring threshold")
  }

  if (rho < -0.9999 || rho > 0.9999) {
```

```
  return(1e+11)
}

Sig <- matrix(c(1, rho, rho, 1), ncol = 2)



if (!exists(".Random.seed", mode = "numeric", envir = globalenv()))
  sample(NA)
oldSeed <- get(".Random.seed", mode = "numeric", envir = globalenv())

ind <-  which(apply(z, 1, function(x) {sum(!is.na(x)) == dim(z)[2]}))

z <- z[ind,]

tmp <- apply(z, 1, function(t) {(sum(t > uz))})

ind_part_cens <- c(1:dim(z)[1])[tmp > 0 & tmp < dim(z)[2]]
ind_full_cens <- c(1:dim(z)[1])[tmp == 0]
ind_no_cens <- c(1:dim(z)[1])[tmp == dim(z)[2]]

if(length(ind_no_cens) > 0){
  ll1 <- -sum(mvnfast::dmvn(z[ind_no_cens, ],
                            mu = rep(0, ncol(z)),
                            sigma = Sig,
                            log = TRUE)) +
    sum(dnorm(z[ind_no_cens,], log = TRUE))
} else{ll1 <- 0}
```

```
ll2 <- foreach::foreach(j = ind_part_cens, .combine = 'c') %dopar% {
  cens <- which(z[j, ] <= uz)
  nocens <- which(z[j, ] > uz)
  Sig11 <- Sig[cens, cens] - Sig[cens, nocens] %*%
    (solve(Sig[nocens, nocens]) %*% Sig[nocens, cens])
  Sig11 <- as.matrix(Sig11)
  if(!isSymmetric.matrix(Sig11)){
    Sig11 <- (Sig11 + t(Sig11))/2
  }
  mu11 <- c(Sig[cens, nocens] %*%
                  (solve(Sig[nocens, nocens]) %*%
                     z[j, nocens]))
  set.seed(123)
  mvnfast::dmvn(z[j, nocens],
                mu = rep(0, length(nocens)),
                sigma = as.matrix(Sig[nocens, nocens]),
                log = TRUE) +
    log(mvtnorm::pmvnorm(upper = uz[cens],
                         mean = mu11, sigma = Sig11)[1]) -
    sum(dnorm(z[j,nocens], log = TRUE))
}
ll2 <- -sum(ll2)

set.seed(123)
ll3 <- -length(ind_full_cens)*log(mvtnorm::pmvnorm(upper = uz,
```

```
                                                    sigma = Sig)[1])

  assign(".Random.seed", oldSeed, envir = globalenv())
  return(ll1 + ll2 + ll3)
}

#### optmisation ####
mle <- matrix(rep(NA, 2), nrow = 1, ncol = 2)

comb <- function(...) {
  mapply('rbind', ..., SIMPLIFY = FALSE)
}

# loop to obtain the MLEs and negative log-likelihood estimates
# for the Gaussian copula
loop <- foreach::foreach(i = 1:length(unif_df),
                         .combine = 'comb',
                         .multicombine = TRUE) %dopar% {
                           data <- unif_df[[i]]
                           data <- as.matrix(data)
                           fit <- optim(0.6,
                                        nll_Gaussian,
                                        method = "BFGS",
                                        thresh = 0.95,
                                        datU = data)

                           mle[, 1] <- fit$par
```

```
                        mle[, 2] <- fit$value
                        list(mle)
                      }

mles <- data.frame(loop)
names(mles) <- c("rho", "nll_Gaussian")

# estimates for each model
#> print(mles)
#     rho       nll_Gaussian
#1  0.9863241  -473.198683
#2  0.9652405  -283.995622
#3  0.9857896  -435.782406
#4  0.8613851   120.186416
#5  0.7097369   312.610705
#6  0.9852277  -425.114979
#7  0.9866764  -430.010106
#8  0.9714839  -303.335174
#9  0.9648448  -276.455859
#10 0.9799662  -366.290839
#11 0.8702448   107.717808
#12 0.7319095   290.910204
#13 0.9795275  -368.535953
#14 0.9728336  -291.429854
#15 0.9687396  -273.059565
#16 0.9794567  -392.146757
#17 0.9131043     9.867028
```

```
#18 0.7461897   273.535119
#19 0.9646458  -275.517008
#20 0.9483830  -154.114057
#21 0.9821864  -395.823026
#22 0.8788641    86.939471
#23 0.7298333   292.799895
#24 0.9866249  -448.133902
#25 0.9772714  -319.506604
#26 0.9836648  -421.980756
#27 0.8159237   185.917185
#28 0.8591979   124.030247
#29 0.8366267   169.222046
#30 0.9059431     9.707709
#31 0.7095187   312.061879
#32 0.6863212   334.474167
#33 0.7474842   272.522087
#34 0.9811539  -345.494980
#35 0.9701906  -281.419806
#36 0.9578660  -191.293670
```

```
######################################################################
# PROGRAM: cusum-betareg-control-chart.R
# USAGE: Computation of the CUSUM beta regression control chart
#         using the quantile residual
# AUTHOR: Cristine Rauber Oliveira
######################################################################

# getting the dataset from the rattle package
data(weatherAUS)
head(weatherAUS)
tail(weatherAUS)

# checking the dimension of the dataset
dim(weatherAUS)

#checking the locations available in this dataset
table(weatherAUS$Location)

# obtaining the data for Sydney, which is the city we are interested in
# analysing the relative humidity
df_rh <- na.omit(weatherAUS[weatherAUS$Location == "Sydney",])
attach(df_rh)

# phase I data: we use these observations to fit the model and get
# the parameter estimates
df_rh_p1 <- df_rh[1:845, c(3,4,5,6,7,15,17,19)]
```

```
# phase II data: we monitor these observations after fitting the model to
# the phase I data and getting the parameter estimates
df_rh_p2 <- df_rh[846:1690, c(3,4,5,6,7,15,17,19)]

# feature names
names(df_rh_p1)

# summary of each feature
summary(df_rh_p1)

# transforming the target to a uniform scale (the beta regression model
# only applies to the target in the unit interval (0,1))
df_rh_p1$Humidity3pm <- df_rh_p1$Humidity3pm/100
df_rh_p2$Humidity3pm <- df_rh_p2$Humidity3pm/100

# fitting the beta regression model to the data in phase I
fit <- betareg(Humidity3pm ~ MinTemp + MaxTemp + Rainfall + Evaporation +
                 Pressure3pm + Cloud3pm | MinTemp + Sunshine +
                 Pressure3pm, data = df_rh_p1)
summary(fit)

# obtaining the design matrices for both submodels
X_mu <- cbind(1, df_rh_p1[,c(1,2,3,4,7,8)])
X_phi <- cbind(1, df_rh_p1[,c(1,5,7)])

# linear predictor for each submodel
```

```
eta_mu <- as.matrix(X_mu)%*%fit$coefficients$mean
eta_phi <- as.matrix(X_phi)%*%fit$coefficients$precision

shape1 <- exp(eta_mu)/(1+exp(eta_mu))
shape2 <- exp(eta_phi)

# parameters of the beta regression distribution
p <- shape1*shape2
q <- shape2-(shape1*shape2)

rh <- df_rh_p1$Humidity3pm

# quantile residual
residual <- qnorm(pbeta(rh, p, q))

# plot of the residuals
plot(residual, caption = NULL, sub.caption = NULL, ylim = c(-4,4),
     ylab = "Quantile residual", xlab = "Index")
abline(h = c(-3,3), lty = 2)
abline(h = 0, col = "red")

# qqplot of the residuals
qqnorm(residual, caption = NULL, main = "", ylab = "Empirical quantiles",
       xlab = "Theoretical quantiles")
qqline(residual)

# design matrices for the data in phase II
```

```
X_mu2 <- cbind(1, df_rh_p2[,c(1,2,3,4,7,8)])
X_phi2 <- cbind(1, df_rh_p2[,c(1,5,7)])

# linear predictor for each submodel
eta_mu2 <- as.matrix(X_mu2)%*%fit$coefficients$mean
eta_phi2 <- as.matrix(X_phi2)%*%fit$coefficients$precision

shape1_2 <- exp(eta_mu2)/(1+exp(eta_mu2))
shape2_2 <- exp(eta_phi2)

p2 <- shape1_2*shape2_2
q2 <- shape2_2-(shape1_2*shape2_2)

# rh of phase II data
rh2 <- df_rh_p2$Humidity3pm

set.seed(1)

# residuals for the phase II data
res2 <- qnorm(pbeta(rh2, p2, q2))

# control limits
control_limits <- cusum(res2, center = mean(residual),
                        std.dev = sd(residual),
                        decision.interval = 3*1.842887,
                        se.shift = 1, plot = F)
```

```
DI <- control_limits$decision.interval
cusum_pos <- as.vector(control_limits$pos)
cusum_neg <- as.vector(control_limits$neg)*(-1)

out_model1 <- c(which(cusum_neg > DI))
out_model2 <- c(which(cusum_pos > DI))

obs1 <- cusum_neg[out_model1]
obs2 <- cusum_pos[out_model2]

# red dots are observations out of control
plot(cusum_pos, xaxs = "r", type = "o", lwd = 1, pch = 1,
     ylab = expression("Cumulative Sum"),
     xlab = expression("Observations"),
     ylim = c(-0.4, 15))
points(cusum_neg, type = "o", xaxs = "r", lwd = 1, pch = 1)
abline(h = DI, lwd = 1, lty = 1)
points(out_model1, obs1, pch = 16, col = "red")
points(out_model2, obs2, pch = 16, col = "red")
```

Bibliography

Adegoke, N. A., Smith, A. N. H., Anderson, M. J., Sanusi, R. A., and Pawley, M. D. M. (2019). Efficient homogeneously weighted moving average chart for monitoring process mean using an auxiliary variable. *IEEE Access*, 7:94021–94032.

Albarracin, O. Y. E., Alencar, A. P., and Lee Ho, L. (2018). CUSUM chart to monitor autocorrelated counts using negative binomial garma model. *Statistical Methods in Medical Research*, 27(9):2859–2871.

Alencar, A. P., Lee Ho, L., and Albarracin, O. Y. E. (2017). CUSUM control charts to monitor series of negative binomial count data. *Statistical Methods in Medical Research*, 26(4):1925–1935.

Allen, J., Sauer, H., Frank, L., and Reiff, P. (1989). Effects of the march 1989 solar activity. *Eos, Transactions American Geophysical Union*, 70(46):1479–1488.

Anderson, C. W., Lanzerotti, L. J., and Maclennan, C. G. (1974). Outage of the l4 system and the geomagnetic disturbances of 4 august 1972. *Bell System Technical Journal*, 53(9):1817–1837.

Anholeto, T., Sandoval, M. C., and Botter, D. A. (2014). Adjusted pearson residuals in beta regression models. *Journal of Statistical Computation and Simulation*, 84(5):999–1014.

Asadzadeh, S., Aghaie, A., and Niaki, S. T. A. (2013). AFT regression-adjusted monitoring of reliability data in cascade processes. *Quality & Quantity*, 47(6):3349–3362.

Aslam, M., Azam, M., and Jun, C.-H. (2014). A new exponentially weighted moving average sign chart using repetitive sampling. *Journal of Process Control*, 24(7):1149–1153.

Aytaçoğlu, B., Driscoll, A. R., and Woodall, W. H. (2022). Controlling the conditional false alarm rate for the MEWMA control chart. *Journal of Quality Technology*, 54(5):487–502.

Baker, D. N., Li, X., Pulkkinen, A., Ngwira, C. M., Mays, M. L., Galvin, A. B., and Simunac, K. D. C. (2013). A major solar eruptive event in july 2012: Defining extreme space weather scenarios. *Space Weather*, 11(10):585–591.

Bayer, F. M., Tondolo, C. M., and Müller, F. M. (2018). Beta regression control chart for monitoring fractions and proportions. *Computers & Industrial Engineering*, 119:416–426.

Benjamim, M. A., Rigby, R. A., and Stasinopoulos, M. (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association*, 98(461):214–223.

Bolduc, L. (2002). GIC observations and studies in the hydro-québec power system. *Journal of Atmospheric and Solar-Terrestrial Physics*, 64(16):1793–1802.

Bortot, P., Coles, S. G., and Tawn, J. A. (2000). The multivariate gaussian tail model: An application to oceanographic data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(1):31–049.

Boteler, D. H. (2021). Modeling geomagnetic interference on railway signaling track circuits. *Space Weather*, 19(1):e2020SW002609.

Boteler, D. H., Pirjola, R. J., and Nevanlinna, H. (1998). The effects of geomagnetic disturbances on electrical systems at the earth's surface. *Advances in Space Research*, 22(1):17–27.

Carrington, R. C. (1859). Description of a singular appearance seen in the sun on september 1, 1859. *Monthly Notices of the Royal Astronomical Society*, 20:13–15.

Chen, H. and Huang, C. (2014). The use of a CUSUM residual chart to monitor respiratory syndromic data. *IIE Transactions*, 46(8):790–797.

Cliver, E. W. (2006). The 1859 space weather event: then and now. *Advances in Space Research*, 38(2):119–129.

Cliver, E. W. and Svalgaard, L. (2004). The 1859 solar–terrestrial disturbance and the current limits of extreme space weather activity. *Solar physics*, 224(1):407–422.

Cribari-Neto, F. and Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, 34:1–24.

Crowder, S. V. (1989). Design of exponentially weighted moving average schemes. *Journal of Quality technology*, 21(3):155–162.

Czech, P., Chano, S., Huynh, H., and Dutil, A. (1992). The hydro-quebec system blackout of 13 march 1989: System response to geomagnetic disturbance. In *Proc. EPRI Conf. Geomagnetically Induced Currents*, pages 1–19.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244.

Eroshenko, E. A., Belov, A. V., Boteler, D., Gaidash, S. P., Lobkov, S. L., Pirjola, R., and Trichtchenko, L. (2010). Effects of strong geomagnetic storms on northern railways in russia. *Advances in Space Research*, 46(9):1102–1110.

Espinheira, P. L., Ferrari, S. L. P., and Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics*, 35(4):407–419.

Ewan, W. D. (1963). When and how to use CUSUM charts. *Technometrics*, 5(1):1–22.

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Fournier, B., Rupin, N., Bigerelle, M., Najjar, D., and Iost, A. (2006). Application of the generalized lambda distributions in a statistical process control methodology. *Journal of Process Control*, 16(10):1087–1098.

Gan, F. F. (1991). An optimal design of CUSUM quality control charts. *Journal of Quality Technology*, 23(4):279–286.

Gjerloev, J. W. (2009). A global ground-based magnetometer initiative. *Eos, Transactions American Geophysical Union*, 90(27):230–231.

Gjerloev, J. W. (2012). The superMAG data processing technique. *Journal of Geophysical Research: Space Physics*, 117(A9).

Graham, J. W. (2011). *Data Mining with Rattle and R: The art of excavating data for knowledge discovery*. Springer Science & Business Media.

Green, J. L. and Boardsen, S. (2006). Duration and extent of the great auroral storm of 1859. *Advances in Space Research*, 38(2):130–135.

Haq, A. (2017). New synthetic CUSUM and synthetic EWMA control charts for monitoring the process mean using auxiliary information. *Quality and Reliability Engineering International*, 33(7):1549–1565.

Hawkins, D. M. (1981). A CUSUM for a scale parameter. *Journal of Quality Technology*, 13(4):228–231.

Hodgson, R. (1859). On a curious appearance seen in the sun. *Monthly Notices of the Royal Astronomical Society*, 20:15–16.

Huser, R. and Wadsworth, J. L. (2019). Modeling spatial processes with unknown extremal dependence class. *Journal of the American Statistical Association*, 114(525):434–444.

Hwang, W.-Y. (2021). Deviance residual-based control charts for monitoring the beta-distributed processes. *Quality and Reliability Engineering International*.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. CRC Press, New York, 1 edition.

Jones, M. C. (2009). Kumaraswamy distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, 6(1):70–81.

Keef, C., Tawn, J. A., and Lamb, R. (2013). Estimating the probability of widespread flood events. *Environmetrics*, 24(1):13–21.

Kim, H. and Lee, S. (2021). On residual CUSUM statistic for PINAR(1) model in statistical design and diagnostic of control chart. *Communications in Statistics-Simulation and Computation*, 50(5):1290–1314.

Kimball, D. S. (1960). A study of the aurora of 1859.

Koons, H. C. (2001). Statistical analysis of extreme values in space science. *Journal of Geophysical Research: Space Physics*, 106(A6):10915–10921.

Kozyreva, O. V., Pilipenko, V. A., Belakhovsky, V. B., and Sakharov, Y. A. (2018). Ground geomagnetic field and GIC response to march 17, 2015, storm. *Earth, Planets and Space*, 70(1):1–13.

Lee Ho, L., Fernandes, F. H., and Bourguignon, M. P. (2019). Control charts to monitor rates and proportions. *Quality and Reliability Engineering International*, 35(1):74–83.

Lima-Filho, L. M. A. and Bayer, F. M. (2021). Kumaraswamy control chart for monitoring double bounded environmental data. *Communications in Statistics-Simulation and Computation*, 50(9):2513–2528.

Lima-Filho, L. M. A., Pereira, T. L., Souza, T. C., and Bayer, F. M. (2019). Inflated beta control chart for monitoring double bounded processes. *Computers & Industrial Engineering*, 136:265–276.

Lima-Filho, L. M. A., Pereira, T. L., Souza, T. C., and Bayer, F. M. (2020). Process monitoring using inflated beta regression control chart. *PLOS ONE*, 15(7):e0236756.

Lucas, J. M. (1976). The design and use of V-mask control schemes. *Journal of Quality Technology*, 8(1):1–12.

Lucas, J. M. and Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, 32(1):1–12.

Mayaud, P. (1973). A hundred year series of geomagnetic data, 1868-1967: indices aa, storm sudden commencements. *Iaga Bulletin*, 33:256.

Montgomery, D. C. (2009). *Introduction to Statistical Quality Control*. John Wiley & Sons.

Nelsen, R. B. (2007). *An Introduction to Copulas*. Springer Science & Business Media, New York, 2 edition.

Ngwira, C. M., Pulkkinen, A., Leila Mays, M., Kuznetsova, M. M., Galvin, A. B., Simunac, K., Baker, D. N., Li, X., Zheng, Y., and Glocer, A. (2013). Simulation of the 23 july 2012 extreme space weather event: What if this extremely rare cme was earth directed? *Space Weather*, 11(12):671–679.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

Page, E. S. (1961). Cumulative sum charts. *Technometrics*, 3(1):1–9.

Park, J. and Jun, C.-H. (2015). A new multivariate ewma control chart via multiple testing. *Journal of Process Control*, 26:51–55.

Paulino, S., Morais, M. C., and Knoth, S. (2016). An ARL-unbiased c-chart. *Quality and Reliability Engineering International*, 32(8):2847–2858.

Pereira, G. H. A. (2019). On quantile residuals in beta regression. *Communications in Statistics-Simulation and Computation*, 48(1):302–316.

Perry, M. B. and Wang, Z. (2022). A distribution-free joint monitoring scheme for location and scale using individual observations. *Journal of Quality Technology*, 54(2):144–161.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 54(3):507–554.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250.

Rogers, N. C., Wild, J. A., Eastoe, E. F., Gjerloev, J. W., and Thomson, A. W. P. (2020). A global climatological model of extreme geomagnetic field fluctuations. *Journal of Space Weather and Space Climate*, 10:5.

Rogers, N. C., Wild, J. A., Eastoe, E. F., and Huebert, J. (2021). Climatological statistics of extreme geomagnetic fluctuations with periods from 1 s to 60 min. *Space Weather*, 19(11):e2021SW002824.

Roy, B. and Paul, A. (2013). Impact of space weather events on satellite-based navigation. *Space Weather*, 11(12):680–686.

Sant'Anna, A. M. O. and ten Caten, C. S. (2012). Beta control charts for monitoring fraction data. *Expert systems with applications*, 39(11):10236–10243.

Sanusi, R. A., Abbas, N., and Riaz, M. (2018). On efficient cusum-type location control charts using auxiliary information. *Quality Technology & Quantitative Management*, 15(1):87–105.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. Macmillan And Co Ltd, London.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer Science & Business Media.

Silrergleit, V. M. (1996). On the occurrence of geomagnetic storms with sudden commencements. *Journal of geomagnetism and geoelectricity*, 48(7):1011–1016.

Silrergleit, V. M. (1999). Forecast of the most geomagnetically disturbed days. *Earth, planets and space*, 51(1):19–22.

Simas, A. B., Barreto-Souza, W., and Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, 54(2):348–366.

Siscoe, G. L. (1976). On the statistics of the largest geomagnetic storms per solar cycle. *Journal of Geophysical Research*, 81(25):4782–4784.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1):54.

Sugiura, M. (1964). Annals of the international geophysical year. *Pergamon Press, Oxford*, 35:945.

Thaduri, A., Galar, D., and Kumar, U. (2020). Space weather climate impacts on railway infrastructure. *International Journal of System Assurance Engineering and Management*, 11(2):267–281.

Thomson, A. W. P., Dawson, E. B., and Reay, S. J. (2011). Quantifying extreme behavior in geomagnetic activity. *Space Weather*, 9(10).

Trichtchenko, L. and Boteler, D. H. (2001). Specification of geomagnetically induced electric fields and currents in pipelines. *Journal of Geophysical Research: Space Physics*, 106(A10):21039–21048.

Tsubouchi, K. and Omura, Y. (2007). Long-term occurrence probabilities of intense geomagnetic storm events. *Space Weather*, 5(12).

Wardell, D. G., Moskowitz, H., and Plante, R. D. (1992). Control charts in the presence of data correlation. *Management Science*, 38(8):1084–1105.

Weiß, C. H. and Testik, M. C. (2015). Residuals-based CUSUM charts for poisson INAR(1) processes. *Journal of Quality Technology*, 47(1):30–42.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wik, M., Pirjola, R., Lundstedt, H., Viljanen, A., Wintoft, P., and Pulkkinen, A. (2009). Space weather events in july 1982 and october 2003 and the effects of geomagnetically induced currents on swedish technical systems. In *Annales Geophysicae*, volume 27, pages 1775–1787. Copernicus GmbH.

Wintoft, P., Viljanen, A., and Wik, M. (2016). Extreme value analysis of the time derivative of the horizontal magnetic field and computed electric field. volume 34, pages 485–491. Copernicus GmbH.

Woodall, W. H. and Adams, B. M. (1993). The statistical design of CUSUM charts. *Quality Engineering*, 5(4):559–570.

Xue, L. and Qiu, P. (2021). A nonparametric CUSUM chart for monitoring multivariate serially correlated processes. *Journal of Quality Technology*, 53(4):396–409.