EVA 2023 Data Challenge:

Christian Rohrbeck[1], Emma Simpson[2], Jonathan Tawn[3]
[1] University of Bath; [2] University College London; [3] Lancaster University
January 13, 2023

1 Overview

It is a pleasure to be entrusted with setting the EVA 2023 Data Challenge. This year we have tried to develop problems which capture the variety of contexts we experience in the analysis of environmental extremes data. This involves both univariate and multivariate problems. The univariate extremes problems involve inference for extreme quantiles when faced with additional complications such as covariates; data missing at random; and the need to convert the inference into design levels which account for different losses from over- and under-design. We wanted to assess performance in multivariate extremes in a way that is independent of your marginal extremes abilities. Consequently the multivariate problems relate to data where the univariate marginal distributions are all known. Here the complexity comes from estimating probabilities of extreme events in different dimensions.

As we need to know the truth for assessing your performance we have searched the Earth and other worlds for the ideal data source. Our data come from a rather unique and special place, called Utopia. There is more to Utopia than meets the eye. It has some environmental variables that are very like on Earth, e.g., wind speed, but others that are entirely unique to Utopia, which are sufficiently difficult to spell in English, e.g., the key environmental variable of interest we will simply refer to as $Y_{i,t}$, being the variable Y at time point t and site i. For more details see Section 2.

Utopia has very special properties depending on where we are. Features you will notice in the data, but can take as given without proof are:

1. All observations of the $Y_{i,t}$ random variables are independent of each other over time (for each i) given the full set of covariates.

2. The environmental variables in Utopia either do not exist outside Utopia or are on completely different scales than we are familiar with from Earth, so no knowledge of a specific environmental process known on Earth is relevant for the analysis.

3. Although we have the site indicators of variables, we do not provide details of the locations of the sites. This is partly for data confidentiality; the Leader of the Utopia people, if there is one, wants to avoid releasing information that could be commercially important. Furthermore, the effect of topography and the nature of the environmental processes in Utopia mean that inter-site distance is irrelevant to describing inter-site dependence. So you must treat this as an entirely multivariate and not a spatial problem.

4. Utopia has two islands, called Coputopia and Utopula, both of which have additional special properties to those for Utopia. Specifically, the marginal distributions of the variable Y everywhere are identical over space and time, and these are known to be standard Gumbel distributed.

5. Coputopia is very lightly populated, with only three towns. Although $Y_{i,t}$ has identical margins over i and t on Coputopia, it is not known if any, some or all of the covariates we have available on Coputopia have any influence on the dependence structure of $Y_{i,t}$ across the three towns. However it is known that conditionally on any appropriate covariates, the $Y_{i,t}$ are independent over time.

6. It is known that observations of $Y_{i,t}$ for Utopula are independent and identically distributed over time and exhibit some dependence over i.

7. Despite all the equality over Utopia, the island of Utopula has two regional governments, U1 and U2, with regional government Uj able to provide defence against environmental variable $Y_{i,t}$ to a level s_j at each of its 25 towns for each of $j = 1, 2$, with $s_1 > s_2$. Specifically, let i_j denote site i in the region controlled by government Uj, (i.e., $i_j = 1, \ldots, 25$ and $j = 1, 2$) then for all i_j we have that $\Pr(Y_{i_j,t} > s_j) = \phi_j$, where the values of (ϕ_1, ϕ_2) are given later.

The overall challenge will consist of four sub-challenges, with performance scored separately on each sub-challenge and an overall score obtained from these individual scores. All sub-challenges count equal weight towards the overall score. Further details are provided in Section 4.

Rankings of the teams will be provided for each of the sub-challenges, so if you have time to undertake only a subset of the challenges you will get helpful feedback even though you are unlikely to be highly ranked overall.

2 Data and Covariates

2.1 Accessing data and covariates

All the files described below can be accessed by following the link provided on the EVA 2023 conference website.

2.2 Information for C1 and C2

You are given 70 years of data for Y for a single site in the capital city Amaurot. In Utopia a year is split into 12 months with 25 days each. Utopia has experienced a very stable climate over the observation period, and experts predict that this won't change in the next decades. Consequently, the distributions of the environmental variables can be assumed independent and identically distributed, except that of **Season** and **Atmosphere** which have a cyclical pattern.

Utopia's Meteorological Institute has provided you with a range of variables, which we collectively denote by $\mathbf{x}$, that vary over time and may explain variations in the distribution of Y:

V1,V2,V4,V5: These environmental variables describe atmospheric features that are unique to Utopia - the meteorologists in Utopia quickly gave up their attempt of explaining them to us.

Season: Amaurot experiences two seasons per year that are encoded as "S1" and "S2".

Wind direction: Provided in radians.

Wind speed: On a scale different to that used on Earth.

Atmosphere: A cyclical variable (recorded monthly) which is known to repeat exactly the same values every 70 years and describes fluctuations in the difference of atmospheric pressure between northern and southern Utopia.

Values for variables **V1**,...,**V4**, **Wind direction** and **Wind speed** are missing completely at random and given as `NA` in the file "Amaurot.csv".

The file containing the set of covariates $(\mathbf{x}_i : i = 1, \ldots, 100)$ for which conditional quantiles are required for **C1** are given in the file "AmaurotTestSet.csv".

2.3 Information for C3

When considering the island of Coputopia, we are interested in certain joint probabilities of the environmental variable across the three towns; these are denoted by Y_1, Y_2, Y_3.

The data recording facilities in Coputopia are less advanced than in Amaurot, so the majority of the covariate information given for the capital cannot be provided for Coputopia's three towns. You may, however, assume that the seasons are the same on Coputopia as on Amaurot and that the variable **Atmosphere** may still be relevant.

Data for Coputopia are provided in the file "Coputopia.csv". This dataset includes daily observations of Y at the three towns in Coputopia (labelled **Y1**,**Y2**,**Y3** in the dataset), as well as the **Season** and **Atmosphere** covariates. Observations are again available for a 70 year period.

2.4 Information for C4

There are 50 sites located on the island of Utopula. The two regional governments U1 and U2 collected data on the variable $Y_{i_j,t}$ ($i_j = 1, \ldots, 25, j = 1, 2$, and $t = 1 \ldots, 10000$) for all 50 sites and a total of 10,000 days. The data collected by the regional governments U1 and U2 can be found in the files "UtopulaU1.csv" and "UtopulaU2.csv", respectively. The observation for $Y_{i_j,t}$ of government Uj corresponds to the column with label **Y**i ($i = 1, \ldots, 25$) in each of the provided files.

3 Challenges

C1 The government of Utopia wants to improve the resilience of their capital city Amaurot to extreme weather events. Recent years have shown a particular vulnerability of Amaurot to events described by the variable Y which will be considered in this task.

The government want to know how extreme values of Y are affected by the covariates $\mathbf{X}$, described in Section 2.2. They have asked that you build a model for the distribution of $Y \mid \mathbf{X}$ and from this provide estimates. They want this information for 100 different covariate combinations ($\mathbf{x}_i : i = 1, \ldots, 100$) given in "AmaurotTestSet.csv": described in Section 2.2. Specifically, they want an

estimate of the level $q(\mathbf{x}_i)$ where

$$\Pr(Y < q(\mathbf{x}_i) \mid \mathbf{X} = \mathbf{x}_i) = 0.9999.$$

The government has previously received several statistical reports providing only point estimates of these conditional quantiles, which they have found to be unhelpful without any information on their accuracy. Hence, they are asking you for central 50% confidence intervals for the estimates of these extreme conditional quantiles.

Your results should be provided as a 100×3 matrix, where the first column contains your conditional quantile estimates, the second column the lower bound of the associated 50% central confidence interval and the third column the upper bound of this interval.

C2 Utopia's government want to develop a design standard to protect against a really rare event of Y, irrespective of the values of the covariates (i.e., they want this new design to be in place for hundreds of years so they aren't concerned about where in the Atmospheric cycle we are currently). In particular, they want to estimate the quantile q such that

$$\Pr(Y > q) = \frac{1}{300T},$$

where $T = 200$, so if the environmental process Y was i.i.d. this would be a one in T year event.

However, the government are really worried about the losses they could incur from over- or under-estimating q. Over-estimating would mean they have spent more to protect against Y than necessary. Under-estimating q would lead to more regular environmental damage to the city of Amaurot than is expected, thus resulting in large insurance claims. They can accept a small error in the estimate $\hat{q}$ of q but would be much less happy with an under-estimate than an over-estimate otherwise. They have framed this through the loss function

$$L(q, \hat{q}) = \begin{cases} 0.9\,(0.99q - \hat{q}) & \text{if } 0.99q > \hat{q} \\ 0 & \text{if } |q - \hat{q}| \leq 0.01q \\ 0.1\,(\hat{q} - 1.01q) & \text{if } 1.01q < \hat{q}. \end{cases}$$

Use this loss function and the data provided to give an estimated value $\hat{q}$ which you believe minimises the loss function for the government.

C3 The Utopian government want to know the probabilities of there being a combination of extreme and non-extreme events simultaneously in the three towns in Coputopia. As with **C2**, here the government are interested in these extreme events over a long time span - well in excess of the 70 year period of the Atmosphere cycle. Specifically, they want you to provide point estimates $\hat{p}_i$ $(i = 1, 2)$ of the following probabilities for the Y process, to be obtained using data in the file "Coputopia.csv" described in Section 2.3:

(i) $p_1 := \Pr(Y_1 > y, Y_2 > y, Y_3 > y)$;

(ii) $p_2 := \Pr(Y_1 > v, Y_2 > v, Y_3 < m)$;

where $y = 6$, $v = 7$ and m is the median of the marginal Gumbel distribution, i.e., $m = -\log(\log 2)$.

C4 The Utopian government want to know the probability of there being an exceeedance of the defence standards at all of the 50 towns in Utopula simultaneously, i.e.,

$$\Pr(Y_{i_j,t} > s_i : i_j = 1, \ldots, 25; j = 1, 2).$$

Specifically, they want you to provide an estimate of these probabilities for the Y process, to be obtained using the data in the files "UtopulaU1.csv" and "UtopulaU2.csv" described in Section 2.4. Provide a point estimate for this probability (which is known to be non-zero) in the cases where

(i) The current set-up is considered, with the design standards giving greater protection for sites in government area U1 than in area U2. Specifically, s_1 (and s_2) is the marginal level exceeded once in a year (in a month) on average, respectively. Thus, $\phi_2 = 12 \times \phi_1$ with $\phi_1 = 1/300$, and so $s_1 = 5.702113$ and $s_2 = 3.198534$. The associated probability and its estimate are denoted p_1 and $\hat{p}_1$.

(ii) A new set-up is considered, ensuring equality of U2 defences with those of U1 defences. That is, $\phi_1 = \phi_2 = 1/300$, i.e., $s_1 = s_2 = 5.702113$, both

corresponding to the marginal level exceeded once in a year. The associated probability and its estimate are denoted p_2 and $\hat{p}_2$.

4 Assessing Performance

4.1 Scoring methods for each sub-challenge

C1 Scoring uses the number, c_{50}, of the 100 submitted 50% confidence intervals that contain the true conditional quantile. The score is given by the value of $C_{50} = |c_{50} - 50|$, with smaller values of C_{50} being better. Only in the case where teams tie will we look at the widths of the confidence intervals. In that case, the team with the narrowest average interval will be ranked highest of these tied teams.

C2 Scoring here is based on the value of the loss $L(q, \hat{q})$ for your estimated quantile $\hat{q}$, where the truth is q. The smaller the value of $L(q, \hat{q})$ the better. Only when teams tie will we then look to use how close $\hat{q}$ is to q to complete the ranking.

C3 For the two required estimates $(\hat{p}_1, \hat{p}_2)$ of the true probabilities (p_1, p_2) we will use the metric

$$P_{12} = \sum_{i=1}^{2} \left| p_i \log\left(\frac{p_i}{\hat{p}_i}\right) + (1 - p_i) \log\left(\frac{1 - p_i}{1 - \hat{p}_i}\right) \right|.$$

Smaller values of P_{12} are better.

C4 Here the scoring metric used is identical to that in **C3**.

For each of the four sub-challenges the team with the ith lowest score will be given the ith ranking for the sub-challenge.

4.2 Converting rankings to points

The leaders of Utopia are big fans of the *Eurovision Song Contest* and have asked that the scoring system of this competition is also used in the Data Challenge for EVA 2023. So, to account for this, for each sub-challenge the following points will be awarded in relation to each of the ranked sub-challenges **C1**-**C4**.

Ranking	Points Awarded
1st	12
2nd	10
3rd	8
4th	7
5th	6
6th	5
7th	4
8th	3
9th	2
10th	1
$\geq$ 11th	0

4.3 Combining sub-challenge points into an overall ranking

The following are our rules for combining the points from the four individual sub-challenges into an overall performance score which will be used to decide on the winners of the EVA 2023 Data Challenge.

- Only the points from the individual sub-challenges are used in deciding on the overall performance.
- Equal weighting is to be given to all four sub-challenges.
- The overall points for a team will be the sum total over their points from the four sub-challenges.
- Teams will be ranked based on the highest overall points total.
- In the event of a tie, the overall ranking of these "equal" teams will be determined by the team highest score on their best (then second best, etc.) sub-challenge scores. If any teams are equal after that, the team with the smallest value for P_{12} in **C4** will be ranked highest of these tied teams.

5 Submission(s)

5.1 Registration

To register for the competition the competitors will simply need to send an email to the conference organisers via the address **eva23.board@unibocconi.it**.

It is important to register as early as possible as we will only send any updates addressing queries about the challenge to those registered.

5.2 Format

- Submissions should be provided as a single Rdata file and sent to **cr777@bath.ac.uk**.
- Please use your team name to name the Rdata file.
- The file should contain four R objects labelled **AnswerC1** for **C1**, **AnswerC2** for **C2**, etc. For **C1**, the answer has to be in the form of a 100×3 matrix as described in Section 3. For **C2** the answer is a single value while for each of **C3** and **C4** two values have to be submitted as a vector (and in the same order as introduced in the question text).
- Over-rounding of your results may have consequences on your ranking.

5.3 Preliminary submission

- We will allow one round of preliminary submissions. **The deadline for submissions is Sunday April 16, 2023.** Each submission will be scored and receive a ranking (e.g., 5th of 8 entries) separately for each sub-challenge. This information will be returned privately for each team by **April 30, 2023**.
- If a team has an entry for a sub-challenge which is considered "very far away" relative to the top teams on the challenge, we promise to tell them this in addition to reporting their ranking.

5.4 Final submission

The deadline for the final submission is 23:59 UTC on May 31, 2023.

6 Rules

1. There is no limit to the number of teams or team members. However, each participant can only be part of one team, not of several teams.
2. Only the final submission will be taken into account to rank the teams.
3. Submission of preliminary predictions is not mandatory, but highly encouraged.
4. Late submissions will not be considered.
5. Failure to comply with the above rules may result in disqualification.

7 Rewards

1. The rankings will be published on the EVA 2023 website and in the *Extremes* journal. Teams can choose not to appear in the published rankings. The winners will be officially announced during the EVA 2023 conference.
2. The best-ranked teams will be invited to present their work during an invited session at the EVA 2023 conference, organised by members of the Bocconi University from June 26 to June 30 2023.
3. After the EVA 2023 conference, all the teams will be invited to submit a paper describing their approach for publication in the journal *Extremes*. The submitted papers will undergo the usual peer-review process with the same quality standards and criteria of acceptance.