Big Data: Progress in Automating Extreme Risk Analysis

Nadine Gissibl* Claudia Klüppelberg* Johanna Mager*

June 22, 2016

Abstract

Big data can be a curse and a blessing for the statistician. We report in this paper about some positive effect of big data: big data may open the way for more reliable risk analysis, simply because more extreme data are available. However, big data also require a fully automated analysis. We present here a method, which can easily be implemented and used for large numbers of statistical attributes. We apply this method to safety issues at airplane landings.

1 Introduction

There has been much debate about use and abuse of big data in recent years, Klaus Mainzer's excellent book [9] bears witness to this. Big data are often defined via the three Vs: *Volume* (data come in terabytes, are automatically recorded in files and tables, are often transactions), *Velocity* (they come in real time, are near time recorded or streamed) and *Variety* (they come structured or unstructured, or both).

On the negative side all of us are exposed to a supervision machinery unprecedented in history. Details of our lives are propagated to and within social networks without any respect or ethical conduct. Business concepts are based on masses of data which are either freely available or can be bought from illegitimate sources with the expressed goal of product placement.

On the positive side, however, scientists have never had so much information to study rare events like natural catastrophes (earthquakes, hurricanes, floods), technical risks like flight operation problems, nuclear power plant safety, or financial risks like the subprime crisis of 2007 or the ongoing European Sovereign-debt crisis.

Extreme value statistics has in the past always suffered from lack of data resulting in non-robust and inefficient estimators and prediction of rare events with often huge errors. As a consequence, extreme value statisticians performed each data analysis like a piece of art, supervising every step of the analysis with a catalog of provisions to avoid severely wrong and often disastrous conclusions. Other scientists took resort to standard statistical tools based on Gaussian models, which were often used for the whole data set, and extrapolated to extreme events, although extreme events can only be assessed statistically properly, when previous extreme events are analysed. Moreover, it has long been known that Gaussian models in almost all cases underestimate extreme events severly.

Now with huge data sets available and computers able to simulate and analyse masses of non-standard data originating from large networks or space-time measurements, risk assessment

*Center for Mathematical Sciences, Technische Universität München, 85748 Garching, Boltzmannstrasse 3, Germany, e-mail: {n.gissibl@tum.de , cklu@tum.de , johanna.mager@mytum.de}

by extreme value statistics is again on the rise and in greater demand than ever before. New interesting problems arise, since environmentalists, engineers and economists ask for user friendly extreme value methods based on a high degree of automation.

One of the attractions of modern extreme value methods is the embedded dimension reduction of the data. Since risk events can only be found in a small amount of data, even in a huge data set, extreme value statistics automatically leads to a reduction of the dimensionality and complexity of the data.

Whereas there exist computer packages for extreme value statistics implemented in MATLAB and `R`, which allow for graphical assessment of the extreme value analysis, it still requires the eye of the analyst to take the decision, which events should be included in the analysis. This task is usually named "threshold selection" and is one of the most critical points in extreme value statistics. With hundreds or even thousands of different variables measured over grids, networks or simply as a high-dimensional vector, such a method is no longer feasible.

This has been recognized by a number of extreme value statisticians, who have suggested methods for automatic threshold selection. We shall show one of these methods at work for assessing the runway overrun risk at airplane landings based on operational flight data.

Our paper is organised as follows. In Section 2 we introduce extreme risk models, where in Subsection 2.1 we present some dimension reduction methods based on extreme value concepts. Subsection 2.2 is then devoted to the Peaks-over-Threshold method, giving a statistical model for all observations exceeding a high threshold. This sets the stage for Section 3, where we present the automated threshold selection method, which opens up the way to risk assessment based on big data. In Section 4 we show our method at work in a technical risk analysis exemplified for the safety of airplane landings.

2 Extreme risk models

2.1 Dimension reduction

Assume that we have a large number of observations for a high-dimensional vector $X = (X_1, \ldots, X_d)$, which contributes to some risk in a system. In the simple case that all components measure the same kind of risk, for instance negative relative price changes of some financial assets, then, when we plot all such data together, only the most risky assets determine the high risk in the system. Note that throughout we think of risk as a positive quantity, and high risk corresponds to large values.

This knowledge allows us to perform an often substantial dimension reduction: We only need to consider the most risky assets, all the others contribute to the risk in the financial system only marginally. We can formulate this in mathematical terms. If extreme events correspond to large data values, like large financial losses for component i, then the behaviour of the distribution tail $\mathbb{P}(X_i > x)$ for large x matters only. For a common risk model this distribution tail is algebraically decreasing; i.e., $\mathbb{P}(X_i > x) = K_i x^{-\alpha}$ for large values of x. Now assume that among the components of X we have several with such algebraic tails, perhaps with different α-s, and some may have even faster (e.g. exponentially) decreasing tails. When we now assess the risk of the vector X by the sum of its components, and $\mathbb{P}(X_j > x)/\mathbb{P}(X_i > x) \to 0$ as $x \to \infty$, for all $j \neq i$, then (regardless of the dependence structure between the components of X), $\mathbb{P}(\sum_{j=1}^{d} X_j > x)/\mathbb{P}(X_i > x) \to 1$ as $x \to \infty$. In other words, the risk with the slowest decreasing tail (which indeed generates the highest risk values) completely determines the distribution tail

of the sum of extreme risks.

If there are several components (let's say components, or call them risk factors, with indices $i \in A \subset \{1, \dots, d\}$) with the same behaviour in the distribution tail (a statistical test could check this), then for every $i^* \in A$ we obtain $\mathbb{P}(X_i > x)/\mathbb{P}(X_{i^*} > x) \to C_i > 0$ as $x \to \infty$ for $i \in A$ and by Lemma A3.28 of [6] (for independent $X_1, \dots, X_d$),

$$\frac{\mathbb{P}(\sum_{j=1}^d X_j > x)}{\mathbb{P}(X_{i^*} > x)} \to \sum_{i \in A} C_i, \quad x \to \infty. \tag{2.1}$$

We observe a similar effect, when we have data classified with respect to some covariable Y, and want to calculate, for disjoint $B_1, \dots, B_n$ whose union covers all possible values for Y,

$$\mathbb{P}(X > x) = \sum_{i=1}^n \mathbb{P}(Y \in B_i)\mathbb{P}(X > x \mid Y \in B_i) =: \sum_{i=1}^n p_i \overline{F}_i(x), \tag{2.2}$$

where we have set $p_i := \mathbb{P}(Y \in B_i)$ and $\overline{F}_i(x) := \mathbb{P}(X > x \mid Y \in B_i)$. Then under the same condition as above, denoting $\overline{F}(x) := \mathbb{P}(X > x)$ for $x > 0$,

$$\frac{\overline{F}(x)}{\overline{F}_{i^*}(x)} \to \sum_{i \in A} p_i C_i, \quad x \to \infty. \tag{2.3}$$

Hence, for large values of x we can focus on all components with index in A and no others. This dimension reduction will also play a role in the aviation application of Section 4.

Example 2.1 (Sum of exponential and Pareto risks)**.** Assume that we have two different risks, one being exponentially distributed with distribution tail $P(X_1 > x) = e^{-x}$ and the other Pareto with distribution tail $P(X_2 > x) = (1+x)^{-2}$ for $x > 0$. Figure 1 shows a simulation of a random sample of X_1 (left) and of X_2 (middle). Comparing the high peaks from the middle and right hand plot we see immediately that *high* risks originate mainly from the Pareto risk X_2.

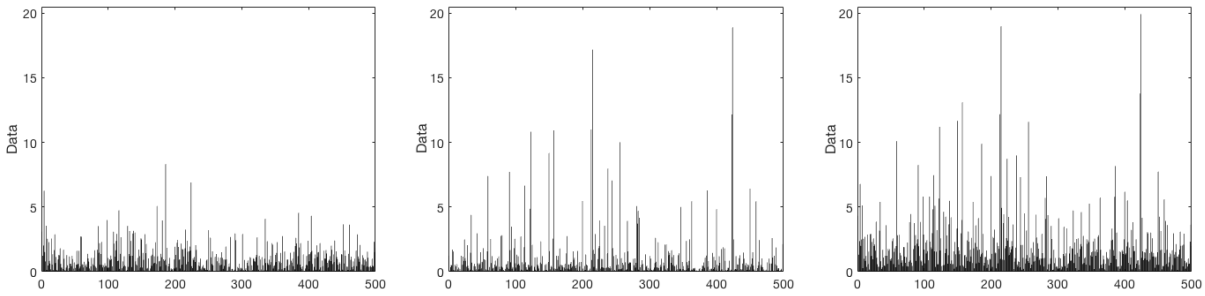

Figure 1: 500 data points simulated from the distributions given in Example 2.1: exponential distribution (left) and Pareto distribution (middle). The right hand plot shows the respective sums of both data.

2.2 Let the tails speak for themselves: the POT method

This recipe goes back to William H. DuMouchel in the 1970ies, propagated by Richard Smith in the 1980ies, emphasising that a wrong model for the extreme data gives wrong answers for the far out tail estimation, leading to wrong risk estimation. Consider for instance, Figure 1: if a Pareto distributed risk (as in the middle plot) is wrongly modelled by an exponential distribution (left plot), then the high risk is grossly underestimated.

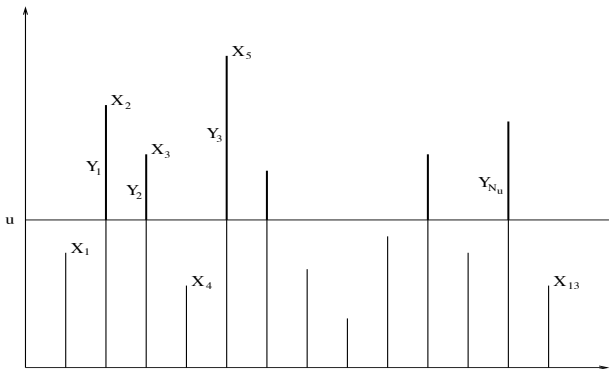

Figure 2: Data $X_1, \ldots, X_{13}$ with corresponding excesses $Y_1, \ldots, Y_{N_u}$.

As we have indicated above, and know from Figure 1, risk based on large values can be assessed by distribution tails $\mathbb{P}(X_i > x)$ for large x. Consequently, we have to estimate this for large values of x regardless of the behaviour of the data around their mean.

Since extreme risk is present only in large observations, we use exactly those, and the method is called *Peaks-over-Threshold* or *POT method*. We start with a high threshold u and note that an observation larger than $u+y$ for some $y > 0$ is only possible, if the observation is in the first place larger than u; this means one needs an *exceedance* of u. The observation itself has then necessarily an *excess over the threshold* u; cf. Figure 2.

It is now important for the POT method that for a random variable X with distribution function F and a large threshold u the following approximation holds under weak regularity conditions by the Pickands-Balkema-de Haan Theorem (cf. [6], Theorem 3.4.13 and the succeeding Remark 6),

$$\mathbb{P}(X > u + y \mid X > u) = \frac{\overline{F}(u+y)}{\overline{F}(u)} \approx \overline{H}_{\xi,\beta}(y) = \mathbb{P}(Y > y) \quad \text{for } y > 0, \tag{2.4}$$

where

$$H_{\xi,\beta}(x) = \left\{ \begin{array}{ll} 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-\frac{1}{\xi}}, & \text{if } \xi \in \mathbb{R}\backslash\{0\} \\ 1 - e^{-\frac{x}{\beta}}, & \text{if } \xi = 0 \end{array} \right\} \text{ for } \left\{ \begin{array}{ll} x > 0, & \text{if } \xi \geq 0, \\ 0 < x < -\beta/\xi, & \text{if } \xi < 0. \end{array} \right. \tag{2.5}$$

Moreover, the shape parameter ξ is independent of the threshold u, whereas the scale parameter $\beta = \beta(u)$ may change with u. The class $H_{\xi,\beta}$ for $\xi \in \mathbb{R}$ and $\beta > 0$ is called *generalized Pareto distribution (GPD)*. In a general context, (2.4) says that excesses over a high threshold, which we denote by Y, are (for large enough u) approximately generalized Pareto distributed. (This may seem at first sight suprising. However, such results are prominent in probability theory. The central limit theorem is the most known example: if independent random variables have the same distribution with finite variance, then the partial sums properly centered and scaled converge to a normal distribution. A similar result holds for partial maxima of random variables (cf. [6], Chapters 2 and 3).)

For sufficiently large u we approximate the excess distribution by a GPD $H_{\xi,\beta}$ as in (2.4) and obtain the tail approximation

$$\mathbb{P}(X > x) = \mathbb{P}(X > u)\mathbb{P}(X > x \mid X > u) \approx \overline{F}(u)\overline{H}_{\xi,\beta}(x-u), \quad x > u. \tag{2.6}$$

The choice of the threshold is delicate and creates a so-called bias-variance problem in the following sense: from a statistical point of view more data give better parameter estimates (with smaller variance) suggesting to take a low threshold; however, a too low threshold may approximate observations by a GPD, which cannot yet be approximated by a GPD (giving high bias).

For what follows we assume that we have independent and identically distributed (iid) observations $X_1, \dots, X_n$ all with distribution function F. Let us also assume for the moment that we know the threshold u such that observations above it can be well approximated by a GPD. In order to identify the observations exceeding u we order the observations $X_1, \dots, X_n$ as

$$\min\{X_1, \dots, X_n\} = X_{n:n} < \cdots < X_{1:n} = \max\{X_1, \dots, X_n\}. \tag{2.7}$$

Then observations larger than u correspond to the upper values $X_{k:n} < \cdots < X_{2:n} < X_{1:n}$ for some $k \in \{1, \dots, n\}$, in particular, $X_{k+1:n}$ for $k < n$ can serve as a threshold u. Moreover, the conditonal distribution of the exceedances $Y_1, \dots, Y_k$ given $X_{k+1:n} = u$ is approximated by the distribution function $H_{\xi,\beta}$ as in (2.5).

Since for all parameters ξ, β the GPDs are continuous and increasing functions, the integral transform yields for the ordered exceedances

$$H_{\xi,\beta}(0) = H_{\xi,\beta}(X_{k+1:n} - u) < H_{\xi,\beta}(Y_{k:n}) < H_{\xi,\beta}(Y_{k-1:n}) < \cdots < H_{\xi,\beta}(Y_{1:n}),$$

which are in distribution equal to ordered uniform random variables: $0 < U_{k:k} < U_{k-1:k} < \cdots < U_{1:k} < 1$.

This relation allows us to assess the goodness of fit of the upper order statistics by a GPD. To this end, note that, if the model is approximately correct, then the two random variables $H_{\xi,\beta}(Y_{i:n})$ and $U_{i:k}$ have the same distribution (denoted by $\stackrel{\mathrm{d}}{=}$) resulting in

$$H_{\xi,\beta}(Y_{i:n}) \;=\; 1 - \Big(1 + \xi\frac{X_{i:n} - u}{\beta}\Big)^{-\frac{1}{\xi}} \stackrel{\mathrm{d}}{=} U_{i:k} \quad \text{for } i = 1, \dots, k, \tag{2.8}$$

where we interpret the case of $\xi = 0$ as the limit for $\xi \to 0$ giving the exponential distribution function with mean β. If we replace the uniform order statistics by their expectations, i.e. $U_{i:k}$ by $\mathbb{E}[U_{i:k}] = 1 - \frac{i}{k+1}$ for $i = 1, \dots, k$, and the parameters ξ and β by their maximum likelihood estimates (MLEs) $\hat{\xi}$ and $\hat{\beta}$ based on the upper k order statistics of the observations, we find from (2.8),

$$Y_{i:n} := X_{i:n} - u \approx \frac{\hat{\beta}}{\hat{\xi}}\Big(\big(\frac{i}{k+1}\big)^{-\hat{\xi}} - 1\Big) =: \hat{Y}_{i:n} \quad \text{for } i = 1, \dots, k. \tag{2.9}$$

It is well-known that the asymptotic normality of the MLEs requires $\xi > -0.5$ (cf. [11]). So whenever we use the asymptotic normality of the MLEs we have to assume this.

3 Threshold selection

It is intrinsic to the problem that statistical estimation methods of tails and quantiles are as a rule very sensitive to substantial changes in the extreme data, since every data point is relevant among the not too many data points.

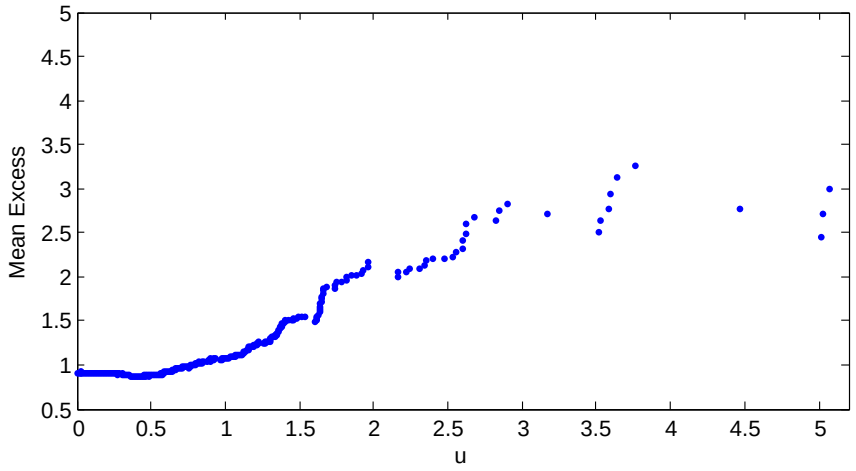

Figure 3: Simulated random numbers of independent random variables from the distribution function $\mathbb{P}(X \leq x) = \mathbf{1}(x \leq \Phi^{-1}(1-p))\Phi(x) + \mathbf{1}(x > \Phi^{-1}(1-p))(1-p+pH_{\xi,\beta}(x-\Phi^{-1}(1-p)))$ for $p = 0.1$, $\xi = 0.5, \beta = u\xi = 0.64$, where Φ denotes the standard normal distribution. Depicted is the mean excess plot. The true threshold (known by choice of the simulation method) is $u = z_{0.9} = 1.2816$ with $z_{0.9}$ being the 90% standard normal quantile.

When only one or very few risk types have to be considered, then a useful graphical tool for threshold selection is based on the *mean excess function*, given for a positive random variable X by

$$e(u) = \mathbb{E}[X - u \mid X > u], \quad u > 0.$$

For X with distribution function $H_{\xi,\beta}$ such that $\xi < 1$ (only then $\mathbb{E}[X]$ exists) and $\beta > 0$, the mean excess function is given by

$$e(u) = \frac{\beta + \xi u}{1 - \xi}, \quad \beta + \xi u > 0.$$

The function is linear in u with slope $\frac{\xi}{1-\xi}$ and intercept $\frac{\beta}{1-\xi}$.

When we define an empirical version $e_n(u)$ of the mean excess function and search for linearity beyond a high threshold, then from this threshold on a GPD makes a good model. Again choosing $u = X_{k+1:n}$, the empirical mean excess function is given by

$$e_n(u) = \frac{1}{k}\sum_{i=1}^{k}(X_{i:n} - u).$$

A mean excess plot consists of the points

$$\{(X_{i:n}, e_n(X_{i:n})) : i = 1, \ldots, n\} \in \mathbb{R}^2.$$

The mean excess plot should show an approximately linear behavior, when the underlying data follow a GPD. To choose an appropriate threshold for the POT method, we search for the smallest u where the plot is approximately linear. This guarantees that we use as many data points as possible to obtain GPD parameter estimates with minimal variance. This plot gives a first impression, whether a GPD approximation for high threshold excesses of the data is reasonable. For a simulation example see Figure 3.

As a second step the parameters ξ and β are estimated, where we focus on MLE (different estimation methods can be found in [6], Chapter 6). A meanwhile classical approach is to plot

the estimates of the shape parameter $\xi = \xi(u)$ for different thresholds $u = X_{k+1:n}$ (equivalently, different k) and find the best threshold by eye inspection. Since ξ is independent of the threshold, we choose a threshold value (not too high to avoid high variance, and not too low to avoid a bias), where the estimates $\xi(u)$ are stable.

When we have, however, a large number of risk types, then an automatic reliable procedure for estimating an appropriate threshold is called for. Various methods have been suggested in the literature; cf. Beirlant et al. [2], Section 4.7. We extend a method proposed in [5] (cf. also [10]) for a Pareto distribution to the GPD.

The basic idea is to minimize the mean squared prediction error with respect to the threshold u; i.e. to minimize $\Gamma(k) := \Gamma(X_{k+1})$ given by

$$\Gamma(k) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\Big[\Big(\frac{\hat{Y}_{i:n} - \mathbb{E}[Y_{i:n}]}{\sigma_i}\Big)^2\Big] \tag{3.1}$$

with respect to k, where $\sigma_i^2 = \mathrm{Var}(Y_{i:n})$, and $Y_{i:n}$ and $\hat{Y}_{i:n}$ are as in (2.9). Since we cannot compute the moments $\mathbb{E}[Y_{i:n}]$ and $\mathrm{Var}(Y_{i:n})$ explicitly, we have to reformulate the problem. We reformulate $\Gamma(k)$ as follows:

$$\begin{aligned}\Gamma(k) &= \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\Big[\Big(\frac{Y_{i:n} - \hat{Y}_{i:n}}{\sigma_i}\Big)^2\Big] + \frac{2}{k}\sum_{i=1}^{k}\frac{\mathrm{Cov}(Y_{i:n}, \hat{Y}_{i:n})}{\sigma_i^2} - \frac{1}{k}\sum_{i=1}^{k}\frac{\mathrm{Var}(Y_{i:n})}{\sigma_i^2} \\ &= \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\Big[\Big(\frac{Y_{i:n} - \hat{Y}_{i:n}}{\sigma_i}\Big)^2\Big] + \frac{2}{k}\sum_{i=1}^{k}\frac{\mathrm{Cov}(Y_{i:n}, \hat{Y}_{i:n})}{\sigma_i^2} - 1. \end{aligned} \tag{3.2}$$

The idea now is to replace $\Gamma(k)$ using some empirical versions of the moments.

Since tail estimation can be very sensitive to slight changes in the observations, we use classical methods of Robust Statistics to find an expression for the prediction error; cf. [7] and references therein. For simplification we set $\theta = (\beta, \xi)$ and, instead of working with the ideal GPD H_θ, in Robust Statistics, for some given value $\varepsilon > 0$, we define suitable distributional neighbourhoods about this ideal model. Here we restrict the method to neighbourhoods consisting of all distributions $H_\varepsilon = (1-\varepsilon)H_\theta + \varepsilon\delta_x$, where δ_x is the Dirac measure in x; i.e., it puts mass 1 on the point x. The Influence Function (IF) of an estimator T at x specifies the infinitesimal influence of the individual observation on the estimator by

$$IF(x; T, H_\theta) = \lim_{\varepsilon \to 0} \frac{T(H_\varepsilon) - T(H_\theta)}{\varepsilon}.$$

When we estimate now the parameter $\theta = (\beta, \xi)$ via a functional T evaluated at the empirical distribution, this robust approach allows us to specify the infinitesimal influence of the individual observations on the estimator (cf. (2.8)). This idea is now extended to the minimization of $\Gamma(\cdot)$ in (3.2) by estimating σ_i^2 and $\mathrm{Cov}(X_{i:n}, \hat{X}_{i:n})$ by robust methods based on different threshold values u corresponding to k. This means that for large k and $u = X_{k+1}$,

$$\sigma_i^2 = \mathrm{Var}(Y_{i:n}) \approx \frac{1}{k}\int_0^\infty IF(x; Y_{i:n}, H_\theta)^2 H_\theta(dx), \tag{3.3}$$

$$\mathrm{Cov}(Y_{i:n}, \hat{Y}_{i:n}) \approx \frac{1}{k}\int_0^\infty IF(x; Y_{i:n}, H_\theta) IF(x; \hat{Y}_{i:n}, H_\theta) H_\theta(dx). \tag{3.4}$$

In our model, for $\xi \neq 0$ it is possible to compute the influence functions $IF(x; Y_{i:n}, H_\theta)$ and $IF(x; \hat{Y}_{i:n}, H_\theta)$, where $Y_{i:n}$ and $\hat{Y}_{i:n}$ are defined in (2.9). Influence functions corresponding to $\xi = 0$ can be obtained as limits for $\xi \to 0$.

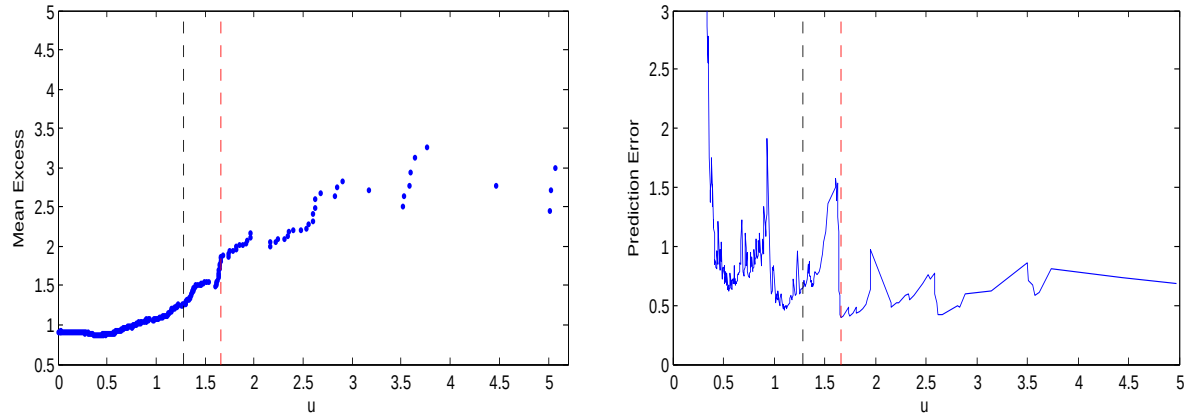

Figure 4: Simulated random numbers of independent random variables from the distribution function $\mathbb{P}(X \leq x) = \mathbf{1}(x \leq \Phi^{-1}(1-p))\Phi(x) + \mathbf{1}(x > \Phi^{-1}(1-p))(1-p+pH_{\xi,\beta}(x-\Phi^{-1}(1-p)))$ for $p = 0.1$, $\xi = 0.5, \beta = u\xi = 0.64$, where Φ denotes the standard normal distribution. The left plot shows the mean excess plot as in Figure 3 and the right plot depicts $\hat{\Gamma}$. The estimated threshold (red dashed line) is close to the true one (black dashed line, known by choice of the simulation method), which is $u = z_{0.9} = 1.2816$ with $z_{0.9}$ being the 90% standard normal quantile.

Empirical versions of the expressions of (3.3) and (3.4) can be obtained by introducing the MLE $\hat{\theta} = (\hat{\xi}, \hat{\beta})$ for $\theta = (\xi, \beta)$ and the estimated order statistics using $H_{\hat{\theta}}(Y_{i:n}) = U_{i:k}$ (cf. (2.8)). Then we minimize the empirical prediction error $\Gamma(\cdot)$, which takes the form

$$\hat{\Gamma}(k) = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{Y_{i:n} - \hat{Y}_{i:n}}{\hat{\sigma}_i}\right)^2 + \frac{2}{k}\sum_{i=1}^{k}\frac{\widehat{\mathrm{Cov}}(Y_{i:n}, \hat{Y}_{i:n})}{\hat{\sigma}_i^2} - 1. \tag{3.5}$$

This means we determine

$$k_{\mathrm{opt}} = \mathrm{argmin}_{k\in\{1,\ldots,n\}}\hat{\Gamma}(k). \tag{3.6}$$

This procedure uses the MLEs $\hat{\xi}, \hat{\beta}$ as input parameters and their asymptotic normality; hence, we require that $\xi > -0.5$ as indicated in Section 2.2.

So far we have introduced statistical risk models, which are appropriate to model risk variables by distribution tails, giving the probability that some specific risk exceeds a certain threshold. For the estimation of this probability we start with a semiparametric tail approximation, which holds for most probabilistic risk models above a high threshold and yields to reliable approximations of high risks. We also have presented an automatic threshold selection method, which estimates the high risk probabilities for large numbers of different risk quantities in an automated way. In the next section we will show this method at work in a technical risk analysis.

4 Technical risk analysis

We apply extreme value statistics to the safety of airplane landings. One specifically risky event is the so-called *runway overrun* (RO), which describes the fact that an airplane is unable to stop before the end of the runway. Such an event happened for instance on December 29, 2012 in Moscow; cf. `http://avherald.com/h?article=45b4b3cb` for details. A case study can be found in [1].

We set extreme value statistics to work in order to estimate the risk of such serious incidents. From conversation with experts we know that not every variable defining a RO is appropriate for estimating the occurrence probability of a RO. As an appropriate risk variable Max Butter [4] proposed the *maximum deceleration needed to stop* (DNS) within the runway. At each time during the landing, the hypothetical deceleration needed to stop before the end of the runway is computed, and then the maximum over the whole landing process is taken. A RO occurs when the DNS is higher than the *maximal deceleration physically possible* (DPP), i.e. in this case the DNS value is physically not possible.

Obviously, the landing process depends on various variables, and one of the most relevant risk factors is the *runway condition* (RWY Cond), where we distinguish between a "dry" and a "wet" runway. Engineers calculate the DPP for a dry runway as lying between 0.54g and 0.60g, depending on the specific runway, and for a wet runway between 0.33g and 0.35g, again the precise number depends on the specific runway.

Our goal is to estimate the occurrence probability of a RO for dry and wet runways

$$\mathbb{P}(\text{RO} \mid \text{RWY Cond} = \text{dry}) \quad \text{and} \quad \mathbb{P}(\text{RO} \mid \text{RWY Cond} = \text{wet}). \tag{4.1}$$

4.1 The data and risk factors

Our data set contains around 500.000 operational flight data, recorded mostly from February 2013 to February 2015. Table 4.1 gives a short overview of the available parameters of each flight. One basic parameter is the *aircraft type*. Examples of aircraft types are the A320, one of the famous narrow-body jets from the Airbus family, and the A388, the world's largest commercial passenger aircraft.

The *runway ID* gives the airport, the degrees the runway is pointing to the north pole, and a letter defining the position of the runway (R=right, L=left, C=center). For example, an A320 landing at MUC 08L heads 80 degrees northeast on the left runway of the Munich Airport. MUC 26R is the same physical runway, but approached from the other side; i.e., 260 degrees southwest, the right runway.

The *landing distance available* (LDA) is the usable length of the runway for landing. Sometimes we observe different LDAs for the same physical runway, caused for instance by an obstacle on the runway.

The *inflight landing distance* (ILD) is the calculated landing distance which can be accomplished by an average line pilot adhering to standard technique. The ILD is being published by the manufacturer of the aircraft and takes into account the aircraft weight, the runway geometry, weather and runway condition and other environmental aspects.

First we classify our data to make sure that we work with identically distributed samples, when estimating the distribution tail. We do this step by step based on the different attributes from Table 4.1.

We start with the attribute aircraft type. Figure 5 shows the boxplots of the DNS for landings on a dry runway grouped by different aircraft types. The medians (red lines) as well as the first and the third quartiles (the bottom and top of the blue boxes) are clearly distinct for each aircraft type, indicating that the distributions differ for different aircraft types.

Pairwise Kolmogorov-Smirnov tests (e.g. [3]) confirm that we have to differentiate between all 11 aircraft types. Similar classifications apply to all given attributes. However, we have to be aware that too detailed classification may lead to so few data within classes that we cannot estimate the GPD parameters in a statistically reliable way.

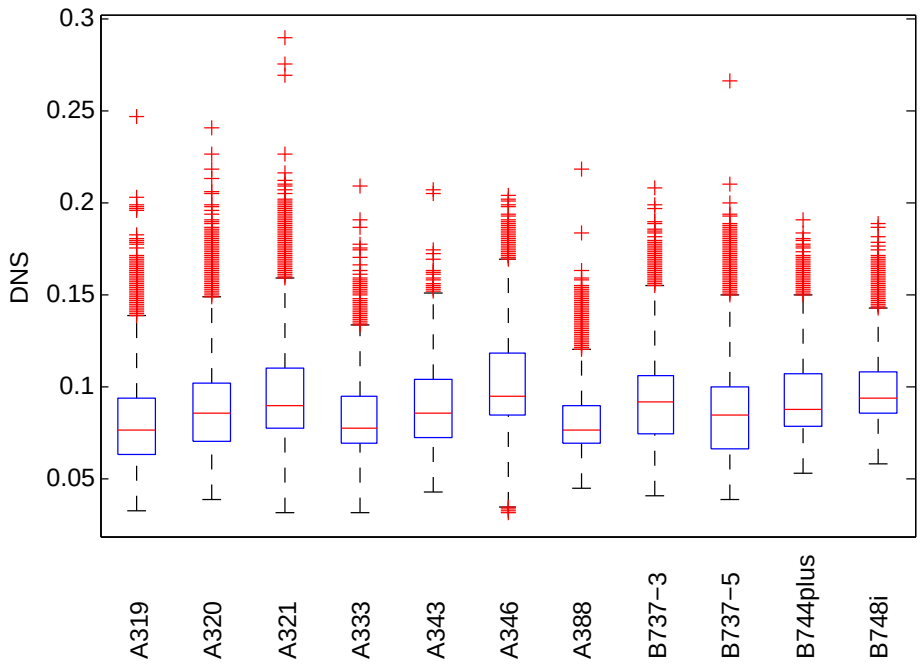

Figure 5: RWY Cond = dry: boxplots of DNS grouped by aircraft type.

Attribute	Description
Aircraft type	Aircraft type
Runway ID	Airport and runway identifier
DNS	Maximum deceleration needed to stop during roll-out measured in units of gravity ($1\text{g} = 9.81\text{m/s}^2$)
LDA	Landing distance available in meter
ILD	Inflight landing distance in meter: an estimation of the required landing distance
RWY Cond	Runway condition (dry, wet)

Table 4.1: Overview of the different parameters given for each flight.

Our data have to be classified by at least two of the most relevant attributes: The aircraft type as well as the runway ID have an influence on the distribution of the DNS. In this classification, however, the problem described above occurs and some classes would contain too few observations, in particular in case of wet runways. So finally we classify 11 aircraft types and 5 LDA categories (which summarize landings with similar LDAs).

4.2 Estimation

By definition, a runway overrun (RO) is an extreme event, and none was observed during the data records. So there is no way to estimate the occurrence probability of a RO with classical methods of statistics. It is, however, a typical problem to solve with extreme value statistics, which allows us to extrapolate risk assessment beyond the data range. We shall use the POT method as presented in Section 2 with the optimal threshold $u = X_{k+1:n}$ and k selected automatically within each data class by the optimization procedure from Section 3.

In the following we summarize the estimation algorithm and recall that the prediction error criterion leading to (3.6) is different for $\xi \neq 0$ and $\xi = 0$. This causes certain difficulties, which our algorithm has to deal with.

Algorithm

For a sample $X_1, \ldots, X_n$ of iid random variables and its corresponding order statistics $X_{n:n} < \cdots < X_{k+1:n} < X_{k:n} < \cdots < X_{1:n}$ perform the following steps.

(1) **Find the MLEs**
Find the MLEs $\hat{\xi} = \hat{\xi}(k)$ and $\hat{\beta} = \hat{\beta}(k)$ for all $k = k(n) \in [\min(40, \lfloor 0.02n \rfloor), \lfloor 0.2n \rfloor]$. This means, the percentage of the number of large sample values k taken into account ranges from 2% to 20%. If the upper 2% of the sample contains less than 40 values, for stability reasons, we take 40 extreme data points as the smallest k. On the other hand, we do not want to go too close to the center of the data. Hence, we do not consider more than 20% of the largest sample values. This is also motivated by the fact that for nearly all our data sets the optimal value for k was less than 20% of the upper data values. The estimator $\hat{\xi}$ will always give values different from 0, although close to 0 will be possible, indicating that the true ξ may be zero. We come back to this point below. We have also to ensure that the MLEs are asymptotically normal, requiring that $\xi > -0.5$. This is, however, not a problem for our data, since many estimates result in small positive or very small negative estimates.

(2) **Estimate** $\Gamma(k)$
Estimate the prediction error (3.1) under the assumption that $\xi \neq 0$: compute $\hat{\Gamma}(k)$ by (3.5) (details are given in [8], Theorem 4.9):

$$\begin{aligned}
\hat{\Gamma}(k) = &\sum_{i=1}^{k} \frac{1}{\hat{\beta}^2} \Big(\frac{i}{k+1}\Big)^{2\hat{\xi}} \Big(\frac{k+1}{i} - 1\Big)^{-1} \Big(X_{i:n} - X_{k+1:n} - \frac{\hat{\beta}}{\hat{\xi}} \Big(\Big(\frac{i}{k+1}\Big)^{-\hat{\xi}} - 1\Big)\Big)^2 \\
&+ \frac{2}{k} \sum_{i=1}^{k} \Big(\frac{i}{k+1}\Big)^{2\hat{\xi}} \Big(\frac{k+1}{i} - 1\Big)^{-1} \frac{(1+\hat{\xi})^2(1+2\hat{\xi})}{\hat{\xi}^2} \Big(\frac{\big(\frac{i}{k+1}\big)^{-\hat{\xi}} - 1}{\hat{\xi}}\Big)^2 \\
&+ \frac{4}{k} \sum_{i=1}^{k} \Big(\frac{i}{k+1}\Big)^{\hat{\xi}} \Big(\frac{k+1}{i} - 1\Big)^{-1} \frac{(1+\hat{\xi})(1+2\hat{\xi})}{\hat{\xi}^2} \frac{\big(\frac{i}{k+1}\big)^{-\hat{\xi}} - 1}{\hat{\xi}} \log\Big(\frac{i}{k+1}\Big) \\
&+ \frac{2}{k} \sum_{i=1}^{k} \Big(\frac{k+1}{i} - 1\Big)^{-1} \frac{(1+\hat{\xi})^2}{\hat{\xi}^2} \log^2\Big(\frac{i}{k+1}\Big) - 1.
\end{aligned}$$

(3) **Find the optimal threshold**
Find the number of excesses k such that the prediction error criterion (3.5) is minimal:

$$k_{opt} = \arg\min_k \hat{\Gamma}(k) \quad \text{and set} \quad u_{opt} = X_{k_{opt}+1:n}.$$

Denote by $(\hat{\xi}, \hat{\beta})$ the MLE of the GPD $H_{\xi,\beta}$ based on $X_{k:n}, \ldots, X_{1:n}$ for $\xi \neq 0$ and $k = k_{opt}$.

(4) **Perform a likelihood ratio test for** $\xi = 0$
Choose the $k = k_{opt}$ largest values of the sample and estimate the parameter β of the GPD $H_{0,\beta}$ based on $X_{k:n}, \ldots, X_{1:n}$. Denote the MLE by $\tilde{\beta}$.

Perform a likelihood ratio test for $\mathrm{H}_0 : \xi = 0$ with level $\alpha \in (0,1)$ on $X_{k:n}, \ldots, X_{1:n}$: reject H_0 if the likelihood ratio statistic

$$D := 2\Big\{ -\Big(\frac{1}{\hat\xi} + 1\Big) \sum_{i=1}^{k} \log\Big(1 + \hat\xi \frac{X_{i:n} - u}{\hat\beta}\Big) + \sum_{i=1}^{k} \frac{X_{i:n} - u}{\tilde\beta} + k \log \frac{\tilde\beta}{\hat\beta} \Big\} > \chi^2_{1-\alpha},$$

where $\chi^2_{1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2 distribution with one degree of freedom. We choose the level $\alpha = 0.05$, which gives $\chi^2_{0.95} = 3.84$.

We consider the two possible outcomes of the test:

- If H_0 is rejected: Then $\hat\xi, \hat\beta$ are the MLEs obtained in step (1), based on the fitted GPD of the k_{opt} largest values and $u_{opt} = X_{k_{opt}+1:n}$ as in steps (2) and (3) above.
- If H_0 is not rejected: Then we assume that $\xi = 0$ (likelihood ratio tests for k close to k_{opt} also do not reject H_0 for all our data sets). Finally, now for $\xi = 0$, we estimate the prediction error (3.1) $\Gamma^0(k)$ by (3.5) (details are given in [8], Theorem 4.14):

$$\begin{aligned}\hat\Gamma^0(k) = &\sum_{i=1}^{k} \frac{1}{\tilde\beta^2} \Big(\frac{k+1}{i} - 1\Big)^{-1} \Big(X_{i:n} - X_{k+1:n} + \tilde\beta \log\Big(\frac{i}{k+1}\Big)\Big)^2 \\ &+ \frac{2}{k} \sum_{i=1}^{k} \Big(\frac{k+1}{i} - 1\Big)^{-1} \log^2\Big(\frac{i}{k+1}\Big) - 1.\end{aligned}$$

We find

$$k^0_{opt} = \arg\min_k \hat\Gamma^0(k) \quad \text{and set} \quad u_{opt} = X_{k^0_{opt}+1:n}.$$

Denote now by $\tilde\beta$ the MLE of the GPD $H_{0,\beta}$ based on $X_{k:n}, \ldots, X_{1:n}$ and $k = k^0_{opt}$.

(5) **Estimate the GPD for the excesses**
Recall (2.4) and set

$$\begin{aligned}
\hat\xi < 0 \qquad & \overline{H}_{\hat\xi,\hat\beta}(x) = \Big(1 + \hat\xi \frac{x - u_{opt}}{\hat\beta}\Big)^{-\frac{1}{\hat\xi}}, \quad x \in \Big(u_{opt}, u_{opt} - \frac{\hat\beta}{\hat\xi}\Big), \\
\hat\xi = 0 \qquad & \overline{H}_{0,\hat\beta}(x) = \exp\Big\{ -\frac{x - u_{opt}}{\tilde\beta} \Big\}, \quad x > u_{opt}, \\
\hat\xi > 0 \qquad & \overline{H}_{\hat\xi,\hat\beta}(x) = \Big(1 + \hat\xi \frac{x - u_{opt}}{\hat\beta}\Big)^{-\frac{1}{\hat\xi}}, \quad x > u_{opt}.
\end{aligned}$$

(6) **Estimate the excess probability for the risk variable X**
Recall (2.6), estimate $\mathbb{P}(X > u)$ by its empirical version and obtain

$$\widehat{\mathbb{P}(X > x)} = \widehat{\overline{F}}(x) = \frac{k_{opt}}{n} \overline{H}_{\hat\xi,\hat\beta}(x - u_{opt}) \quad \text{for} \quad x > u_{opt}.$$

This algorithm is applied to estimate (DNS is measured in units of gravity g):

$$\mathbb{P}(\text{RO} \mid \text{RWY Cond} = \text{dry}) = \mathbb{P}(\text{DNS} > 0.54) \text{ and } \mathbb{P}(\text{RO} \mid \text{RWY Cond} = \text{wet}) = \mathbb{P}(\text{DNS} > 0.33),$$

as in (4.1) for all different classes determined in Section 4.1. The precise probabilities can be found in [8]. As to be expected, some classes lead to higher probabilities than others, which allows for an assessment of the riskiness of the specific values of the corresponding attributes.

Moreover, there are classes with less than 300 observations, which leads to unreliable parameter estimates. There are two possibilities to deal with this problem: consider only such classes with at least 300 observations, or merge such classes, whose distributions are not too different. Since we want to use most of the data, we use the classification by aircraft type and landing distance available (LDA) as already mentioned in Section 4.1. It is possible to summarize the results of the different classes by the theorem of total probability (cf. (2.2)) resulting in

$$\mathbb{P}(\text{RO}|\text{RWY Cond}) = \sum_{i=1}^{n_{\text{T}}} \sum_{j=1}^{n_{\text{LDA}}} \mathbb{P}(\text{RO}|\text{T}_i \cap \text{LDA}_j \cap \text{RWY Cond})\mathbb{P}(\text{T}_i \cap \text{LDA}_j|\text{RWY Cond}),$$

where T_i stands for aircraft type i with $i = 1, \ldots, n_{\text{T}} = 11$ and LDA_j for LDA category j with $j = 1, \ldots, n_{\text{LDA}} = 5$. RWY Cond is again either dry or wet. This way we use most of the data and obtain the estimates

$$\mathbb{P}(\text{RO} \mid \text{RWY Cond} = \text{dry}) = 3.72 \times 10^{-7} \quad \text{and} \quad \mathbb{P}(\text{RO} \mid \text{RWY Cond} = \text{wet}) = 4.98 \times 10^{-6}.$$

When analysing these results more closely, we find that—for landings on a dry runway—the total probability is mainly driven by two classes for which relatively high probabilities have been estimated, such that (2.3) implicitly applies. These two classes involve the B744i-plus landings on runways with LDA between 3000m and 3500m and the B737-5 landings on runways with LDA between 3500m and 4000m.

4.3 Conclusion

We have applied extreme value statistics to assess the safety of airplane landings, focussing on the event of a runway overrun. Given the large number of data available we have developed an automated threshold selection to identify the relevant events, which we used for statistical estimation of the occurrence probability of a runway overrun. Although we had a large number of flight data available, we still had to merge certain data, which may not have been identically distributed. We have tried to keep track of the consequences of this simplification, but there is still room for improvement. Furthermore, we have used the maximum deceleration needed to stop during roll-out (DNS) as risk variable, which seems to be the relevant quantity to study in this context. This novel risk quantity is currently discussed among flight engineers. Further studies based on more detailed data will certainly help to clarify this notion. We also hope that this study helps to propagate the powerful statistical extreme value methods further in the engineering sciences.

Acknowledgements

We are grateful to Johan Segers for sharing his insight into different threshold selection methods with us. The first author acknowledges support from ISAM of TUM Graduate School at Technische Universität München. This research was funded by the German LUFO IV/4 project SaMSys – Safety Management System in Order to Improve Flight Safety.

References

[1] E.S. Ayra. Risk analysis of runway overrun excursions at landing: a case study. `http://www.agifors.org/award/submissions2013/EduardoAyra.pdf`, 2013.

[2] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes: Theory and Applications*. Wiley, New York, 2004.

[3] P.J. Bickel and K.A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Volume I*. Chapman & Hall/CRC, New York, 2 edition, 2001.

[4] M. Butter. DNS as risk variable for a runway overrun. Personal Communication, 2015.

[5] D.J. Dupuis and M-P. Victoria-Feser. A robust prediction error criterion for Pareto modelling of upper tails. *Canadian Journal of Statistics*, 34(4):639–658, 2006.

[6] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin, 1997.

[7] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics - The Approach Based on Influence Functions*. Wiley, New York, 1986.

[8] J. Mager. *Automatic Threshold Selection of the Peaks-Over-Threshold Method*. Master's thesis, Technische Universität München, 2015.

[9] K. Mainzer. *Die Berechnung der Welt*. C.H. Beck, München, 2014.

[10] P. Ruckdeschel and N. Horbenko. Robustness properties of estimators in generalized Pareto models. Bericht 182, Fraunhofer Institut für Techno- und Wirtschaftsmathematik, 2010.

[11] R.L. Smith. Estimating tails of probability distributions. *Annals of Statistics*, 15:1174–1207, 1987.