Sería bueno agregar una sub-sección independiente que contenga la descripción del caso de estudio.

the gaps in which this research aims to contribute are related to the use of alternative data sources, and how to meet the downscaling challenge, considering that the ~~main drawback of the research is the~~ lack of field measurement data coming from weather stations.

1.4 Research Aim and Objectives

The Main aim of this research is the estimation of wind extreme velocities to be used as input loads for the design of ~~different types of~~ structures, considering ~~its~~ their risk categories, and covering any place in the whole study area.

Specific objectives are:

1. Complement the lack of field measured wind data, with other sources of information, then, analyze and compare different time series, to select and use the best data source (or combination of sources) ~~for research~~, based on objective criteria, for instance similitude, completeness, coverage, etcetera, to achieve in this way a formal support for the decision made in this regard, in case of downscaling issue.
2. Select and apply a suitable probabilistic method to infer wind maps for infrastructure design, ~~that allows to fulfill wind load requirements defined for the respective authority in the study area~~.
3. Estimate ~~needed~~ extreme wind values for the stations in the selected input data source, considering non-hurricane approaches.
4. Allow the comparison of wind extreme values estimations, using different methods, ~~in order~~ to verify and calibrate output results.
5. Generate ~~needed~~ continuous non-hurricane wind maps, using the most suitable spatial interpolation technique, considering the specific characteristics of the input data source~~, advantages, and disadvantages of the selected methods~~.
6. Combine output maps from non-hurricane analysis with existing hurricane studies to ~~allow the inclusion of~~ obtain final maps ~~in the~~ for structural design ~~standard~~ purposes.

Separar en otra oración

1.5 Research Question

Main question of this research is directed to calculate future extreme velocities (return levels) for infrastructure design, then the research question could be

What extreme velocities (return levels) need to be used as load design forces for structures of different use category, in the study area?

If we remember that, for the ~~case study area (Colombia), there are~~ predefined requirements or mean return intervals - MRI to design structures depending of it use category, and that this MRI values are 700, 1700, and 3000 years, the research question could be more specific.

What extreme velocities (return levels) will be equaled or exceeded with a probability equal to $\frac{1}{MRI}$ in a given year?

Integrar esto en la introducción

agregar redacción

Muy coloquial

Usualmente la pregunta de investigación es solo una.

What extreme velocities (return levels) will be equaled or exceeded only one time in the period defined for this specific MRIs: 700, 1700, and 3000 years?

If we consider not only the annual exceedance probability $\frac{1}{MRI}$, but also the exposure time (compound probability), understood as the time the structure will be in use, then the question will be

What extreme velocities (return levels) will have a occurrence compound probability of 67%, when the exposure time of the structure will be equal to the main return intervals 700, 1700, and 3000 years?

1.6 Outline

Main sections of thesis document are 1) Introduction, 2) Data, 3) Theoretical Framework, 4) Methodology, 5) Results and Discussion, 5) Conclusions, and 6) ~~five~~ Annexes, from A to E.

After introduction, in second section **Data**, main information about data sources IDEAM, ISD, and ERA5 are described, including at the end, details for data download and organization, topic that is complemented for ERA5, with the content of the Annex C~~ - ERA5 Data Download and Integration~~.

Theoretical framework section is dedicated to introduce statistical concepts that are basis for the investigation, both in **probability distributions** (~~density function - pdf, distribution function - cdf, percent point function - ppf, and hazard function - hf~~) and in **extreme analysis** (~~annual exceedance probability - P_e, recurrence interval - MRI, and compound exceedance probability - P_n~~). Later, it is described in more detail, topics related to **extreme value analysis** (peaks over threshold with generalized Pareto - POT-GPD, and peaks over threshold with Poisson process - POT-PP), and at the end, a summary report is done about **wind load requirements** for the study~~, which is foundation for addressing the research~~.

The chapter the

In methodology ~~are described~~ main processes needed to meet the objectives and answer the research question, which, broadly speaking, are data standardization, downscaling support, POT-PP, spatial interpolation and integration with hurricane data.

(1)

Results and discussion section shows ~~first~~ all results for data standardization and comparison to support the downscaling issue, ~~second~~, all POT-PP results ~~are set out~~ for one ISD station. (3) ~~then~~ all output maps for ISD and ERA5 data sources ~~are exhibited~~, including discussions ~~without bias, about the good and the bad~~ of those finals results. These discussions are complemented by the ~~following~~ conclusions section. (2)

To finalize the document, a series of appendices were created, to facilitate the reproducibility of the research. Appendix A contains *research R code*, ~~but~~ It is necessary to keep in mind that the code provided by Dr. Adam Pintar to do the de-clustering and thresholding in POT-PP is not there because its publication and distribution is not authorized. Appendix B contains all *results in digital format*. Appendix C compliments the information needed to *download*

Because

and organize data from the sources, mainly with details related to ERA5. ~~As~~ the document for the thesis was done using package ´thesisdown´, which is based in ´bookdown´, the most important *document R code* to create the thesis document, mainly graphics, is shown in Appendix D. Then at the end, in Appendix E, an *user manual* is presented, in order to provide instruction to apply the same methodology in a different case study~~, and provided using R code~~.

En general, la redacción tiene demasiadas palabras, repetición y errores de puntuación.
Las oraciones son demasiado largas.
Por favor corregir estos aspectos en el resto del documento.

make up: es maquillar

Chapter 2

Data

IDEAM

was obtained from

Input data ~~is made up of~~ three different sources a) ~~IDEAM~~ Institute of Hydrology, Meteorology and Environmental Studies of Colombia `http://www.ideam.gov.co`, b) ~~ISD~~ Integrated Surface Database `https://www.ncdc.noaa.gov/isd`, and c) ~~ERA5~~ climate reanalysis `https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5`.

ERA5

ISD

ampliar

Table 2.1: Datasets description

Institution	Dataset	Details
IDEAM	Historical records at weather stations	IDEAM is responsible for the instalation, maintenance and management of all kind of weather stations located everywhere along the country.
NOAA	ISD	ISD (Integrated Surface Database. NOAA's National Centers for Environmental Information - NCEI) Lite: A subset from the full ISD dataset containing eight common surface parameters in a fixed-width format free of duplicate values, sub-hourly data, and complicated flags.
ECMWF	ERA5	ERA5 is a reanalysis dataset with hourly estimates of atmospheric variables with horizontal resolution of 0.25° (33 kilómeters), ~~this is~~ equally spaced ~~cells~~ every 0.25 degrees

i.e. cells

Table 2.2: Datasets variables

Dataset	Variables	Description
IDEAM	vv_aut_2	Instantaneous wind velocity each two (2) minutes
	vv_aut_10	Instantaneous wind velocity each ten (10) minutes
ISD	v5	Maximun hourly five seconds (5-s) wind gust velocity
ERA5	fg10	10 metre wind gust since previous post-processing
	fsr	Forecast Surface Roughness

Table 2.3: Variables units and time

Variable	Units	Time	Stations
vvmx_aut_60	meters per second	Variable from 2001 until today. Irregular time series.	203
Wind speed	meters per second	Variable from 1941 until today. Note: There is too much variability in time (start, end, and time range) for each station. Irregutal time series.	101
fg10	meters per second	1979-Today	3381
fsr	meters per second	1979-Today	3381

Ideal data source to create extreme wind speeds maps should be field observed data from IDEAM, ~~but~~ there are not enough number of stations around the study area to represent all the local wind variability in a huge country with multiple variety of climates and and changing thermal floors. ~~but there are other important motivation to include different sources trying to improve output results:~~

but the ~~data~~ ~~for~~ IDEAM data have the following ~~deficiencies~~: deficiencies: (1)

- ~~As just mentioned, low quantity of IDEAM stations~~
- (2) There are uncertainties related to the way IDEAM anemometers are registering data, then comparison with other data sources are needed to be able to do appropriate data standardization, needed as a prerequisite to the analysis.
- (3) There is no time continuity in the registration of IDEAM data. Historical time series are different and variable in each station.

Importance of ISD database for this study is based on the fact that post-processed ISD database has wind extreme values, and it was used to create extreme wind maps for United States. ISD allows comparison with IDEAM records to take better decisions in order to ~~do~~ ~~needed~~ data standardization.

Despite that ERA5 data are not observed data (but forecast), its main advantage is ~~data~~ availability ~~to assess the local climatic variance every 33 square kilometers.~~

conduct their resolution and availability. (§3 & 2)

Unir.

2.1 IDEAM

Historical observed wind speeds from 203 stations in Colombia are managed by the official environmental authority IDEAM. Table 2.4 shows a sample of five IDEAM stations. Figure 2.1 shows a map of IDEAM stations.

Table 2.4: IDEAM Stations sample

Name[Code]	Latitud	Longitud
EMAS - AUT [26155230]	5.09	-75.51
SAN BENITO - AUT [25025380]	9.16	-75.04
AEROPUERTO ALFONSO LOPEZ - [28025502]	10.44	-73.25
TIBAITATA - AUT [21206990]	4.69	-74.21
ELDORADO CATAM - AUT [21205791]	4.71	-74.15

Incluir toda la tabla en la misma hoja

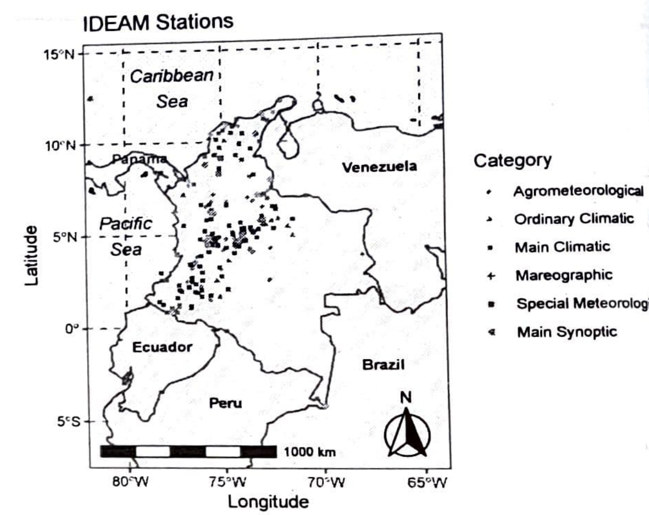


Figure 2.1: IDEAM Stations. Colombia

Figure 2.2 shows data: for

~~Following, the time series, auto~~correlation function. ~~and partial autocorrelation~~ function for IDEAM station "21205791" will ~~be displayed~~.

Evitar parrafos de una sola oración

Alex, todas las figuras deben estar referidas en el texto. Las figuras ~~son~~ no hacen parte del texto por lo tanto su posicion en un documento es variable y textos como estos son incorrectos: "La siguiente figura muestra..."

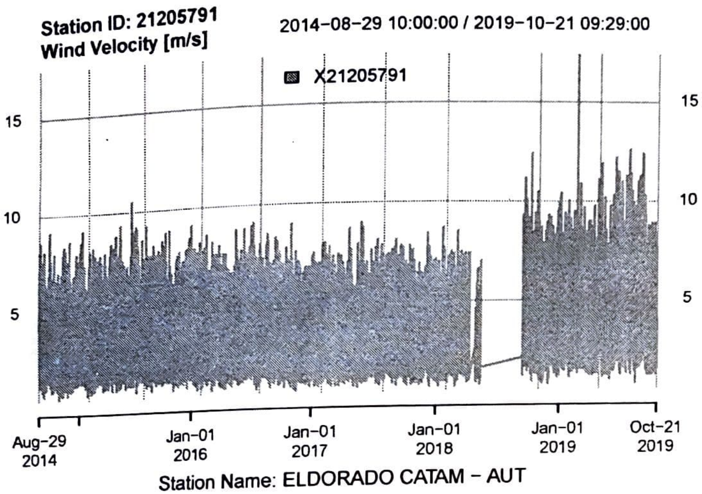


Figure 2.2: ~~IDEAM Station - Time Series~~

Time series of IDEAM Station ......

2.2 ISD

ISD is a database with environmental variables among then extreme wind speeds. ISD has data for the whole planet, and is based on observed data at meteorological stations in each country, which means that for Colombia is based on IDEAM data. Main advantage is data availability at neighbor countries and specialized post-processing made by NOAA's National Centers for Environmental Information - NCEI in United States, which facilitates its use. Table 2.5 shows a sample of five ISD stations. Figure 2.3 shows a map of ISD stations. 5 and 2.4 and data from ISD station "802590".

Table 2.5: ISD Stations sample

Code	Name	Latitud	Longitud
804400	BARINAS	8.62	-70.22
800810	ALTO CURICHE	7.05	-76.35
801000	BAHIA SOLANO / JOSE MUTIS	6.18	-77.40
802590	ALFONSO BONILLA ARAGON INTL	3.54	-76.38
803150	BENITO SALAS	2.95	-75.29

El formato de las figuras se puede mejorar manteniendo el mismo tamaño de fuente en todo el documento. ~~Idealmente,~~ Idealmente, el tipo de fuente debería ser el mismo que el documento.

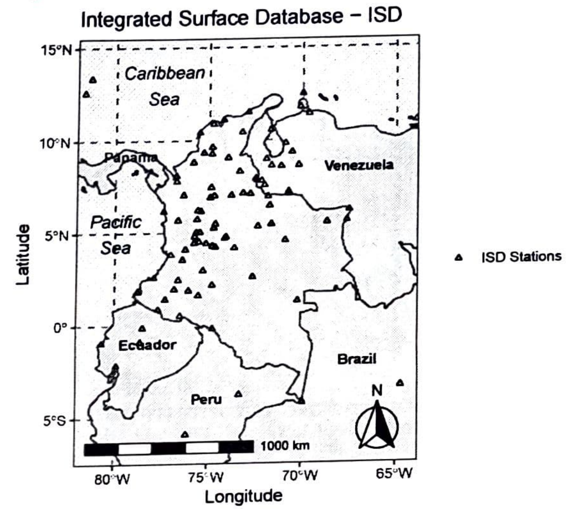


Figure 2.3: ISD Stations. Colombia and surrounding~~s~~ countries.

~~Following, the time series, autocorrelation function, and partial autocorrelation function, for ISD station "802590" will be displayed.~~

mejorar las descripciones de las figuras.

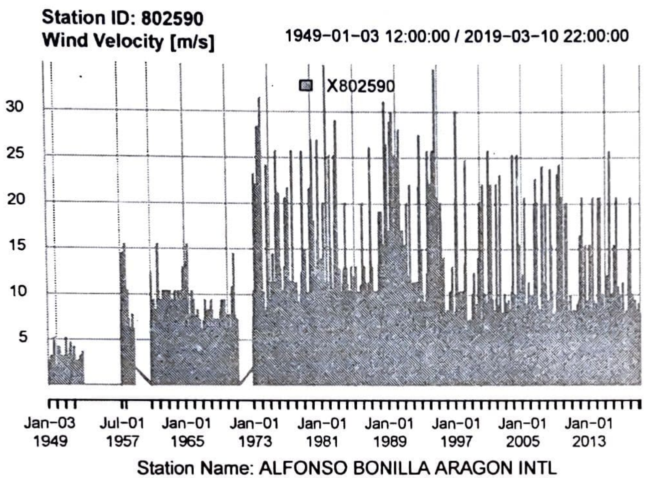


Figure 2.4: ISD Station - Time Series

2.3 ERA5

ERA5 is forecast reanalysis data processed by the *European Centre for Medium-Range Weather Forecasts* - ECMWF with wind speeds time series in square cells *matrix of pixels* of 0.25 degrees (33 km) covering the whole planet. For the study area was extracted a raster of 69 rows by 49 columns in format netCDF. Figure 2.5 shows a map of ERA5 stations (cells centers).

it

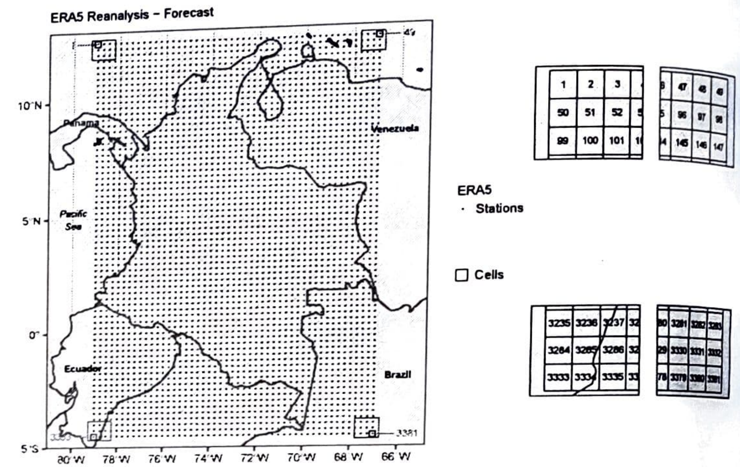


Figure 2.5: ERA5 Cells and Stations (cells centers). 49 cols by 69 rows. Cell size 0.25 decimal degrees (aprox 28 km in Colombia). Station IDs from 1 (lon=-79, lat=12.5) to 3381 (lon=-67, lat=-4.5)

2.4 Data Download and Data Organization

All data sources had different mechanisms for downloading. For IDEAM, the official procedure is through a written request using the e-mail *atencionalciudadano@ideam.gov.co*, then they will provide a link to get the information. For ISD, all files are available in the ftp site `ftp://ftp.ncdc.noaa.gov/pub/data/noaa/isd-lite/`, organized in folders by years, then a gzip file is available, with the station name in the format *ID-99999-YYYY.gz*, where ID is the USAF-ISD station identifier. For ISD, there are many files by station, one file for each year with data available. For ERA5 it is possible to make a request using a Python script, but since there is a size limit for downloading, it is necessary to split the request, then, use console commands to create an unified netCDF file. Files with all IDEAM and ISD stations are available in ~~the~~ Annex A ~~- Results - Digital Files~~. For the Python code, and the commands to join netCDF files of ERA5 data source, see the Annex C ~~- ERA5 Data Download and Integration~~.

Mejorar gramática de las descripciones.

no colocar los nombres.

Chapter 3

Theoretical Framework

"Este error se llama "run on": Dos oraciones independientes no están separadas por una coma"

3.1 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.

3.1.1 Probability Density Function - *pdf*

Pdf defines the probability that a continuous variable falls between two points, this is, in *pdf* the probability is related to the area below the curve (integral) between two points, as for continuous 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.

Si es un acrónimo debe ir en mayúsculas. Quitar los guiones.

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

Equation (3.1) is the Gumbel *pdf*.

is

$$f(x) = \frac{1}{\beta}\exp\left\{-\frac{x-\mu}{\beta}\right\}\exp\left\{-\exp\left\{-\left(\frac{x-\mu}{\beta}\right)\right\}\right\}, \quad -\infty < x < \infty \tag{3.1}$$

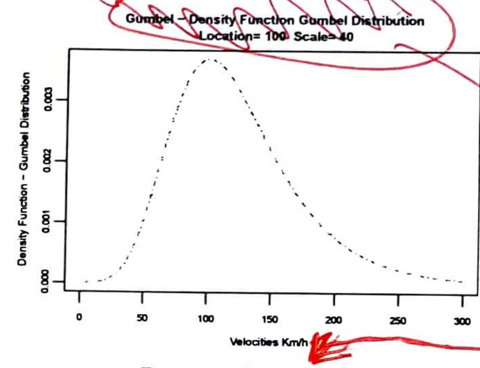
where $\exp\{.\}$ is $e^{(.)}$, β is the scale parameter, and μ is the location parameter. Location (μ) has the effect to shift the *pdf* to left or right along 'x' axis, thus, if location value is changed the effect is a movement of *pdf* to the left (small value for location), or to the right (big 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 3.1 shows *pdf* with location (μ) = 100 and scale (β) = 40, using Equation (3.1).

PDF

Porque no está referenciado ni numerado?

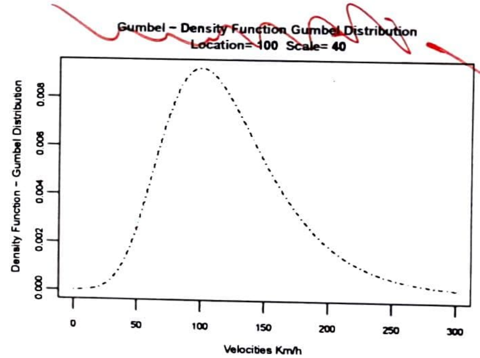
Mejorar redacción. Oraciones demasiado largas y sin conectores



Colocar este texto aquí.

Figure 3.1: Gumbel pdf

Figure 3.2 shows *pdf* with location (μ) = 100 and scale (β) = 40, using function `dgumbel` of the package `RcmdrMisc`



No usar doble descripción. Colocar esto aquí.

Figure 3.2: Gumbel pdf - dgumbel function

3.1.2 Cumulative Distribution Function - *cdf*

CDF

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

CDF

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

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

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

Equation (3.2) is the Gumbel *cdf*.

$$\mathrm{F(x)} = \exp\left\{-\exp\left[-\left(\frac{\mathrm{x}-\mu}{\beta}\right)\right]\right\}, \quad -\infty < \mathrm{x} < \infty \tag{3.2}$$

Figure 3.3 shows Gumbel *cdf* with location (μ) = 100 and scale (β) = 40, using Equation (3.2). As previously done with *pdf*, similar result can be achieved using function `pgumbel` of package `RcmdrMisc`.

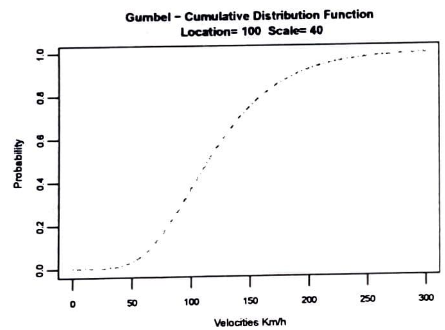


Figure 3.3: Gumbel cdf

3.1.3 Percent Point Function - ~~ppf~~ PPF

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 (3.3) is the Gumbel *ppf*.

$$\mathrm{G}(\alpha) = \mu - \beta\ln(-\ln(\alpha)) \quad 0 < \alpha < 1 \tag{3.3}$$

Figure 3.4 shows Gumbel *ppf*, using Equation (3.3). Similar result can be achieved using function `qgumbel` of package `RcmdrMisc`.

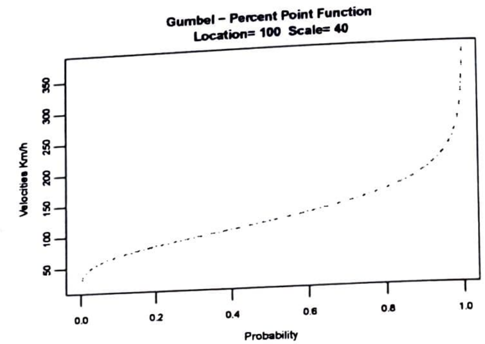


Figure 3.4: Gumbel ppf

3.1.4 Hazard Function - ~~hf~~ HF

SF

Using $S(x) = 1 - F(x)$ as survival function ~~sf~~, the probability that a variable takes a value greater 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 (3.4) is the Gumbel *ppf*.

$$\mathrm{h(x)} = \frac{1}{\beta}\frac{\exp(-(\mathrm{x}-\mu)/\beta)}{\exp(\exp(-(\mathrm{x}-\mu)/\beta)) - 1} \tag{3.4}$$

Figure 3.5 shows Gumbel *hf*, using Equation (3.4).

Esto no es claro!

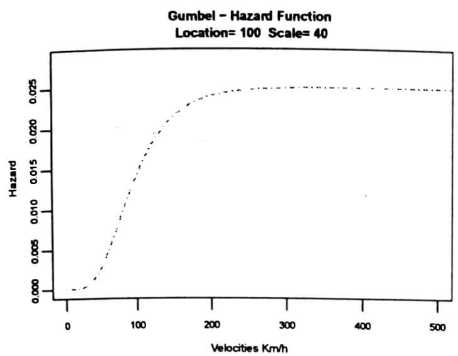


Figure 3.5: Gumbel hf

Esta oración no tiene sujeto.

3.2 Statistical Concepts For Extreme Analysis

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 exceedance probability, mean recurrence interval MRI, exposure time, and compound probability for any given exposure time and MRI. As an hypothetical example, a simulated database of extreme wind speed will be used. This database is supposed to have 10.000 years of ~~simulated~~ wind speeds.

3.2.1 Annual Exceedance Probability - P_e

cual pérdida

No es claro

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 ~~grater~~ or equal (in this case equal) to the highest 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 exceedance probability Pe with next formula

Greater

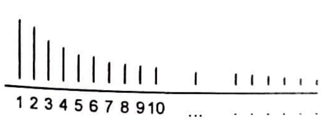


Figure 3.6: Sorted Winds by Magnitude - ~~wind simulation database~~

velocities

Porque pérdida?

No es muy claro el flujo de tus ideas.

el texto va sin formato de ecuación

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

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

porque se habla de pérdida

3.2.2 Return Period ~~-~~ or Mean Recurrence Interval - MRI

Continuing with the previous section, if the inverse of the exceedance probability is taken, the return period (in years) is obtained. The return period or Mean Recurrence Interval - MRI is associated with an specific return level (wind extreme velocity). MRI is the numbers of years (N) needed to obtain 63% of change that the corresponding return level will occur at least one time in that period. The return level is expected to be exceeded on average once every N-years. The annual exceedance probability of the return level corresponding to N-years of MRI, is $P_e = \frac{1}{MRI} = \frac{1}{N}$.

For an specific wind extreme event A, the probability that the event will occur in a period equal to MRI years is 63%. If we analyze for the same period a strongest wind extreme event B, its occurrence probability will be ~~lest~~ less than 67%. If the purpose of this research is to design infrastructure considering wind loads, the structure will be more resistant to wind if we design with stronger winds, this is high MRIs, and low annual exceedance probability. Common approach for infrastructure design, considering any type of load (earthquake, wind, etc) is to choose high MRI according to the importance/use/risk/type of the structure. For highly important structures, like hospitals or coliseums, where the risk of collapse must be diminished, the MRI used to design is higher in comparison to common structures (for instance a normal house), which implies less risks for its use and importance.

$$P_e = \begin{cases} 1 - \exp\left(-\frac{1}{MRI}\right), & for\ MRI < 10\, years \\ \frac{1}{MRI}, & for\ MRI \geq 10\, years \end{cases}$$

esta ecuación no se menciona en el texto

3.2.3 Compound Exceedance Probability - Pn

If time of exposure is consider, understood as time the structure will be in use, it is possible to have a compound probability P_n, where n is the exposure period. P_n is the probability that the extreme wind speed will be equaled or exceeded at least one time in n years, and is related with the occurrence probability, but also is possible to calculate the non-occurrence compound probability (probability that the event will not occur).

$$P_n = \begin{cases} 1 - \left(1 - \frac{1}{MRI}\right)^n, & occurrence\ probability \\ \left(1 - \frac{1}{MRI}\right)^n, & non-occurrence\ probability \end{cases}$$

De hecho, creo que el capítulo 3 no se necesita y se puede borrar todo e incluir los aspectos mas relevantes en el capítulo 4.

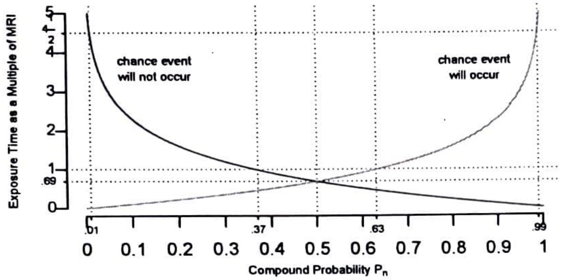


Figure 3.7: Compound Probability

If it is consider exposure time as a multiple of return period, the resulting Figure 3.7, shows that:

- When exposure time is .69% of the return period, then probability (occurrence and non-occurrence) will be 50%
- As was stated previously, when exposure time is equal to return period, then the probability that the extreme wind speed (return level) occur is 63%, and 37% for the non occurrence probability.
- If exposure time is 4.5 times the return period, there is a 99% of change that the return level will occur.

The example discussed here was presented as an instrument to introduce important concepts, nonetheless,there are specialized approaches to deal with extreme value analysis which will be discussed in Extreme Value Analysis Overview and more in detail in Peaks Over Threshold - Poisson Process. In summary, is necessary to fit the data over a specific threshold to an extreme value distribution, and P_c will be $1 - F(y)$, with F(y) as the *cdf*, and MRI as $\frac{1}{1-F(y)}$.

3.3 Extreme Value Analysis Overview

Analysis of extreme values is related with statistical inference to calculate probabilities of extreme events. Main methods to analyze extreme data are epochal, 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, typically, 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 approaches, the most common one is Generalized Pareto 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 appropriate probability distribution model, among them the most used are, - Extreme Value Type I

En mi concepto el capítulo 3 debería llamarse Extreme Value Analysis e incluir solo 3.3

No parece correcto: En POT hay 2 dimensiones: el tiempo y la magnitud. Uno puede trabajarlas separadas con Pareto ó Poisson ó unidas con Poisson 2D

En el documento de Pintor esto bastante bien explicado

algunos autores solo mencionan dos métodos; clásico y POT

No usar guiones

No combinar ideas no relacionadas en la misma oración.

(Gumbel), Extreme Value Type II (Fréchet), Weibull, Generalized Pareto - GPD, and Generalized Extreme Value - GEV.

Distribution models are fitted based in the estimation of its parameters, commonly 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 coefficient, and - elemental percentiles (for GPD and GEV)

or

Once candidate parameters are available, it is necessary to assess the goodness of fit of the selected model, using one of the next methods: - Kolmogorov-Smirnov (KS) ~~goodness of fit test, and~~ Anderson-Darling ~~goodness of fit~~ tests. Here a visual assessment 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 exceedance probability, MRI is its inverse, see Simiu & Scanlan (1996) for more details about MRI. If y is solved from previous 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 speeds 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 approaches. Most important to consider are `evd`, `extremes`, `evir`, `POT`, `extremeStat`, `ismev`, and `Renext`.

3.3.1 POT USING GPD AND 1D POISSON PROCESS

In POT using Pareto distribution, the magnitude of the observations above the threshold are assumed a) to be independent random variables with the same generalized Pareto as probability distribution, σ as scale, and ξ as tail length, and corresponding times are assumed b) to follow a one dimensional homogeneous Poisson process with γ as parameter. The *cdf* of POT-GPD is $F(y) = 1 - \left(1 - \xi\frac{y-b}{\sigma}\right)_+^{-\frac{1}{\xi}}$, where b is the threshold. In both GPD (magnitude), and $1D$ Poisson process (time), it is not possible to differentiate between thunderstorm and non-thunderstorm wind types.

POT { Magnitude: GPD
Time: Poisson 1D

In POT-GPD, to calculate return levels (RL), Y_N, corresponding to the N-years return period, next equation is used,

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

W

Where G is the quantile function (ppf), and the value of the probability passed to the G function, has to be modified with the λ parameter. λ is the number of wind speed events over the threshold per year.

using a 2D

3.4 Peaks Over Threshold Poisson Process POT-PP

According to Pintar et al. (2015) the stochastic Poisson Process - PP is mainly defined by its intensity function. As the intensity function is not uniform over the domain, the PP considered here is non-homogeneous, and due to the intensity function dependency of magnitude and time, ~~it~~ is also bi-dimensional. PP was described for the first time in Pickands (1971), then extended in Smith (1989).

$$\lambda\,(y,t) \begin{cases} \lambda_t(y), \; for\,t\,in\,thunderstorm\,period \\ \lambda_{nt}(y), \; for\,t\,in\,non-thunderstorm\,period \end{cases} \tag{3.5}$$

Generic Equation (3.5) shows the intensity function, which is defined in the domain $D = D_t \cup D_{nt}$, and allow to fit the PP at each station to the observed data $\{t_i, y_i\}_{i=1}^{I}$, for all the times (t_i) of threshold crossing observations, and its corresponding wind speeds magnitudes (y_i). Thus, only data above the threshold (POT) are used.

Intensity function of the PP is defined in Smith (2004), by:

$$\frac{1}{\psi_t}\left(1 + \zeta_t \frac{y - \omega_t}{\psi_t}\right)_{+}^{-\frac{1}{\zeta_t}-1} \tag{3.6}$$

Where, at a given time t, parameter $shape = \zeta_t$ controls the tail length of the intensity function, and the other two parameters ω_t and ψ_t define the location and scale of the intensity function.

Hay muchas ecuaciones que no se estan numerando.

Integrar las ecuaciones al texto

La ecuación debe aparecer después de mencionarla no antes

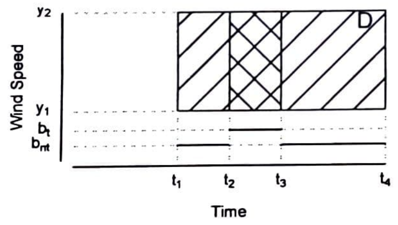


Figure 3.8: Domain off the Poisson Process - PP

Figure 3.8 represent the domain D of PP. 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 magnitude, only thunderstorm data above its threshold b_t, and only non-thunderstorm data above its threshold b_{nt} are used.

Thunderstorms 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 sub-domain 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 PP can be seen in Figure 3.9. 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 PP.

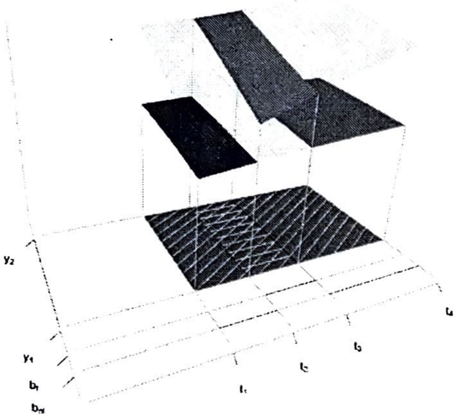


Figure 3.9: Volume under surfaces represents the mean of PP

To fit the intensity function to the data, the method of maximum likelihood is used to estimate its parameters, *scale* $= \psi$, *location* $= \omega$, and *shape* $= \varsigma$, the selected vector of parameters η are the $\hat{\eta} = (\hat{\psi}, \hat{\omega}, \hat{\varsigma})$ values that maximizes next function

$$\mathrm{L}(\eta) = \left(\prod_{i}^{I} \lambda\,(y_i, t_i)\right) \exp\left\{-\int_D \lambda\,(y,t)\,dy\,dt\right\} \tag{3.7}$$

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

Once the PP is fitted 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 exceedance probability) $P_e = \frac{1}{N}$, and also has a probability that the event does not occur (annual non-exceedance 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}. $Y_N = quantile(y, p = P_{nc}) = G(y, p = P_{nc}) = ppf(y, p = P_{nc})$. Y_N can be understood as the wind extreme value expected to be exceeded on average once every N years.

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

$$\int_{Y_N}^{\infty}\int_0^1 \lambda\,(y,t)\,dydt = \mathrm{A_t}\int_{Y_N}^{\infty} \lambda_t\,(y)\,dy + \mathrm{A_{nt}}\int_{Y_N}^{\infty} \lambda_{nt}\,(y)\,dy = \frac{1}{N} \tag{3.8}$$

no es claro

Eliminar o integrar en la metodología

where A_t, is the multiplication of the average number of thunderstorm per year and the average length of a thunderstorm, taken to be 1 hour as defined in Pintar et al. (2015), and $A_{nt} = 365 - A_t$. The average length of a non-thunderstorm event is variable, and it is adjusted for each station to guarantee that $A_{nt} + A_t = 365$. Value 365 is used only, if operations with time in the dataset are performed in days.

The same thunderstorm event in considered to occur if the time lag distance between successive thunderstorm samples is small than six hours, and for non-thunderstorm this time is 4 days. For PP, all the measurements belonging to the same event (thunderstorm or non-thunderstorm), need to be de-clustered to leave only one maximum value. In other words, the number of thunderstorm in the time series is one plus the number of time lag distances grater than 6 hours, and for non-thunderstorm grater than 4 days.

3.4.1 Threshold Selection

POT-PP needs selection of the best threshold pairs b_t and b_{nt} (see Figure 3.8) that produces the optimal fit. Measurement of this threshold fitting is done through W statistic. If wind variable y, in a POT-PP approach, has a $cdf = U = F(y)$, then $F(y)$ is distributed as Uniform between 0 and 1 - Uniform(0,1), meaning that the transformation $W = -log(1-U)$ is an exponential random variable with mean one (1).

$$\mathrm{cdf = U = F(y) = P(y \leq Y)} = \frac{\int_b^Y \lambda(y,t)\,dy}{\int_b^{\infty} \lambda(y,t)\,dy} \tag{3.9}$$

The procedure to choose the best thresholds pairs based in W transformation, is described in the methodology section called thresholding.

3.5 Wind Loads Requirements

As the output maps of this research will be used as input loads for infrastructure design, the methodology used for its creation, need to be consistent with Colombian official wind loads requirements. Colombian structure design code, from now the design standard, was created and it maintained by the Colombian Association of Seismic Engineering - AIS.

The design ~~standard~~ Standard is mainly based in *minimum design loads and associated criteria for buildings and other structures* ASCE7-16 ~~norm~~, see ~~Engineers~~ (2017). ~~Under these circumstances,~~ The ASCE7-16 standard defines the minimum requirements ~~of the research products~~ for design wind loads in. ~~Especially the~~ chapter 26 ~~"wind loads - general requirements", C26.5 "wind hazard map"~~, and C26.7 ~~"Exposure"~~ pages 733 to 747. Wind speeds requirements of ASCE7-16 are based in the combination of independent non-hurricane analysis, and hurricane wind speeds simulations models. The focus of this research will be the analysis of non-hurricane wind data, however, existing results of hurricane studies will be used to present final maps with both components. In ASCE7-16, for non-hurricane wind speed, the procedure is mainly based on Pintar et al. (2015).

ASCE7-16 (page 734), requires the calculation of wind extreme return levels for specific return periods according to the risk category of the structure to be designed as follows: risk category I

- 300 years, risk category II - 700 years, risk category III - 1700 years, risk category IV - 3000 years. ~~The design standard only requires 700, 1700 and 3000 years.~~ In addition, extreme wind speeds for those MRI need to correspond to 3 second gust speeds at 33 ft (10 meters) above the ground, and exposure category C ~~(open space)~~.

- Risk IV - This are 'indispensable buildings' that involve substantial risk. These structures that can handle toxic or explosive substances.
- Risk III - There is substantial risk because these structures that can handle toxic or explosive substances, can cause a serious economical impact, or massive interruption of activities if they fail.
- Risk II - Category 'by default', and correspond to structures not classified in others categories.
- Risk I - This structures represent low risk for life of people.

esto está desligado del texto!

To standardize wind speeds to gust speeds ASCE7-16 proposes the curve Durst (see C. S. Durst (1960), and Figure 3.10). Durst curve is only valid for open terrain conditions, and it shows in axis y the gust factor $\frac{V_t}{V_{3600}}$; a ratio between any wind gust averaged at t seconds, V_t, and the hourly averaged wind speed V_{3600}, and in the axis x the duration t of the gust in seconds.

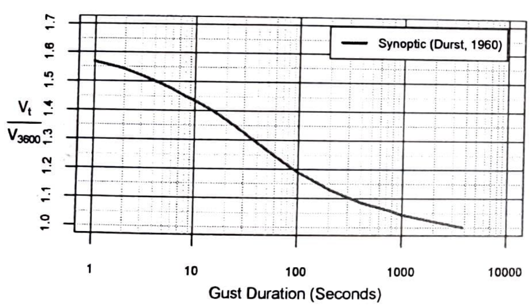


Figure 3.10: Maximum speeds averaged over t (sec), to hourly mean speed. Note: curve values taken visually from the original ~~(use original curve for calculations!)~~

Chapter 4

Methodology

Figure 4.2 shows a schematic representation of the methodology, where more representative ~~steps are~~ identified by numbers (from 1 to 8). This research is focus in non-hurricane data, with three main elements: data, temporal analysis with POT-PP, and spatial analysis for ~~to do spatial~~ interpolation and create return levels RL maps, for MRIs of 700, 1700, and 3000 years. ~~Steps 1, and 3 to 7, need to be done for each available station, see Figure 4.1.~~ With RL in each station, a continuous surface will be created, one for 700 years, next for 1700 years, and finally for 3000 years. An additional element, is the integration with existing hurricane maps to produce final maps, that will be used as input loads for infrastructure design, and will be part of the design standard.

Downscaling support parece ser un elemento adicional

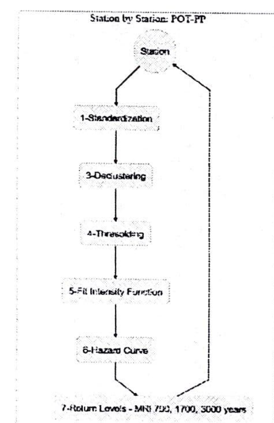


Figure 4.1: Iterative process in methodology

El step 2 no parece que fuera parte de la secuencia sino una verificación de los datos

Las Figuras se deben explicar en orden ó cambiar el orden

Sería bueno explicar un poco más la figura 4.2