What can We Infer From Beyond The Data?
The Statistics Behind The Analysis of Risk Events In The Context Of Environmental Studies

S. Khuluse, S. Das, P. Debba, C. Elphinstone

Logistics and Quantitative Methods, CSIR (Built Environment),
Pretoria, South Africa.

African Digital Scholarship and Curation conference, 2009

Analysis of Risk

The risk and decision analysis framework provides an array of tools to aid the decision making process. Probabilistic Risk Analysis (PRA) involves

- Estimation of the probability of occurrence of a hazardous event.
- Determination of the distribution of the damage.
- Preparation of products that enable prediction of future risk events.

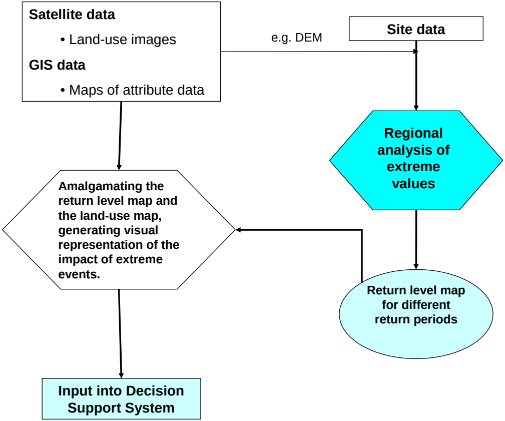
The principle concern is *low* probability - *high* consequence events.

Extreme Value Theory

The methodology provided by Extreme Value Theory (EVT) can be thought of as a tool to estimate the tail area of the distribution.

- Characterizes the probabilistic risk associated with an extreme event.
- In environmental studies, probabilistic risk often interpreted in terms of return intervals.

The role of Extreme Value Analysis in Probabilistic Risk Analysis


Satellite data
• Land-use images
GIS data
• Maps of attribute data
e.g. DEM
Site data
Regional analysis of extreme values
Amalgamating the return level map and the land-use map, generating visual representation of the impact of extreme events.
Return level map for different return periods
Input into Decision Support System

Challenges with the analysis of extreme environmental events

1. Lack of sufficient data
   - Decreasing number of data observation stations.
   - Data is often incomplete, due to instrument failure and damage.
   - Lack of consistency and uniformity in the format of archived data.
2. Issues with extrapolation
   - Often one has to extrapolate too far beyond the range of the data, which cause issues with the precision of the prediction.

Indicators of Fire Danger

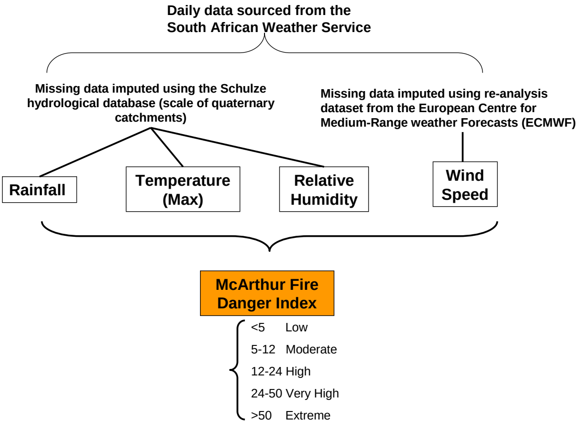
Fire Danger Indices (FDI)

- Indicate fire potential at a particular site and time.
- Are useful input into the National Fire Danger Rating System (NFDRS).

The McArthur FDI

- Is calculated from a combination of meteorological variables.
- Has a closed scale ranging from zero to 100.

Fire Danger Index Data



Daily data sourced from the South African Weather Service

Missing data imputed using the Schulze hydrological database (scale of quaternary catchments)

Missing data imputed using re-analysis dataset from the European Centre for Medium-Range weather Forecasts (ECMWF)

Rainfall

Temperature (Max)

Relative Humidity

Wind Speed

McArthur Fire Danger Index

<5 Low
5-12 Moderate
12-24 High
24-50 Very High
>50 Extreme

Study Objective

To apply the threshold exceedance approach within the Extreme Value Theory (EVT) framework, using the McArthur Fire Danger Index (FDI) time series from 4 sites in the Kruger National Park (KNP). Focus is on the change in parameter estimates, when the observations of the FDI series are assumed to be

- independent and identically distributed (i.i.d.)
- temporally dependent, but identically distributed
- temporally dependent and seasonal

Outline

Characterizing the behaviour of very high FDI values using the *threshold exceedance approach* to EVT.

1. Exploratory analysis
   - Description of the distribution of the FDI data per site
   - Visual inspection of the data for features of seasonality and long-term trend
   - Formal check for clustering of high values of the series
   - Selection of appropriate threshold
2. Modelling the threshold exceedances
   - Assumed the generalized Pareto distribution (GPD) to be suitable for excesses as sample size gets large.
   - Modelled using three cases, from i.i.d. assumption to assuming temporal dependence and seasonality.

Outline

1. Background
2. Methodology
   - Overview of the Theory of Extremes as it Applies to Threshold Exceedances
   - Study Area and Data
   - Exploratory Analysis
3. Results and Discussion
4. Conclusion

Classical Extreme Value Theory

Suppose the observed FDI values are $X_1, X_2, \ldots$. Assume the X_i to be a sequence of independent random variables with common distribution F. The cornerstone of EVT is that, without any knowledge about F, a model exists that describes the behaviour of the largest (or smallest) member of the sample

$$M_n = \max(X_1, X_2, \ldots, X_n)$$

Classical Extreme Value Theory

Conditional on the existence of $\{a_n\}$ and $\{b_n\} > 0$, the*Fisher-Tippett theorem* states that the rescaled sample optima converges in distribution to a variable having a distribution within one of 3 families: I-Gumbel, II-Frechet and III-Weibull. These can be unified into

$$P\left(\frac{M_n - a_n}{b_n} \leq x\right) \longrightarrow \exp\left\{-\left(1 + \xi\frac{x-\mu}{\sigma}\right)_+^{-\frac{1}{\xi}}\right\} \tag{1}$$

provided that $(y_+ = \max(y, 0))$, $-\infty < \mu < \infty$ and $\sigma > 0$.

- This is the Generalized Extreme value (GEV) family of distributions.

Generalized Pareto Distribution

Denote X_i by X. Suppose for large n, the Fisher-Tippett theorem holds. Then, for suitable threshold u,

$$P(X-u|X>u) \sim G(y;\sigma_u,\xi) = 1 - \left(1+\xi\frac{y}{\sigma_u}\right)_+^{\frac{-1}{\xi}} \tag{2}$$

defined on $\{y : y > 0\text{and} \left(1+\xi\frac{y}{\sigma_u}\right) > 0\}$, with

$$\sigma_u = \sigma + \xi(u-\mu) \tag{3}$$

- $G(\cdot)$ defines the Generalized Pareto distribution (GPD) family.

Modelling Dependent Sequence

What happens when the observations are *temporally dependent*?
Modifying the Fisher-Tippett theorem

- Assume existence of a condition that limits long-range dependence, such that
$$G_{\text{dep}}(z) = (G_{\text{indep}}(z))^{\theta}$$
- The extremal index $0 < \theta \leq 1$, measures the degree of clustering of extremes
- θ is loosely defined as the *limiting* mean cluster size
- GPD remains appropriate for threshold excesses, but clusters have to be filtered before fitting the model.
- Simplest method is *Runs declustering*

Non-stationary Extremes

Non-stationarity refers to the marginal distribution of the process not remaining the same as time changes. This could be the result of

- Seasonality
- Long-term trends

There is no general method to model non-stationary extremes.

- Extreme value models are used as templates- with model parameters expressed as statistical functions
- For the GPD, the basic formulation is

$$Y_t \sim GP\left(\hat{\sigma}(t), \hat{\xi}(t)\right) \tag{4}$$

- t is usually (but not restricted) to the time index

Outline

1. Background
2. Methodology
   - Overview of the Theory of Extremes as it Applies to Threshold Exceedances
   - Study Area and Data
   - Exploratory Analysis
3. Results and Discussion
4. Conclusion

The Kruger National Park

Kruger National Park (KNP), 4 sites: Shingwedzi, Letaba, Satara and Skukuza

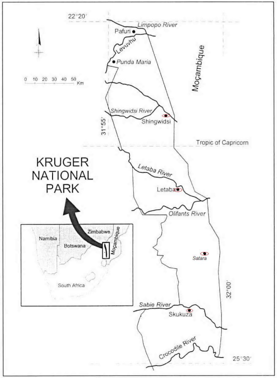


Figure: Map of the KNP- South Africa, with locations of: Shingwedzi, Letaba, Satara and Skukuza

McArthur Fire Danger Index Data

- Complete FDI series from 1960-2007 for 4 sites
- McArthur FDI rating system: <5 low, 5-12 moderate, 12-24 high, 24-50 very high, >50 extreme

Summary statistics	Shingwedzi	Letaba	Satara	Skukuza
Minimum	0.03	0.03	0	0.02
Median	10.42	10.12	9.1	8.83
Mean	11.13	10.91	9.99	9.9
95th Percentile	23.4	22.45	21.5	22.28
99th Percentile	29.61	28.13	27.3	28.42
Maximum	46.74	39.41	44.3	43.32

Table: Summarized description of the observed FDI values (1960-2007) for 4 sites

Outline

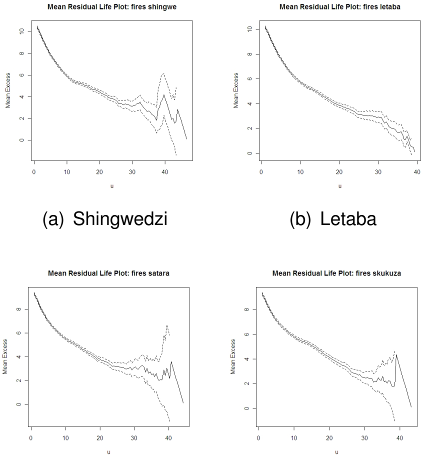
1 Background

2 Methodology
- Overview of the Theory of Extremes as it Applies to Threshold Exceedances
- Study Area and Data
- Exploratory Analysis

3 Results and Discussion

4 Conclusion

Threshold Selection: The mean excess plot



(a) Shingwedzi

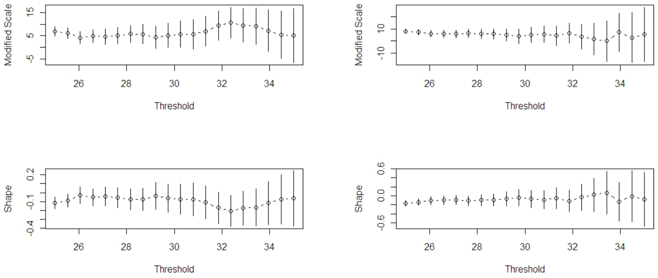
(b) Letaba

(c) Satara

(d) Skukuza

Figure: Mean excess plots of the daily FDI values for the four sites in the KNP

Threshold Selection: Threshold stability plot



(a) Parameter stability plot: Shingwedzi

(b) Parameter stability plot: Skukuza

Figure: Mean excess plots of the daily FDI values for the four sites in the KNP

Seasonality

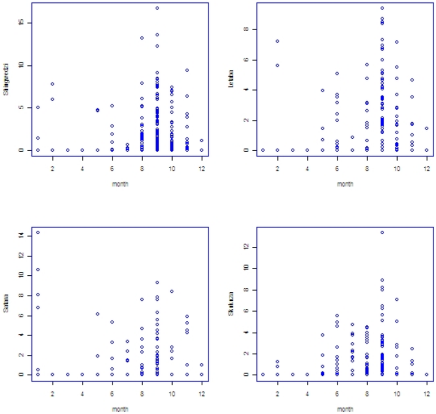
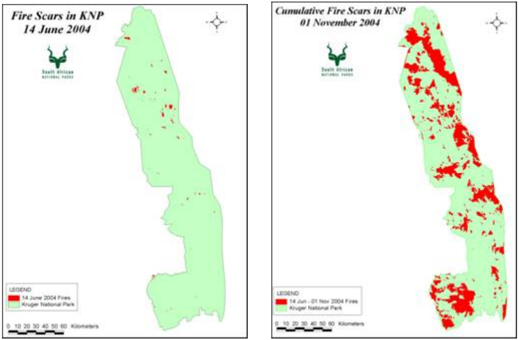


Figure: Monthly scatter plot of McArthur FDI values beyond 30 for each of the 4 sites

Seasonality


Fire Scars in KNP
14 June 2004
South African National Parks
LEGEND
🟥 14 June 2004 Fires
🟩 Kruger National Park
0 10 20 30 40 50 60 Kilometers
Cumulative Fire Scars in KNP
01 November 2004
South African National Parks
LEGEND
🟥 14 Jun - 01 Nov 2004 Fires
🟩 Kruger National Park
0 10 20 30 40 50 60 Kilometers

(a) Accumulated fire scars by June

(b) Accumulated fire scars by November

Figure: Remote sensing maps showing accumulated fire scars for 2004 in the Kruger National Park

Comparing estimates obtained when independence and identical distribution of the FDI values is assumed against the case where temporal dependence is assumed.

Sites		log-lik.	$\hat{\sigma}$(s.e.($\hat{\sigma}$))	$\hat{\xi}$(95% prof. c.i.)	10-yr r.l.(95% c.i.)
Shingwedzi	(a)	-337.94	3.23 (0.39)	-0.03 (-0.18; 0.18)	40.65 (39.07; 43.76)
	(b)	-189.64	3.61 (0.63)	-0.03 (-0.03; 0.31)	37.87 (36.43; 40.09)
Letaba	(a)	-194.73	4.06 (0.54)	-0.37 (-0.55; -0.16)	37.33 (36.66; 38.48)
	(b)	-164.17	4.30 (0.61)	-0.41 (-0.55; -0.18)	36.96 (36.23; 38.01)
Satara	(a)	-153.44	3.21 (0.55)	-0.07 (-0.26; 0.26)	38.01 (36.53; 40.81)
	(b)	-98.56	4.41 (0.91)	-0.19 (-0.59; 0.19)	36.34 (34.60; 38.22)
Skukuza	(a)	-194.59	2.72 (0.37)	-0.08 (-0.21; 0.15)	37.41 (36.24; 39.48)
	(b)	-137.39	3.62 (0.57)	-0.17 (-0.33; 0.09)	36.56 (35.38; 38.15)

Table: Summary of results from fitting the GPD to FDI excesses, where in: (a) Exceedances were assumed to be independent (b) The FDI series was assumed to be stationary

Fitting the GPD to the data for the July-December period. The aim is to compare these to the initial case where all the entire series was used.

Sites	log-lik.	$\hat{\sigma}$(s.e.($\hat{\sigma}$))	$\hat{\xi}$ (95% prof. c.i.)	10-yr r.l.(95% c.i.)
Shingwedzi	-181.69	3.09 (0.56)	0.06 (0.06; 0.41)	37.65 (36.13; 40.21)
Letaba	-150.38	4.37 (0.64)	-0.42 (-0.56; -0.18)	37.01 (36.67; 38.10)
Satara	-82.26	4.68 (1.08)	-0.43 (-0.55; 0.01)	35.51 (35.10; 36.81)
Skukuza	-134.98	3.33 (0.56)	-0.13 (-0.30; 0.17)	36.79 (35.53; 38.67)

Table: Summary of results from fitting the GPD to FDI excesses for the months July-December, under the assumption of stationarity of the series

- The merit in investigating temporal dependence and non-stationarity for environmental series, is obtaining more precision in return level estimation
- Incorrect estimates of the return level could lead to either wastage or lack of resources.
- In the study, only a description of temporal dependence was presented.
- In practice it may be important to to model the dependence structure between observations.
- Interesting
    - Data for the July-December period led similar parameter estimates as considering all exceedances over the year.
    - Incorporating sinusoidal functions for seasonality led to no improvement in model fit for the 2 regions North of KNP. In contrast, significant improvement was observed for regions in the south.

Is there spatial variation in very high McArthur FDI levels over the Kruger National Park?

- An extension of this research would be to describe the extent of spatial variation and it's impact on the estimates of the 10-year return level, to gain a more complete understanding of fire potential over the Kruger National Park.
- The challenge is that more data sites would have to be used for the spatial inference to be reliable.

- With regards to the data, collaborative efforts are underway to set-up observation networks, integrate environmental data from different sources and to archive the data in centralized repositories for easy access to end-users.
- With availability of infrastructure such as the Center for High Performance Computing in South Africa, head-way can be made for complex analysis using satellite data, which require more computing time.

Acknowledgements

- Sally Archibald for compiling and granting us access to the FDI data.
- The Global Risk Analysis SRP- Fire Risk Team for their insights.

For Further Reading

1. Beirlant J., Goegebeur Y., Segers J. and Teugels J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley.
2. Coles S.G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer.
3. Ecological Circuits (2009). South African Environmental Observation Network (SAEON) Editor: van Zwieten C., Issue 2, ISSN-2071-3290.
4. International Council for Science Regional Office for Africa (2007). *Science Plan on Natural and Human-Induced Hazards and Disasters in sub-Saharan Africa*.
5. IPCC- Working Group II (2007). Editors: Parry M.L., Canziani O.F., Palutikof J.P., Van der Linden P.J. and Hanson C.E. *Climate Change 2007: Impacts, Adaptation and Vulnerability*. Cambridge University Press.
6. Paté-Cornell M.E. and Dillon R.L. (2006). *The Respective Roles of Risk and Decision Analyses in Decision Support*. Decision Analysis, 3(4):220-232.

‘

Thank you!