Spatio temporal analysis of extreme wind velocities for infrastructure desing. Case study Colombia

A Thesis
Presented to
The Division of Instituto for Geoinformatics - IFGI
University of Münster

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Geospatial Technologies

Alexys Herleym Rodríguez Avellaneda

Jan 2020

Approved for the Division
(Faculty of Geosciences)

Dr. Edzer Pebesma

Dr. Juan C. Reyes\Dr. Sara Ribero

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

Introduction

Placeholder

Chapter 1

Data

Placeholder

1.1 IDEAM

1.2 ISD

1.3 ERA5

1.4 Data Download and Organization

1.5 Data Standarzation

Chapter 2

Theoretical Framework

2.1 Probability Concepts

Poisson process is an stochastic method that relies in the concepts of probability distributions. The main functions related to probability for extreme value analysis will be described below.

2.1.1 Probability Density Function - *pdf*

Pdf defines the probability that a continuos variable falls between two points, this is, in *pdf* the proability is related to the area below the curve (integral) between two points, as for continuos probability distributions the probability at a single point is zero. The term density is directly related to the probability of a portion of the curve, if the density function has high values the probability will be greater in comparison with the same portion of curve for low values.

$$\int_a^b f(x)dx = Pr[a \leq X \leq b]$$

Equation (2.1) is the Gumbel *pdf*.

$$\mathrm{f(x)} = \frac{1}{\beta} \exp\left\{-\frac{\mathrm{x}-\mu}{\beta}\right\} \exp\left\{-\exp\left\{-\left(\frac{\mathrm{x}-\mu}{\beta}\right)\right\}\right\}, \quad -\infty < \mathrm{x} < \infty \tag{2.1}$$

where $\exp\{.\} \mapsto \mathrm{e}^{\{.\}}$, β is the scale parameter, and μ is the location parameter. Location (μ) has the effect to shift the *pdf* to left or right along 'x' axis, thus, if location value is changed the effect is a movement of *pdf* to the left (small value for location), or to the right (big value for location). Scale has the effect to stretch ($\beta > 1$) of compress ($0 < \beta < 1$) the *pdf*, if scale parameter is close to zero the pdf approaches a spike.

Figure 2.1 shows *pdf* with location (μ) = 100 and scale (β) = 40, using equation (2.1).

```
location = 100
scale = 40
.x <- seq(0, 300, length.out=1000)
pdfG <- function(x) {
  1/location *exp(-(x-location)/scale)*exp(-exp(-(x-location)/scale))
  }
.y = pdfG(.x)
plot(.x, .y, col="green", lty=4,
     xlab="Velocities Km/h", ylab="Density Function - Gumbel Distribution",
     main=paste("Gumbel - Density Function Gumbel Distribution\n", "Location=",
     round(location,2), " Scale=", round(scale,2)), type="l",
     cex.axis = 0.5, cex.lab= 0.6, cex.main=0.7, cex.sub=0.6)
```

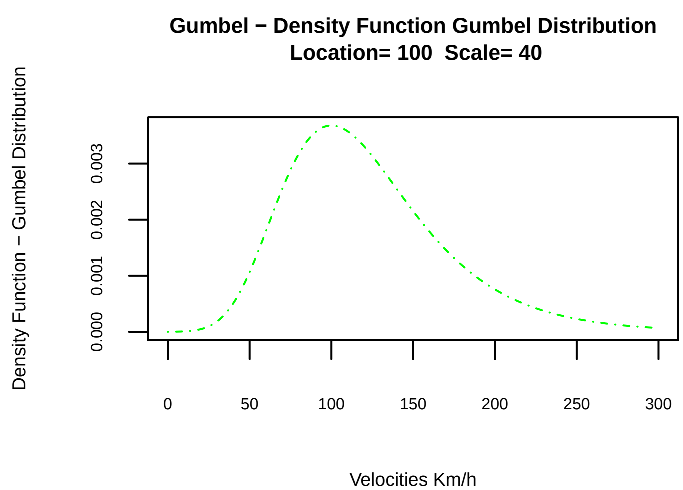

Figure 2.1: Gumbel pdf

Figure 2.2 shows *pdf* with location $(\mu) = 100$ and scale $(\beta) = 40$, using function `dgumbel` of the package `RcmdrMisc`

```
location = 100
scale = 40
.x <- seq(0, 300, length.out=1000)
dfG = dgumbel(.x, location=location, scale=scale)
plot(.x, dfG, col="red", lty=4,
     xlab="Velocities Km/h", ylab="Density Function - Gumbel Distribution",
     main=paste("Gumbel - Density Function Gumbel Distribution\n", "Location=",
     round(location,2), " Scale=", round(scale,2)), type="l",
     cex.axis = 0.5, cex.lab= 0.6, cex.main=0.7, cex.sub=0.6)
```

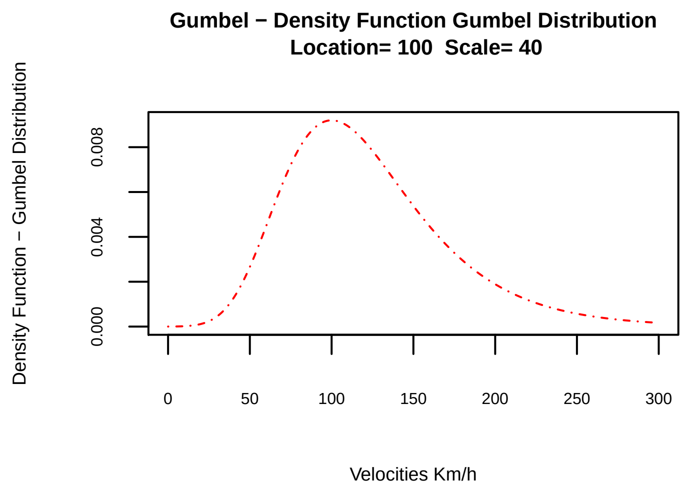

Figure 2.2: Gumbel pdf - dgumbel function

2.1.2 Cumulative Distribution Funtcion - *cdf*

Cdf is the probability of taking a value less than or equal to x. That is

$$F(x) = Pr[X < x] = \alpha$$

For a continuous variable, *cdf* can be expressed as the integral of its *pdf*.

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

Equation (2.2) is the Gumbel *cdf*.

$$\mathrm{F(x)} = \exp\left\{-\exp\left[-\left(\frac{\mathrm{x}-\mu}{\beta}\right)\right]\right\}, \quad -\infty < \mathrm{x} < \infty \tag{2.2}$$

Figure 2.3 shows Gumbel *cdf* with location (μ) = 100 and scale (β) = 40, using equation (2.2). As previously done with *pdf*, similar result can be achieved using function `pgumbel` of package `RcmdrMisc`.

```
location = 100
scale = 40
.x <- seq(0, 300, length.out=1000)
cdfG <- function(x) {
  exp(-exp(-(x-location)/scale))
  }
.y = cdfG(.x)
plot(.x, .y, col="green", lty=4,
     xlab="Velocities Km/h", ylab="Probability",
     main=paste("Gumbel - Cumulative Distribution Function\n", "Location=",
     round(location,2), " Scale=", round(scale,2)), type="l",
     cex.axis = 0.5, cex.lab= 0.6, cex.main=0.7, cex.sub=0.6)
```

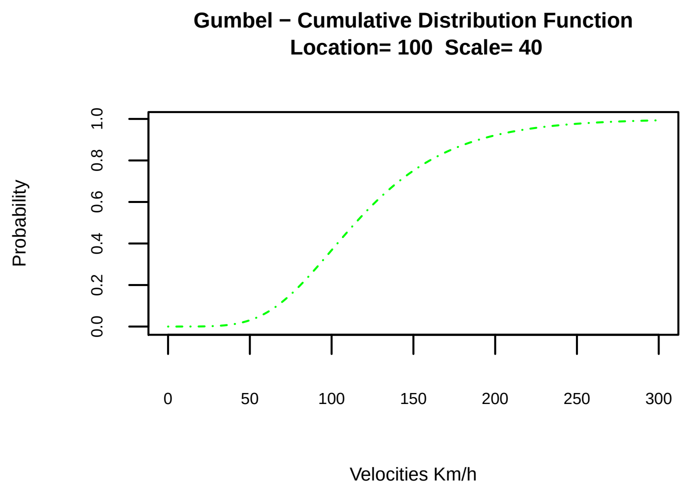

Figure 2.3: Gumbel cdf

2.1.3 Percent Point Function - *ppf*

Ppf is the inverse of *cdf*, also called the *quantile* function. This is, from a specific probability get the corresponding value x of the variable.

$$x = G(\alpha) = G(F(x))$$

Equation (2.3) is the Gumbel *ppf*.

$$\mathrm{G}(\alpha) = \mu - \beta \ln(-\ln(\alpha)) \quad 0 < \alpha < 1 \tag{2.3}$$

Figure 2.4 shows Gumbel *ppf*, using equation (2.3). Similar result can be achieved using function `qgumbel` of package `RcmdrMisc`.

```
location = 100
scale = 40
.x <- seq(0, 1, length.out=1000)
ppfG <- function(x) {
  location - (scale*log(-log(x)))
  }
.y = ppfG(.x)
plot(.x, .y, col="green", lty=4,
     ylab="Velocities Km/h", xlab="Probability",
     main=paste("Gumbel - Percent Point Function\n", "Location=",
     round(location,2), " Scale=", round(scale,2)), type="l",
     cex.axis = 0.5, cex.lab= 0.6, cex.main=0.7, cex.sub=0.6)
```

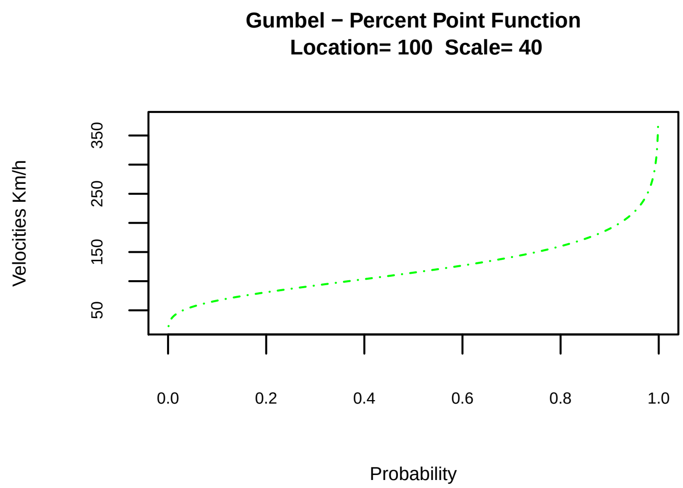

Figure 2.4: Gumbel cdf

2.1.4 Hazard Function - *hf*

Using $S(x) = 1 - F(x)$ as survival function -*sf*, the probability that a variable takes a value greather than x $S(x) = Pr[X > x] = 1 - F(x)$, the *hf* is the ratio between *pdf* and *sf*.

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)}$$

Equation (2.4) is the Gumbel *ppf*.

$$\mathrm{h(x)} = \frac{1}{\beta} \frac{\exp(-(\mathrm{x} - \mu)/\beta)}{\exp(\exp(-(\mathrm{x} - \mu)/\beta)) - 1} \tag{2.4}$$

Figure 2.5 shows Gumbel *hf*, using equation (2.4).

```
location = 100
scale = 40
.x <- seq(0, 3000, length.out=1000)
hfG <- function(x) {
  (1/scale)*(exp(-(x-location)/scale))/(exp(exp(-(x-location)/scale))-1)
  }
.y = hfG(.x)
plot(.x, .y, col="green", lty=4,
     xlab="Velocities Km/h", ylab="Hazard",
     main=paste("Gumbel - Hazard Function\n", "Location=",
     round(location,2), " Scale=", round(scale,2)), type="l",
     cex.axis = 0.5, cex.lab= 0.6, cex.main=0.7, cex.sub=0.6)
```

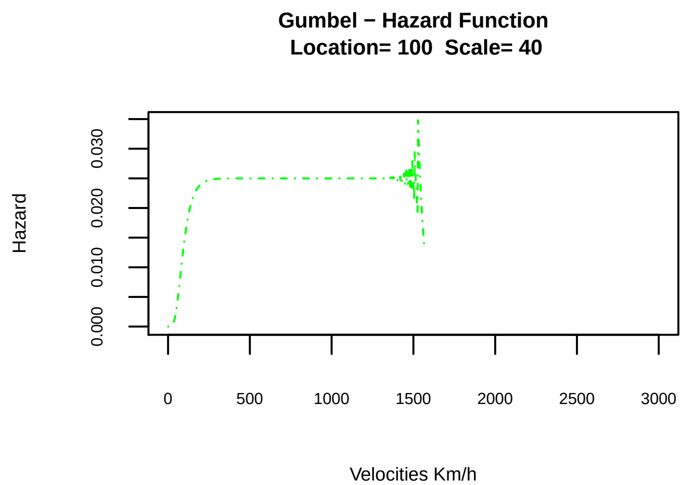

Figure 2.5: Gumbel cdf

```
#library(reliaR)
#plot(.x, hgumbel(.x, mu=location, sigma=scale))
#plot(.x, hra.gumbel(.x, mu=location, sigma=scale))
```

2.2 Introductory concepts for statistical analysis of extreme values

In order to approach the extreme value analysis, some statistical concepts are needed to understand the theoretical framework behind this knowledge area. In this section will be introduced the concepts annual excedance probability, mean recurrence interval - MRI, exposure time, and compound probability for any given exposure time and MRI.

As an hypotetical example, a simulated database of extreme wind speed will be used. This database is supposed to have 10.000 years of simulated wind speeds.

2.2.1 Annual Excedance Probability - Pe

Using the previously described database, a question arises to calculate the probability to exceed the highest probable loss due to the simulated winds. It is possible to conclude that there is only one event grather or equal (in this case equal) to the higest probable causing loss in 10.000 years, and it is the *highest wind*. If we sort the database by wind magnitude in descending order (small winds last), the question is solved calculating the annual excedance probability *Pe* with next formula

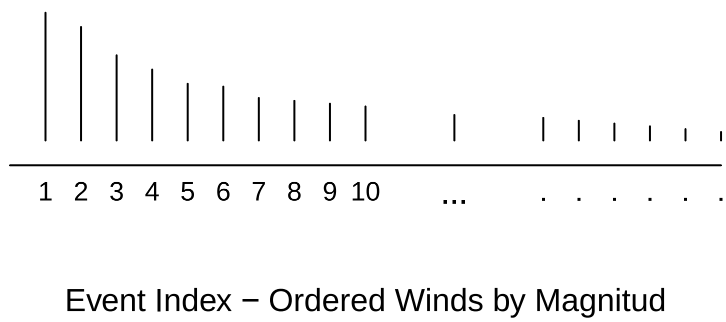

Figure 2.6: Gumbel cdf

$$P_e = \frac{Event\,index\,after\,descending\,sorting}{Years\,of\,simulations} = \frac{1}{10.000} = 0.001 = 0.01\%$$

because the highest wind will be the first in the sorted list. Same exercise can be done with all winds to construct the annual exedance probability curve, that in this case will represent the probability to equal or exceed different probable losses due to wind.

2.2.2 Return Period - Mean Recurrence Interval

Continuing with the previous section, if the inverse of the excedance probability is taken, the return period is obtained. The return period or Mean Recurrence Interval - MRI.

2.2.3 Compound Excedance Probability - Pn

2.3 Extreme Value Analysis Overview

Analysis of extreme values is related with statistical inference to calculate probabilities of extreme events. Main methods to analize extreme data are ephochal, Peaks Over Threshold - POT, and extreme index. The epochal method, also known as block maxima, uses the most extreme value for a specific frame of time, tipically, one year. POT is based in the selection of a single threshold value to do the analysis only with values above the threshold. But there are different POT aproaches, the most commond one is Generalized Paretto Distribution - POT-GPD, but also it is possible to use the Poisson process approach.

In both methods (Epochal and POT), the first step is to fit the data to an appropiate probability distribution model, among them the most used are, - Extreme Value Type I (Gumbel), Extreme Value Type II (Frechet), Weibull, Generalized Pareto - GPD, and Generalized Extreme Value - GEV.

Distribution models are fitted based in the estimation of its parameters, mommonly called location, scale and shape, nonetheless each model has its own parameters names. There are different methods to estimate parameters, among them, - method of moments (modified moments - see Kubler (1994), and L moments - see Hosking & Wallis (1997)), - method of maximum likelihood MLE, see Harris & Stocker (1998), which is problematic for GPD and GEV, - probability plot correlation coeficient, and - elemental percentiles (for GPD and GEV)

Once cadidate parameters are available, it is neccesary to assess the goodness of fit of the selected model, using one of the next methods, - Kolmogorov-Smirnov (KS) goodnes of fit test, and - Anderson-Darling goodness of fit test. Here a visual assesment is also useful using a probability plot or a kernel density plot with the fitted *pdf* overlaid.

The main use of the fitted model is the estimation of mean return intervals - MRI, and extreme wind speeds (return levels),

$$MRI = \frac{1}{1 - F(y)}$$

with $F(y)$ as the *cdf*. If $1 - F(y)$ is the annual excedance probability, MRI is its inverse, see Simiu & Scanlan (1996) for more details about MRI. If y is solved from previos equation using a given MRI of N-years, its value represents the Y_N wind speed return level,

$$Y_N = G\left(1 - \frac{1}{\lambda N}\right)$$

where G is the *ppf* (quantile function) and λ is the number of wind spees over the threshold per year.

The CRAN Task View "Extreme Value Analysis" `https://cran.r-project.org/web/views/ExtremeValue.html` shows available **R** for block maxima, POT by GPD, and external indexes estimation aproaches. Most important to consider are `evd`, `extremes`, `evir`, `POT`, `extremeStat`, `ismev`, and `Renext`.

2.4 Peaks Over Threshold - Poisson Process

According to Pintar, Simiu, Lombardo, & Levitan (2015) the stochastic poisson process is mainly defined by its intensity function. As the intensity function is nos uniform over the domain, the poisson process considered here is non-homogeneous, and due to the intensity function dependance of magnitud and time, it is also bi-dimmensional. Poisson Process was described for the first time in Pickands (1971), then extended in

Smith (1989).

$$\lambda\,(\mathrm{y,t}) \begin{cases} \lambda_t(y), & for\, t\, in\, thunderstorm\, period \\ \lambda_n t(y), & for\, t\, in\, non-thunderstorm\, period \end{cases} \tag{2.5}$$

Generic equation (**??**) shows the intensity function, which is defined in the domain $D = D_t \cup D_{nt}$, and allow to fit the poisson process at each station to the observed data $\{t_i, y_i\}_{i=1}^{I}$ for al the times (t_i) of threshold crossing observations and its corresponding wind speeds magnitudes (y_i). Thus, only data above the threshold is used.

Intensity function of the Poisson Process is defined in Smith (2004),

$$\frac{1}{\psi_t}\left(1+\zeta_t\frac{y-\omega_t}{\psi_t}\right)_+^{-\frac{1}{\zeta_t}-1}$$

Where ζ_t controls the tail lengh of the intensity function at a given time t, but to facilitate the estimation of the parameters then ζ_t is taken to be zero, then doing the limit, the resulting intensity function is the same as the the GEV type I or Gumbel distribution,

$$\frac{1}{\psi_t}\exp\left\{\frac{-(y-\omega_t)}{\psi_t}\right\}$$

In this study, the used intensity functions are shown in ecuation (2.6).

$$\lambda\,(\mathrm{y,t}) \begin{cases} \dfrac{1}{\psi_s}\exp\left\{\dfrac{-(y-\omega_s)}{\psi_s}\right\}, & for\, t\, in\, thunderstorm\, period \\ \dfrac{1}{\psi_{nt}}\exp\left\{\dfrac{-(y-\omega_{nt})}{\psi_{nt}}\right\}, & for\, t\, in\, non-thunderstorm\, period \end{cases} \tag{2.6}$$

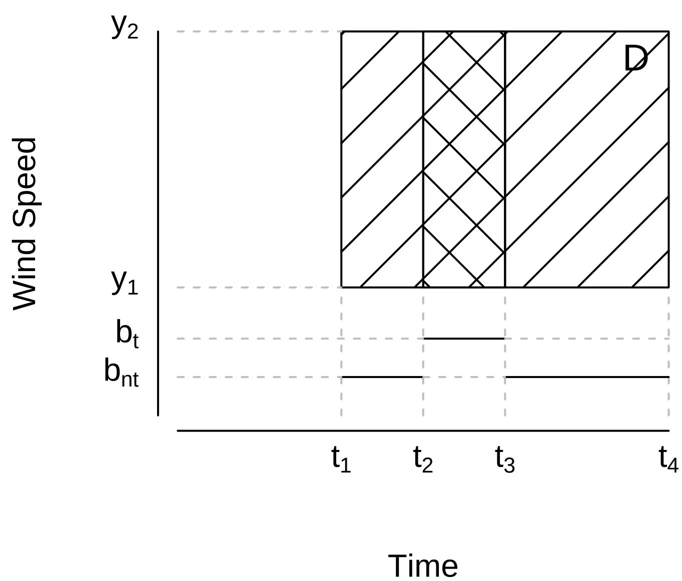

Figure 2.7: Domanin off the Poisson Process

Figure 2.7 represent the domain D of the Poisson process. In time, the domain represents the station service period from first sample t_1 to last sample t_4. D is the union of all thunderstorm periods D_t (from t_2 to t_3), and all non-thunderstorm periods D_{nt} (periods t_1 to t_2 and t_3 to t_4). In magnitud, only thunderstorm data above its threshold b_t, and only non-tunderstorm data above its threshold b_{nt} are used.

Thunderstoms and non-thunderstorms are modeled independently:

1. Observations in domain D follow a Poisson distribution with mean $\int_D \lambda(t, y)\, dt\, dy$
2. For each disjoint subdomain D_1 or D_2 inside D, the observations in D_1 or D_2 are independent random variables.

Visual representation of the intensity function for the Poisson Process can be seen in figure 2.8. In vertical axis, two surfaces were drawn representing independent intensity functions for thunderstorm $\lambda_t(y)$ and for non-thunderstorm $\lambda_{nt}(y)$. The volume under each surface for its corresponding time periods and peak (over threshold) velocities, is the mean of the Poisson Process.

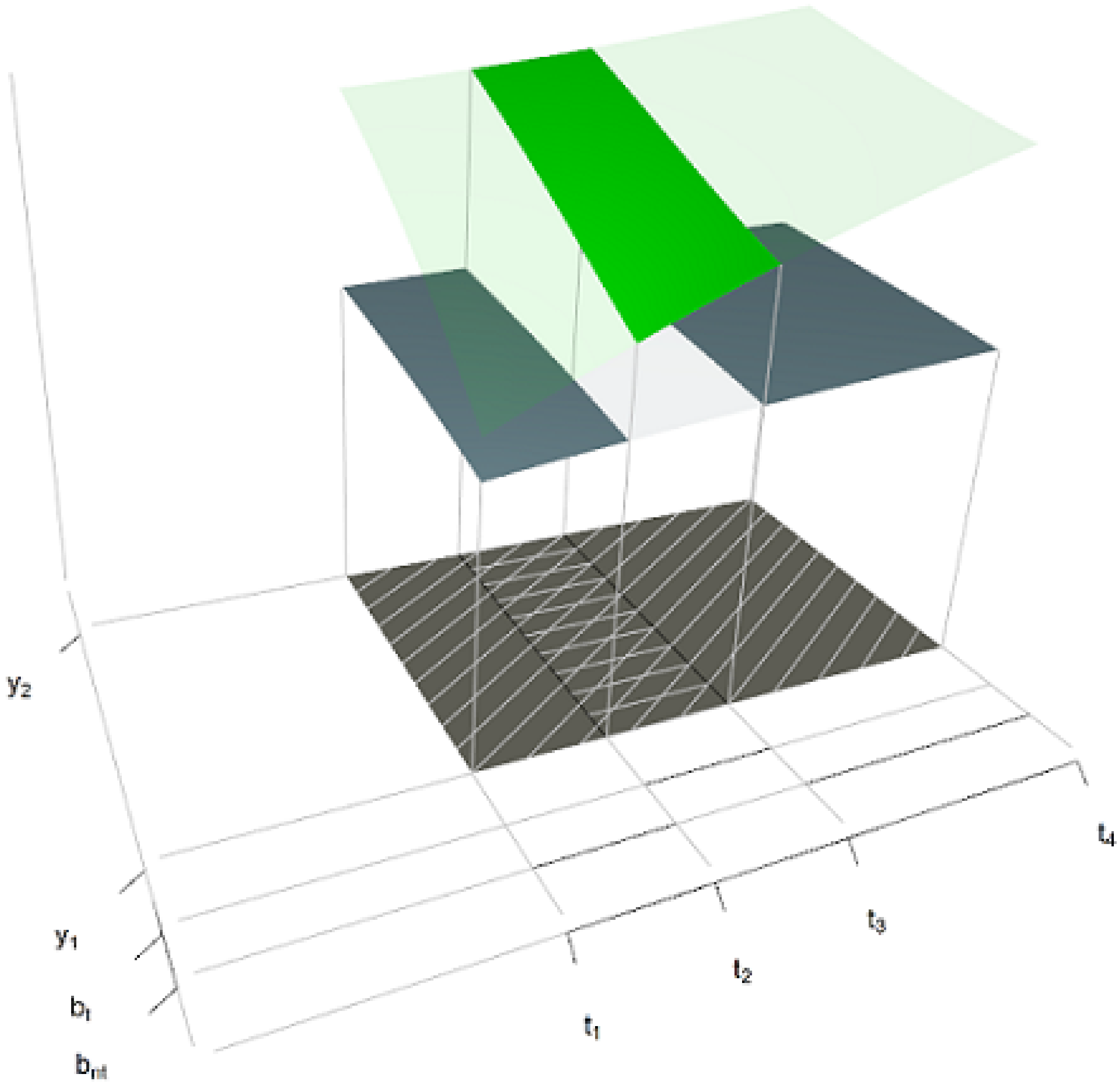

Figure 2.8: Volume under surfaces represents the mean of the Poisson process

The method of maximun likelihood es used to estimate the parameters of the Poisson process, the selected vector of parameters η are the $\hat{\eta}$ values that maximizes the function

$$\mathrm{L}(\eta) = \left(\prod_{\mathrm{i}=1}^{\mathrm{I}} \lambda\left(\mathrm{y_i}, \mathrm{t_i}\right)\right) \exp\left\{-\int_{\mathrm{D}} \lambda\left(\mathrm{y}, \mathrm{t}\right) \mathrm{dy}\, \mathrm{dt}\right\} \tag{2.7}$$

$\hat{\eta}$ values need to be calculated using a numericall approach because there is not analytical solution available.

Once the Poisson process is fittet to the data, the model will provide extreme wind velocities (return levels), for different return periods (mean recurrence intervals).

A Y_N extreme wind velocity, called the return level (RL) belonging to the N-years return period, has a expected frequency to occur or to be exceeded (annual excedance probability)$P_e = \frac{1}{N}$, and also has a probability that the event does not occur (annual non-excedance probability) $P_{ne} = 1 - \frac{1}{N}$. Y_N will be the resulting value of the G (ppf or quantile) function using a probability equal to P_{ne}. $Y_N = quantile(y, p = P_{ne}) = G(x, p = P_{ne}) = ppf(x, p = P_{ne})$. As for this study $\zeta = 0$, the G function to use is the Gumbel quantile function. Y_N can be undestood as the wind extreme value expected to be exceeded on average once every N years.

For different POT approaches, as POT-GPD described –, the value of the probability passed to the G function, has to be modified with the λ parameter, as is described in next equation. λ is the number of wind speed over the threshold per year.

$$Y_N = G\left(y, 1 - \frac{1}{\lambda\, N}\right)$$

For the Poisson process Y_N is also the solution to the next equation, which is defined in terms of the intensity function,

$$\int_{\mathrm{Y_N}}^{\infty}\int_{0}^{1} \lambda\left(\mathrm{y},\mathrm{t}\right)\mathrm{dydt} = \mathrm{A_t}\int_{\mathrm{Y_N}}^{\infty} \lambda_{\mathrm{t}}\left(\mathrm{y}\right)\ \mathrm{dy} + \mathrm{A_{nt}}\int_{\mathrm{Y_N}}^{\infty} \lambda_{\mathrm{nt}}\left(\mathrm{y}\right)\ \mathrm{dy} = \frac{1}{\mathrm{N}} \tag{2.8}$$

where A_t, is the multiplication of the average number of thunderstorm per year and the average lengh of a thunderstorm (taken to be 1 hour as defined in Pintar et al. (2015)), and $A_{nt} = 1 - A_t$. The average length of a non-thunderstorm event is variable, and it is adjusted in each station to guarantee that $A_{nt} + A_t = 1$

The same thunderstorm event in considered to occur if the time lag distance between sucesive thunderstorm samples is small than six hours, and for non-thunderstorm this time is 4 days. For the Poisson process, all the measurements belonging to the same event (thunderstorm or non tunderstorm), need to be declustered to leave only one maximun value. In other words, the number of thunderstorm in the time serie is the number of time lag distances grather than 6 hours, and for non-thunderstorm grather than 4 days.

###Threshold Selection

$$U = F(Y)$$

$$W = -log(1 - U)$$

Chapter 3

Methodology

Placeholder

3.1 Input Data Selection and Standarization

3.1.1 Data Selection

3.1.2 Data Standarization

Anemometer height - 10 m

Surface Roughness - 0.03 m

Averaging Time - 3-s gust

3.1.3 Data Filterng

3.2 Fit data to a POT - Poisson Process

3.2.1 Data Requirements

3.2.2 Exploratory Data Analysis and Data Preparation

Declustering of observations

Exclude no-data periods

Threshold selection

3.2.3 Parameters Estimation

Intensity function

Density function

Distribution function

Maximun likelihood estimation

3.2.4 Velocities at Return Periods

3.3 spatial Interpolation

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

In Chapter 3:

Appendix B

The Second Appendix, for Fun

References

Placeholder

Harris, J. W., & Stocker, H. (1998). Maximum likelihood method. In *Handbook of mathematics and computational science* (p. 824). Springer-Verlag.

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional frequency analysis.* Cambridge University Press. `http://doi.org/10.1017/cbo9780511529443`

Kubler, J. (1994). *Computational Statistics &Amp; Data Analysis*, *18*(4), 473–474. Retrieved from `https://EconPapers.repec.org/RePEc:eee:csdana:v:18:y:1994:i:4:p:473-474`

Pickands, J. (1971). The two-dimensional poisson process and extremal processes. *Journal of Applied Probability*, *8*(4), 745–756. `http://doi.org/10.2307/3212238`

Pintar, A. L., Simiu, E., Lombardo, F. T., & Levitan, M. L. (2015). *Simple guide for evaluating and expressing the uncertainty of NIST MeasuremenMaps of non-hurricane non-tornadic wind speeds with specified mean recurrence intervals for the contiguous united states using a two-dimensional poisson process extreme value model and local regressiont results.* National Institute of Standards; Technology.

Simiu, E., & Scanlan, R. H. (1996). *Wind effects on structures : Fundamentals and applications to design* (3rd ed.). New York : John Wiley. Retrieved from `http://lib.ugent.be/catalog/rug01:001267836`

Smith, R. L. (1989). Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. *Statistical Science*, *4*(4), 367–377. `http://doi.org/10.1214/ss/1177012400`

Smith, R. L. (2004). Extreme values in finance, telecommunications, and the environment (chapman & hall/crc monographs on statistics and applied probability). In B. F. inkenstädt & H. Rootzén (Eds.), (pp. 1–78). Chapman; Hall/CRC. Retrieved from `https://www.amazon.com/Telecommunications-Environment-Monographs-Statistics-Probability/dp/1584884118?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=1584884118`