Spatio-temporal analysis of extreme wind velocities for infrastructure design

Dissertation submitted in partial fulfillment of the requirements

for the Degree of Master of Science in Geospatial Technologies

Jan 2020

Alexys Herleym Rodríguez Avellaneda

alexyshr@gmail.com

https://github.com/alexyshr

Supervised by:

Prof. Dr. Edzer Pebesma

Institute for Geoinformatics

University of Münster - Germany

Co-supervised by:

Prof. Dr. Juan C. Reyes

Department of Civil and Environmental Engineering

Universidad de los Andes - Colombia

Co-supervised by:

Prof. Dr. Sara Ribero

Information Management School

Universidade Nova de Lisboa - Portugal

Declaration of Academic Integrity

I hereby confirm that this thesis on *Spatio-temporal analysis of extreme wind velocities for infrastructure design* is solely my own work and that I have used no sources or aids other than the ones stated.

All passages in my thesis for which other sources, including electronic media, have been used, be it direct quotes or content references, have been acknowledged as such and the sources cited.

February 20, 2020

_____

I agree to have my thesis checked in order to rule out potential similarities with other works and to have my thesis stored in a database for this purpose.

February 20, 2020

_____

Acknowledgements

I would like to thank Prof. Dr. Edzer Pebesma, Prof. Dr. Juan C. Reyes, and Prof. Dr. Sara Ribero, for supervising my work and spending their valuable time for discussions and feedback. It was really a huge advantage for me to have this always available support. It was a pleasure to work beside you. I want to thank Dr Adam Pintar and Engineer Juan David Sandoval for their support and contributions. My mother Ligia made possible all my achievements, because she was always there with love, support, and valuable advice and contributions. I am grateful with all my heart. Thanks to my daughter Nicolle Chaely for its love, support, and always pleasant company. I would like to thank the European Union -'Erasmus Mundus Grant', because their funding allow me to fulfill this dream to go further with my academic and professionals dreams.

I especially want to thank to Dr. Joaquín Huerta Guijarro, because he always was available to help and he was very friendly and receptive. Likewise, I want to thank some family members as Elsa Manrique, Barbara Avellaneda, and Kevin Martinez because they were an important source of motivation and support.

To all the beautiful people that shared with me different activities at the San Antonius Church of Muenster, with special mention of father Alejandro Serrano Palacios and choir friends.

Preface

Models of extreme values are used for designing against the effects of extreme events like earthquakes, winds, rainfall, floods of different types of physical processes, avoiding widespread destruction and loss of lives. This research presents a applied case of univariate extreme value analysis applied to wind velocities for infrastructure design, consequently, the main interest are probable future more extreme wind events that structures need to be able to resist.

This work in its theoretical and methodological component was directed by ASCE7-16 Engineers (2017) considering that output products will be used to update the chapter B.6, wind forces, of the Colombian earthquake resistant standard - NSR-10, maintained by the Colombian Association of Seismic Engineering - AIS by its Spanish acronym. ASCE7-16, defines four risk categories, which implies the use of different wind loads (represented in wind extreme values for different mean recurrence intervals) for structures that belong to each category, 3000 years of MRI for risk IV, 1700 years for risk III, and 700 years for risk II and I.

This research has a particularly new situation regarding to the input data, and it is that not only time series of field measurements from meteorological stations are used (IDEAM data source), but also post-processed information coming from the Integrated Surface Database - ISD (USA database based on IDEAM data source), and forecast reanalysis data from ERA5. This condition demanded a comparison of the different data sources in order to verify the feasibility of using ERA5 and ISD, with a previous process of standardization of wind velocities to reach the needed requirement of 3-s wind gust speed, 10 meters anemometers high and open space condition.

At each station the used method Peaks Over Threshold - Poisson Process, required to identify all the non-thunderstorm events in the non-hurricane dataset through a process of declustering, choose a suitable threshold level to leave for the analysis only the most extreme values available, and then fit to the data a Gumbel extreme value distribution using maximum likelihood to find optimal parameters with the best goodness of fit. With the fitted model, it was possible to calculate return levels for required mean return intervals. Next, a process of spatial interpolation was done using Kriging, what allowed to have three continuous maps for the whole study area. Main interest writing this document, is help to readers to enter speedily with the current details around wind extreme analysis.

Table of Contents

List of Tables

List of Figures

Abstract

For the input non-hurricane, non tornadic data in each available station of the study area (field measurement of forecast models), this research calculate extreme winds or return levels with three different mean recurrence intervals - MRI, 700, 1700, and 3000 years, with a change of being equaled or exceeded only one time in the corresponding MRI period. Then, continuous maps of wind extreme velocities are interpolated to cover the study area, which are mixed with existing wind extreme hurricane studies to be used as input loads for infrastructure design.

Spatio-temporal analysis of historical wind data for infrastructure design, namely, − from wind time series represented in forecast models over rectangular areas or pixels with a virtual station at its center, or field measurements at weather stations in specific coordinates around the study area −, calculate wind extreme magnitudes to be used as design loads of structures of different risk categories (bridges, houses, buildings, hospitals, etc), requires the use of statistical extreme value analysis methodologies to create maps with different mean recurrence intervals (MRI), − short ones for less risky/important structures, and long ones for highly important structures.

Method used to calculate the return levels at each station the Peaks Over Threshold - POT, using a non-homogeneous, bi-dimensional Poisson Process described, recommended by Engineers (2017), and developed and implemented in Pintar, Simiu, Lombardo, & Levitan (2015). To interpolate maps a geostatistical procedure using Kriging was implemented, considering the model with the best goodness of fit from model parameters comparison.

List of Acronyms

pdf	Probability Distribution Function
cdf	Cumulative Distribution Function
ppf	Percent Point Function (Quantile)
hf	Hazard Function
P_e	Annual Exceedance Probability
MRI	Mean Return Interval or Return Period
P_n	Compound Exceedance Probability
IDEAM	Institute of Hydrology, Meteorology and Environmental Studies
ECMWF	European Center for Medium-Range Weather Forecasts
ERA5	ECMWF climate reanalysis dataset
GEVD	Generalized Extreme Value Distribution (EVD, GEV)
EVD	Extreme Value Distribution (GEVD, GEV)
GEV	Generalized Extreme Value Distribution (GEVD, EVD)
GPD	Generalized Pareto Distribution
ISD	Integrated Surface Database
AIS	Seismic Engineering Association
NSR	Seismic Resistant Norm
SEI	Structural Engineering Institute
ASCE	American Society of Civil Engineers
ASCE7-16	ASCE/SEI Design Loads Standard
NOAA	National Oceanic and Atmospheric Administration
NetCDF	Network Common Data Form
NCEI	NOAA's National Centers for Environmental Information
EDA	Exploratory Data Analysis
POT	Peaks Over Threshold
PP	Poisson Process
RMSE	Root Mean Squared Error
PACF	Partial Autocorrelation Function
ACF	Autocorrelation Function
SQL	Structured Query Language
IDW	Inverse Distance Weighted
WGS84	World Geodetic System 1984
RL	Return Level

Chapter 1

Introduction

This research aims to create non-hurricane non-tornadic maps of extreme wind speeds for *three specific recurrence intervals* (700, 1700, and 3000 years) covering the Colombian territory. These maps will be combined with existing hurricane wind speed studies, to be used as input loads due to wind for infrastructure design.

For each station with wind speeds time histories in the input data, extreme wind speed corresponding to each recurrence interval are calculated using a *Peaks Over Threshold* onwards *POT* extreme value model, then wind velocities with the same recurrence interval are *spatially interpolated* to generate continuous maps for the whole study area.

A wind speed linked to a *mean recurrence interval - MRI* of *N-years* (N-years return value or return period) is interpreted as the highest probable wind speed along the period of N-years. The annual probability of equal or exceed that wind speed is 1/N. The annual exceedance probability for all velocity values in 700-years output map will be 1/700, for the 1700-years map will be 1/1700, and 1/3000 for the 3000-years final map.

There are different methods to model extreme value data, among them are a) sample maxima using a *Generalized Extreme Value Distribution* onwards *GEVD* (traditional method), b) POT using a *Generalized Pareto Distribution* onwards *GPD*, c) POT using a two-dimensional Poisson Process, that can be homogeneous, non-homogeneous, stationary, and non-stationary (originally know as *Point Process* approach), and d) POT Poisson-GPD. Following Pintar et al. (2015) in this research a *POT using a non-homogeneous non-stationary two-dimensional Poisson proces* was selected, despite there is no R package available to apply this approach.

1.1 Background

To design one structure, the horizontal forces wind and earthquake play an starring role. For the study area, Colombia, initially the wind force was considered with the decree 1984 as a fixed velocity $100\frac{Km}{h}$, later a continuous map with a return period of 50 years was included in the official design standard of the time (NSR-98), then, with the update to NSR-10, an

additional map with return period of 700 yeas was included.

In the context of this study, extreme wind analysis is concerned with statistical methods applied to very high values of wind as random variable in a stochastic process, to allow statistical inference from historical data, namely, assess from the ordered sample of wind velocities, the probability of wind events that are more extreme than the ones previously observed and included in the mentioned input sample. Classical reference in this matter is Coles (2001), where a detailed study is done about classical extreme value theory and models and threshold models. There are four main approaches to deal with extreme value analysis: - sample maxima associated to a Generalized Extreme Value Distribution - GEV, - exceedances over threshold associated to a Generalized Pareto Distribution - GPD, - the Poisson-GPD, an homogeneous Poisson process for the number of exceedances and a GPD for the excess values, and the exceedances over threshold associated to a non- homogeneous bi-dimensional Poisson process, a Point process approach also known as Peaks Over Threshold - POT - Poisson process. Main details will be discussed here for each method, but as the last one is recommended in Asce2017, a more indeed explanation will be provided in for POT-Poisson Process.

1.1.1 Sample maxima

To work with random variables of sample maximum values, the used probability distribution function *pdf* is the GEV

$$H(y) = \exp\left\{-\left(1+\xi\frac{y-\mu}{\psi}\right)_+^{-\frac{1}{\xi}}\right\},$$

($y+ = max(y,0)$) where μ is the location parameter, $\psi > 0$ is a scale parameter, and ξ is a shape parameter. GEV can be seen as the integration in the same *psf* of the Gumbel distribution (limit $\xi \to 0$), Fréchet distribution ($\xi > 0$), and Weibull distribution ($\xi < 0$).

1.1.2 Exceedances over threshold

If the researcher needs to work only with extreme values above an specific threshold, Pickands (1971) showed that the GEV has a GPD approximation where shape ξ parameter in previous equation is the same parameter for next equation for GPD,

$$G(y,\sigma,\xi) = 1-\left(1+\xi\frac{y}{\sigma}\right)_+^{-\frac{1}{\xi}},$$

Poisson-GPD If a rescale of the variable indexes above the threshold is performed, then the exceedances over threshold approach can be seen as a point process, namely, an homogeneous Poisson Process where:

1. The number of exceedances above the threshold has a Poisson distribution with mean λ

2. The excess values follow a GPD with $N \leq 1$

Its cumulative distribution function *cdf* is

$$F(y) = \exp\left\{-\lambda\left(1+\xi\frac{y-\mu}{\sigma}\right)_{+}^{-\frac{1}{\xi}}\right\},$$

1.2 Research Aim and Objectives

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 risk categories, and covering any place in the whole study area.

Specific objectives are:

1. Analyze and compare three different sources of historical wind time series, to select and use the best data source (or combination or sources) for research, based on objective criteria, for instance similitude, completeness, coverage, etcetera, to achieve this way a formal support for the decision made in this regard.
2. Select and apply an suitable extreme value analysis method that allows to fulfill wind load requirements defined for the respective authority in the study area
3. Estimate extreme wind values for the stations in the selected input data source, for three MRI (700, 1700, 3000 years), considering non-hurricane studies.
4. Generate continuous maps for MRIs 700, 1700, and 3000 years, using the most suitable spatial interpolation technique, considering the specific characteristics of the input data and advantages and disadvantages of the selected methods
5. Combine output maps from non-hurricane analysis, with existing hurricane studies to allow the inclusion of the research study in the NSR-10 norm.

1.3 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?

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?

1.4 Thesis Document Structure

Chapter 2

Data

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`.

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^{o} (33 kilómeters), this is equally spaced cells every 0.25 degrees

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:

- As just mentioned, low quantity of IDEAM stations
- 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.
- 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.

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

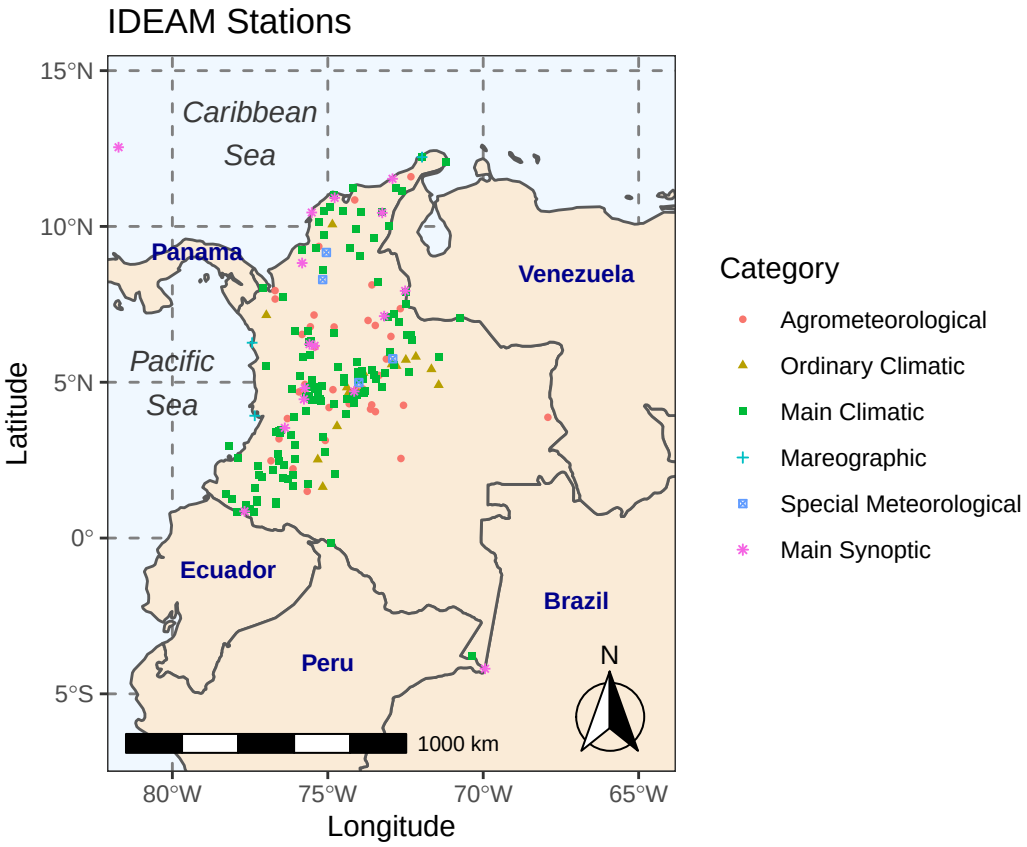

Figure 2.1: IDEAM Stations. Colombia

Following, the time series, autocorrelation function, and partial autocorrelation function, for IDEAM station “21205791” will be displayed.

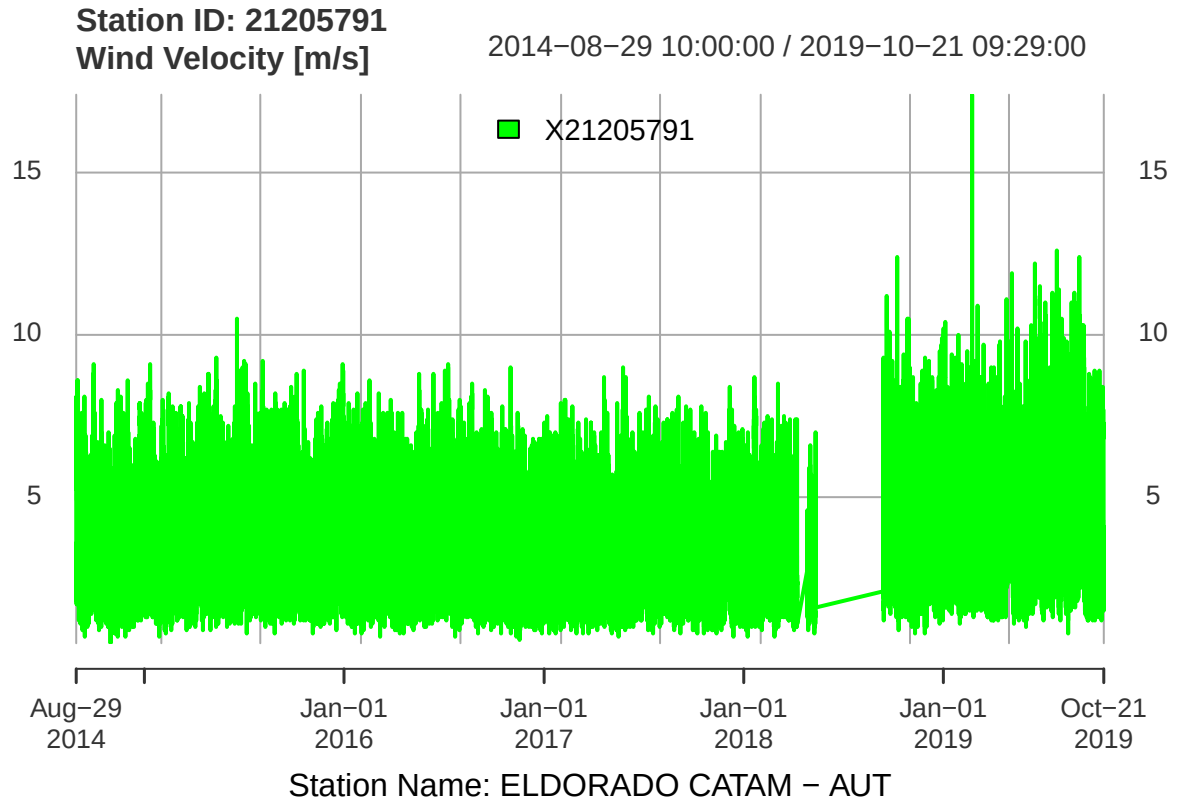

Figure 2.2: IDEAM Station - Time Series

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.

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

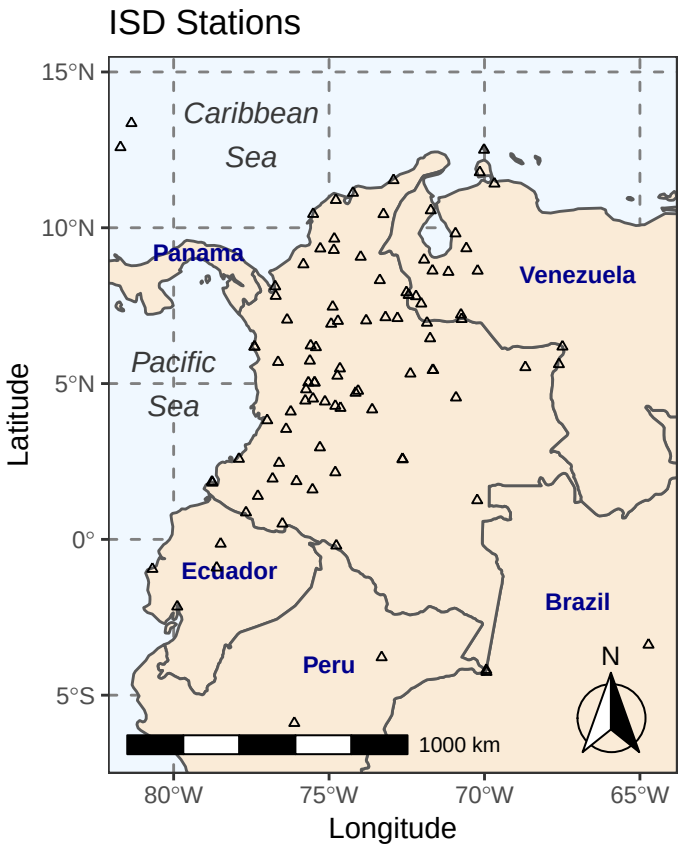

Figure 2.3: ISD Stations. Colombia and surroundings

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

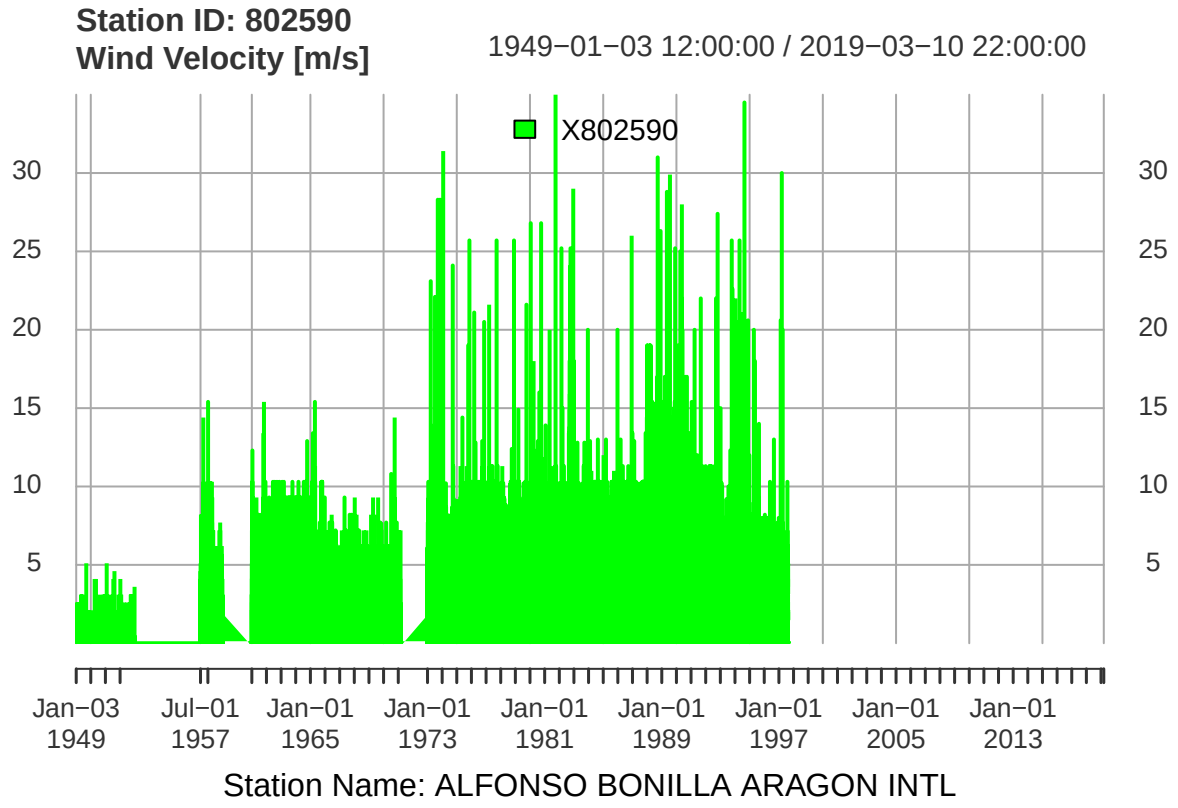

Figure 2.4: ISD Station - Time Series

2.3 ERA5

ERA5 is forecast reanalysis data processed by the *European Center 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).

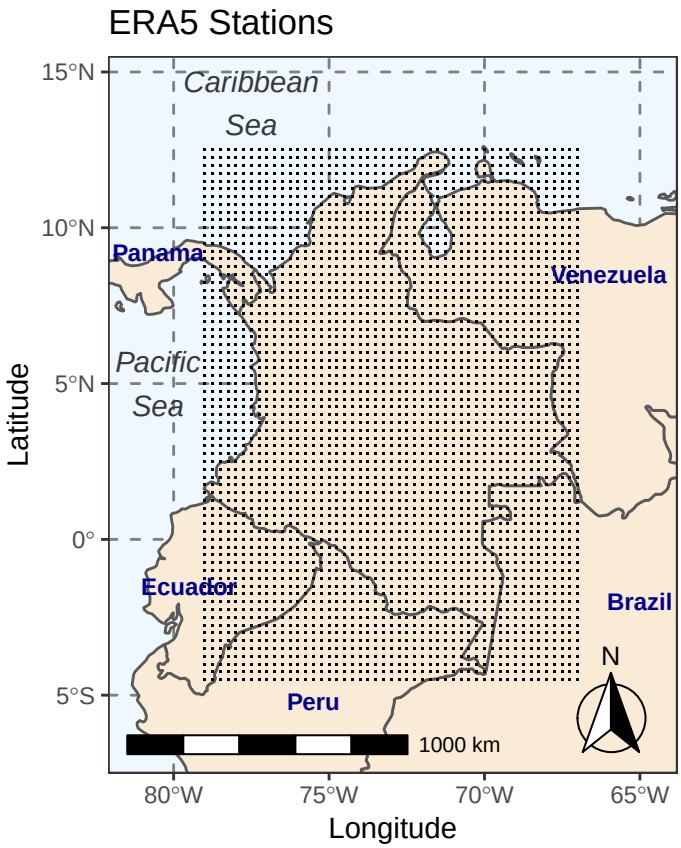

Figure 2.5: ERA5 Stations (cells centers). Colombia bounding box

2.4 Data Download and Organization

2.5 Data Standardization

Analysis of extreme wind speeds requires data standardization as initial step. All input data must be standardized to represent three important conditions: a) anemometer height of 10 meters, b) open space roughness, and c) averaging time of 3-seconds wind gust. Data for analysis must represent 3-s peak wind speeds 10 meters height above the surface, in open terrain. * 10 meters anemometer height * Open space terrain roughness * 3-s gust averaging time

Chapter 3

Theoretical Framework

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.

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

Equation (3.1) is the Gumbel *pdf*.

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

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 3.1 shows *pdf* with location (μ) = 100 and scale (β) = 40, using equation (3.1).

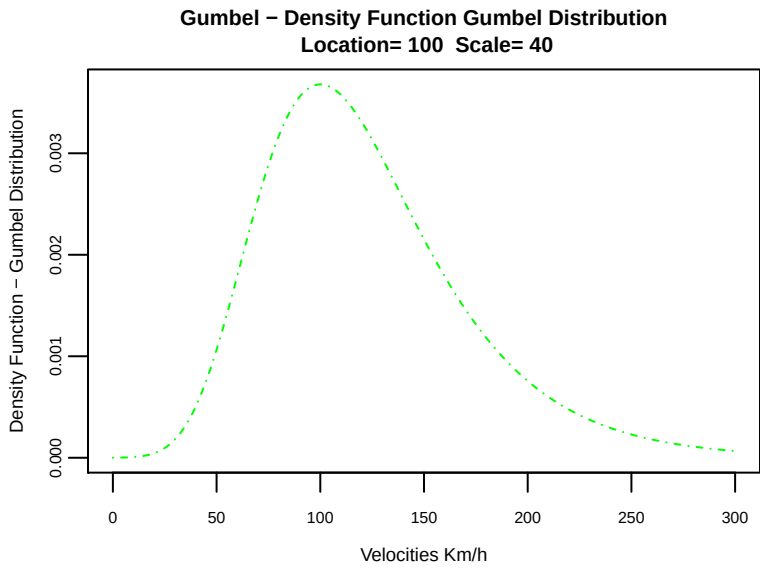

Figure 3.1: Gumbel pdf

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

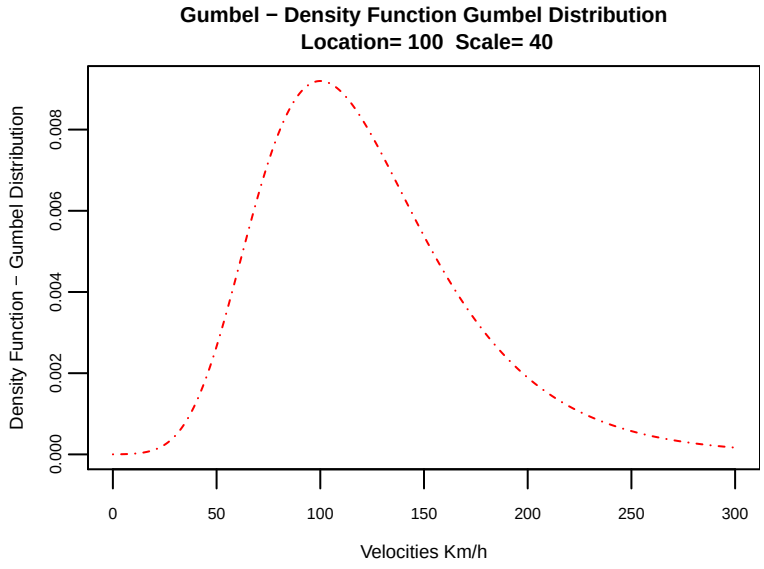

Figure 3.2: Gumbel pdf - dgumbel function

3.1.2 Cumulative Distribution Function - *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 (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`.

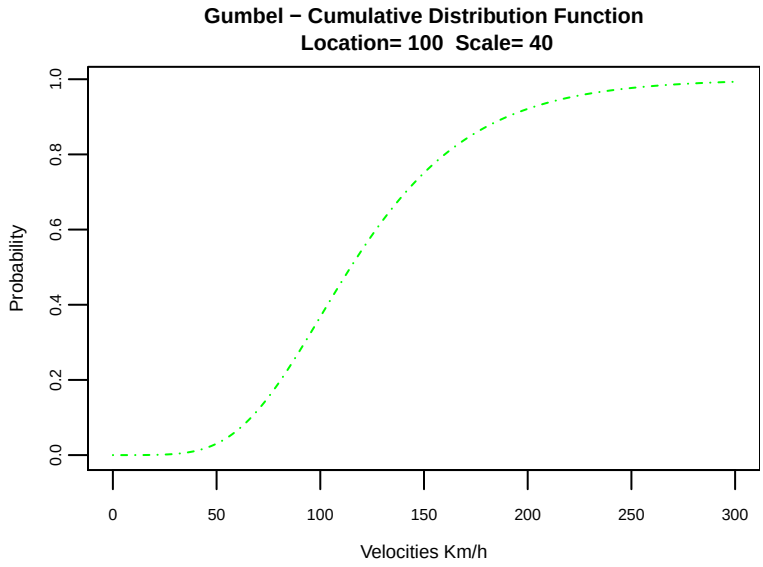

Figure 3.3: Gumbel cdf

3.1.3 Percent Point Function - *ppf*

Ppf 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`.

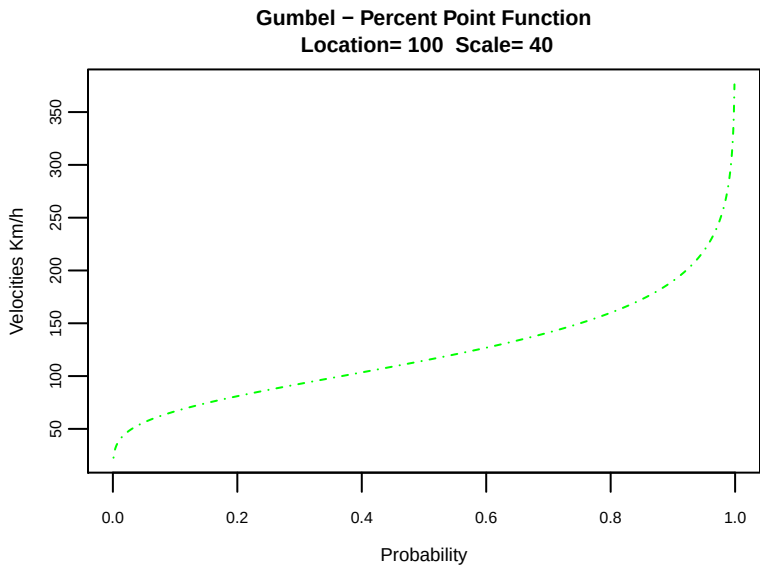

Figure 3.4: Gumbel ppf

3.1.4 Hazard Function - *hf*

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).

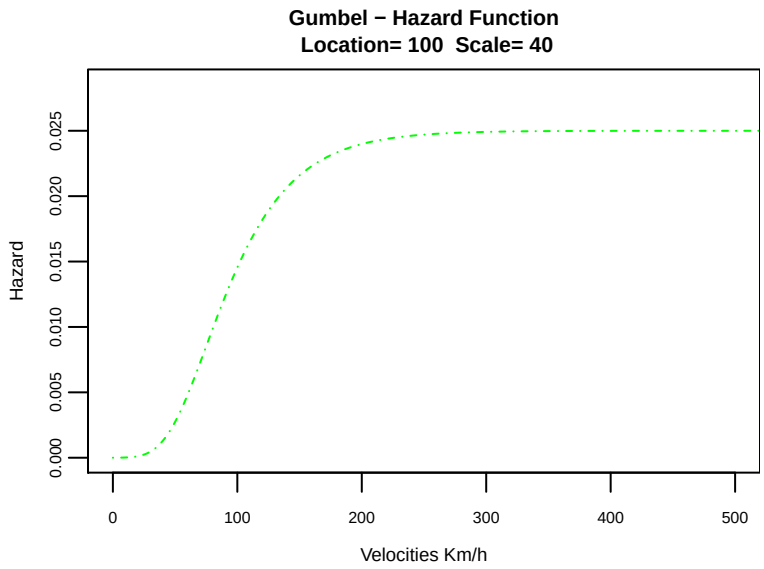

Figure 3.5: Gumbel hf

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

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

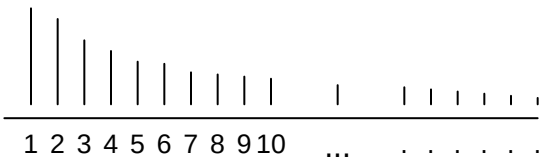

Figure 3.6: Sorted Winds by Magnitude - wind simulation database

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

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.

3.2.2 Return Period - 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 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.

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

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).

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

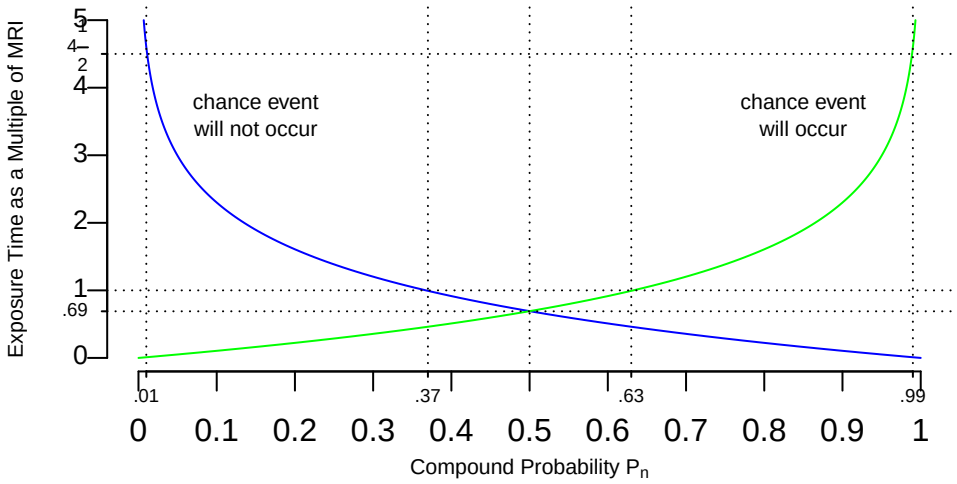

Figure 3.7: Compound Probability

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_e 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

(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)

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 test. 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-GPD

Short description of POT-GPD (this section need to be complemented)

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

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

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.

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).

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

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),

$$\frac{1}{\psi_\mathrm{t}}\left(1+\zeta_\mathrm{t}\frac{\mathrm{y}-\omega_\mathrm{t}}{\psi_\mathrm{t}}\right)_{+}^{-\frac{1}{\zeta_\mathrm{t}}-1} \tag{3.6}$$

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.

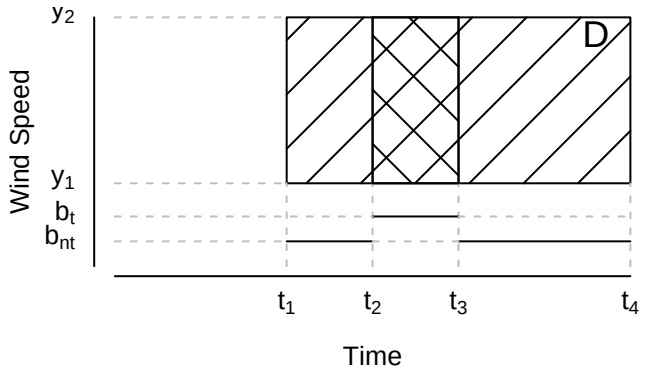

Figure 3.8: Domain off the Poisson Process - PP

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.

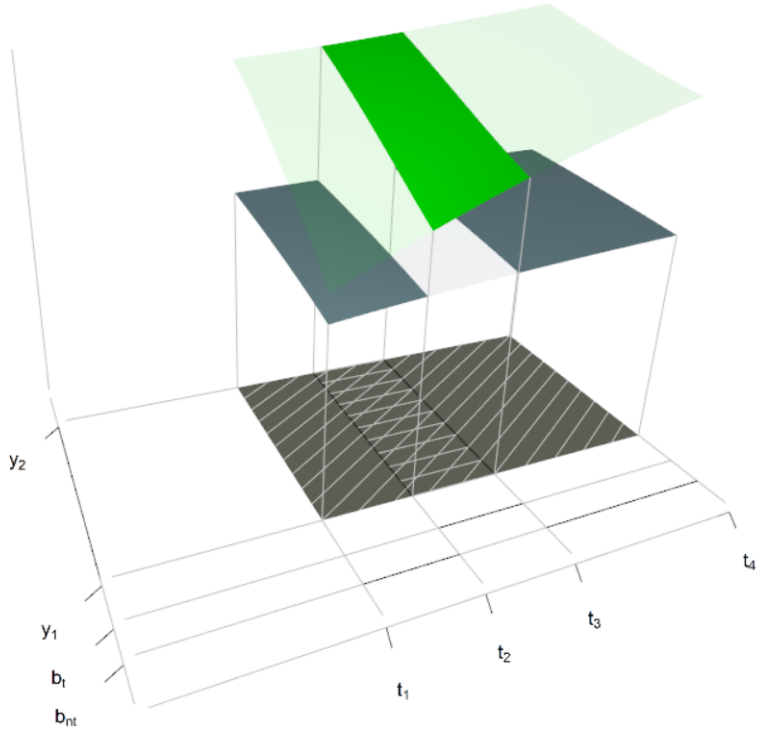

Figure 3.9: Volume under surfaces represents the mean of PP

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* $= \zeta$, the selected vector of parameters η are the $\hat{\eta} = (\hat{\psi}, \hat{\omega}, \hat{\zeta})$ values that maximizes next function

$$\mathrm{L}(\eta) = \left(\prod_{\mathrm{i}=1}^{\mathrm{I}} \lambda\left(\mathrm{y_i}, \mathrm{t_i}\right) \right) \exp\left\{ - \int_{\mathrm{D}} \lambda\left(\mathrm{y}, \mathrm{t}\right) \mathrm{dy}\, \mathrm{dt} \right\} \tag{3.7}$$

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}. $Y_N = quantile(y, p = P_{ne}) = G(y, p = P_{ne}) = ppf(y, p = P_{ne})$. 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,

$$\int_{\mathrm{Y_N}}^{\infty}\int_{0}^{1}\lambda\,(\mathrm{y},\mathrm{t})\,\mathrm{dydt} = \mathrm{A_t}\int_{\mathrm{Y_N}}^{\infty}\lambda_{\mathrm{t}}\,(\mathrm{y})\;\mathrm{dy} + \mathrm{A_{nt}}\int_{\mathrm{Y_N}}^{\infty}\lambda_{\mathrm{nt}}\,(\mathrm{y})\;\mathrm{dy} = \frac{1}{\mathrm{N}} \tag{3.8}$$

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 statistics. 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).

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

The procedure to choose the best thresholds pairs based in W transformation, is described in methodology, section 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 is mainly based in *minimum design loads and associated criteria for buildings and other structures - ASCE7-16* norm, see Engineers (2017). Under these circumstances, ASCE7-16 defines the minimum requirements of the research products. Especially the chapter C26 - "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: 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.

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.

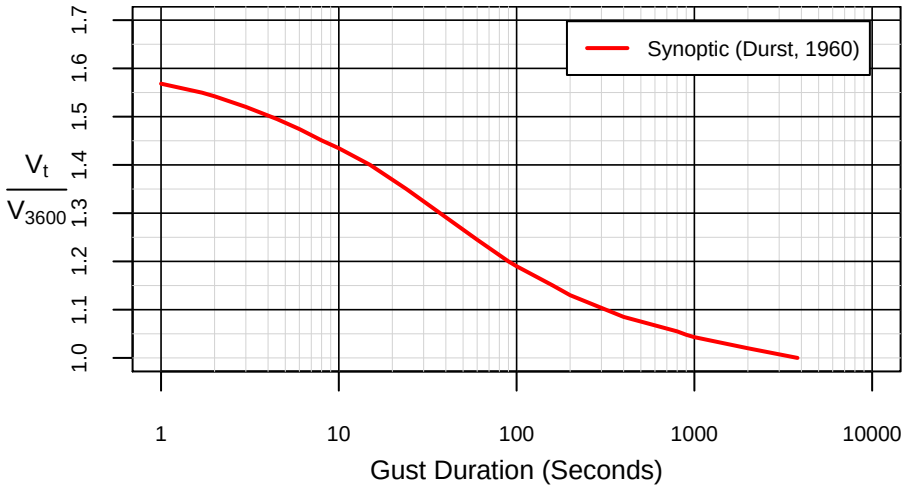

Figure 3.10: Maximum speeds averaged over t (sec), to hourly mean speed. Note: curve values taken visually from the original (use original curve for calculations!)

Chapter 4

Methodology

Figure 4.1 shows a graphic representation of the methodology. This research is focus in non-hurricane data, with three main elements: - data, - temporal analysis with a POT-PP, and - spatial analysis with probabilistic and deterministic methods to do spatial interpolation and create return levels (wind velocities) maps, for MRI of 700, 1700, and 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.

More representative and important steps of the methodology are identified by numbers in figure 4.1, 1) standardization, 2) de-clustering, 3) thresholding, 4) fit intensity function, 5) hazard curve, 6) return levels, and 7) spatial interpolation. Steps 1 to 6, need to be done for each available station to get extreme wind velocities (return levels - RL) for MRI. With RL in each station, a continuous surface will be created, one for 700 years, next for 1700 years and finally for 3000 years. Figure 4.2 schematize the iterative process in the methodology.

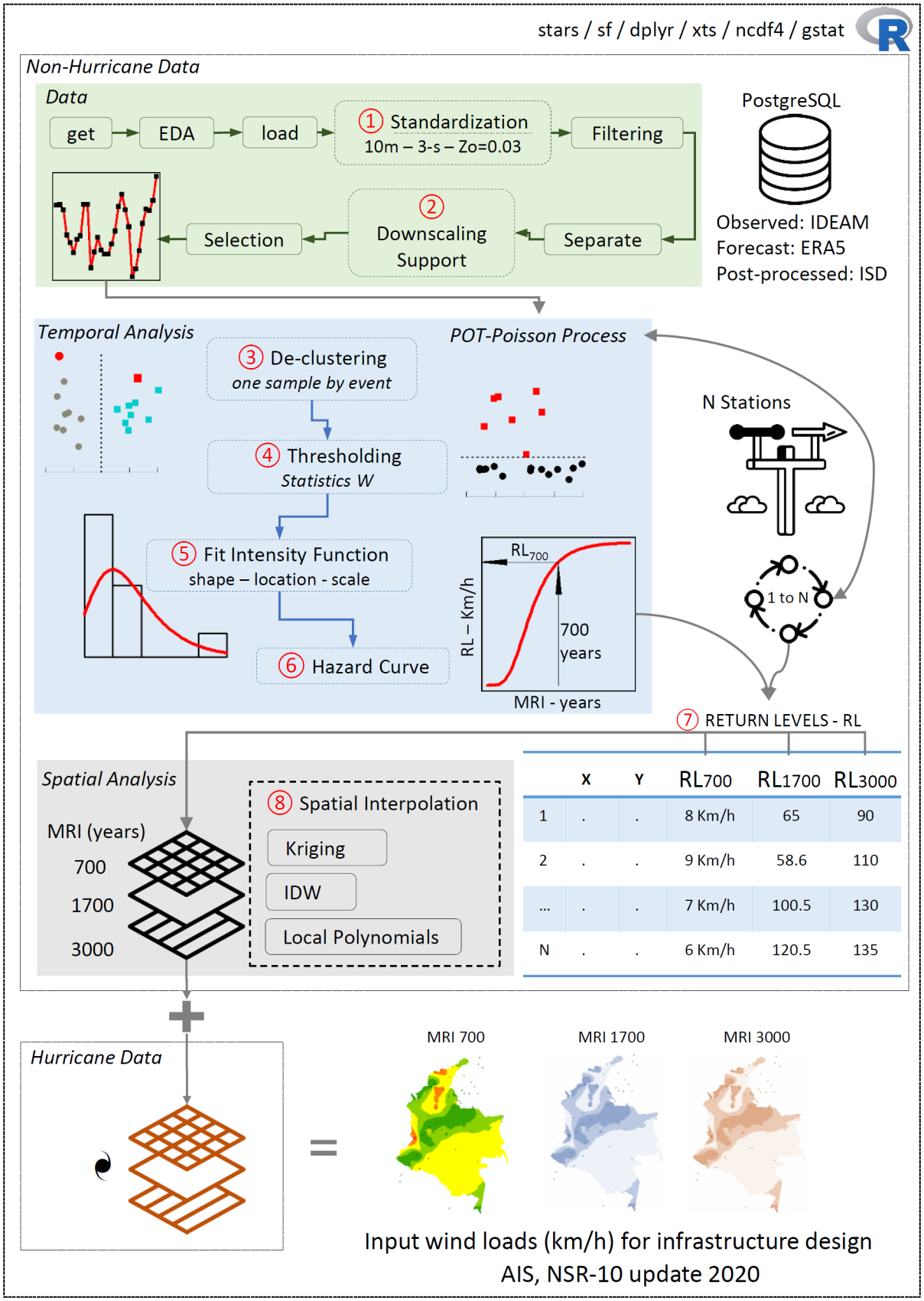

Figure 4.1: Methodology

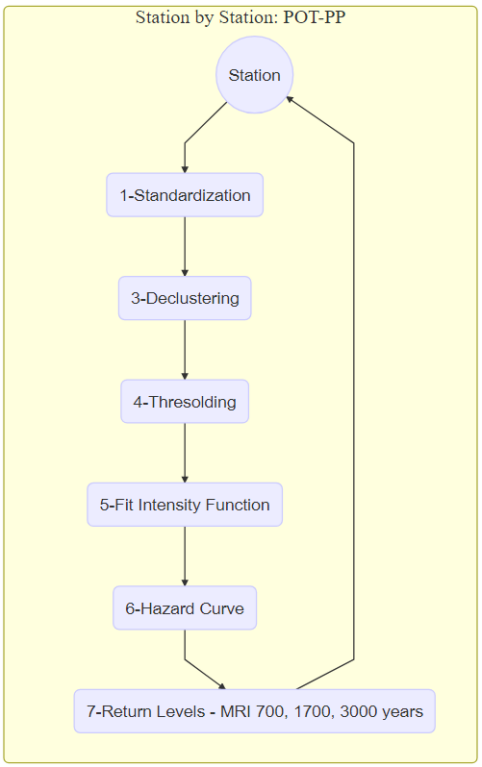

Figure 4.2: Iterative process in methodology

4.1 Data Standardization

Parallel to the standardization activity described below (3-s gust, roughness, and 10 meters anemometer height), it is also important to consider for all stations involved in the analysis:

- *Separating*: As far as possible, identify each record of the time series, as thunderstorm (t) or non-thunderstorm (nt)
- *Filtering*: Remove wind speeds above $200\frac{Km}{h}$ and data pertaining to hurricane events, because the procedure with hurricane requires a different approach and need to be done independently
- Downscaling approach: As it happens in this study, where it is intended to complement the local/regional wind analysis, with data from ISD (output data of a model for extreme winds), and ERA5 reanalysis dataset (large scale forecast data), it is required to probe by means of *comparisons* (exploratory data analysis and statistical measures) that modeled or forecast data are suitable to complement the study.

4.1.1 Anemometer height - 10 m

According to the protocol for field data collection and location of methodological stations - IDEAM (2005), the anemometer (wind sensor) in installed always to a fixed height of 10

meters from the surface, as is shown in figure 4.3, ergo, no height correction is needed.

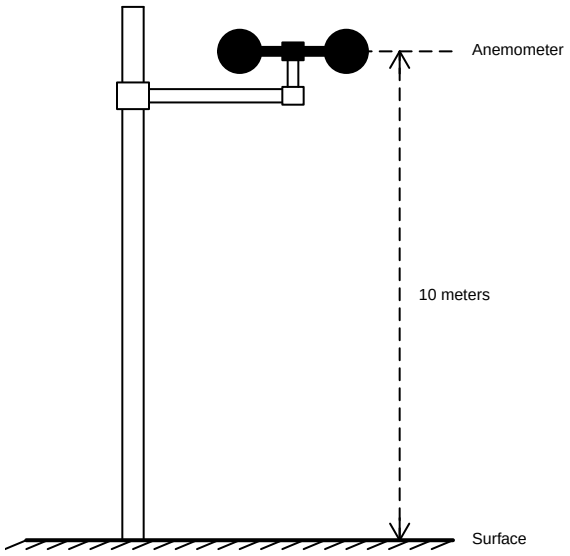

Figure 4.3: Anemometer height - 10 m

4.1.2 Surface Roughness at Open Terrain (0.03 m)

Due to the effects that the terrain has on wind speed, a correction should be applied if the station is located in a geographical space considered "not open terrain". When terrain is open, the roughness corresponds to 0.03 meters. There are some alternative methodologies to calculate the roughness, Masters, Vickery, Bacon, & Rappaport (2010) uses the station data, but the separation of the measurements should not exceed one minute, something difficult to obtain, and Lettau (1969) uses an empirical equation that is recommended in Engineers (2017) (page 743, equation C26.7-1), which was used here,

$$Roughness = z_0 = 0.5 * H_{ob} * \frac{S_{ob}}{A_{ob}}$$

Where H_{ob} is the average height of the obstacles, S_{ob} is the average vertical area perpendicular to the wind of the obstacles, and A_{ob} is the average area of the terrain occupied by each obstruction. Then, the empirical exponent α, gradient height z_g, and exposure coefficient K_z, corresponding to equations C26.10-3, C26.10-4, and C26.10-1.si of Engineers (2017), are used to calculate the correction factor $F_{exposition}$, verifying that z_0 units are in meters.

$$\alpha = 5.65 * z_0^{-0.133}$$

$$z_g = 450 * z_0^{0.125}$$

$$K_z = 2.01 * \left(\frac{z}{z_g}\right)$$

$$F_{exposition} = \frac{0.951434}{K_z}$$

Following NIST (2012), calculation of roughness need to be weighted according to the predominance of wind magnitude in eight directions (north, south, east, west, north-east, north-west, south-east, and south-west), see figure 4.4, using a detailed aerial photo or satellite image inside a radiuos of 800 meters around the station location, as shown in figure **??**, with south direction highlighted.

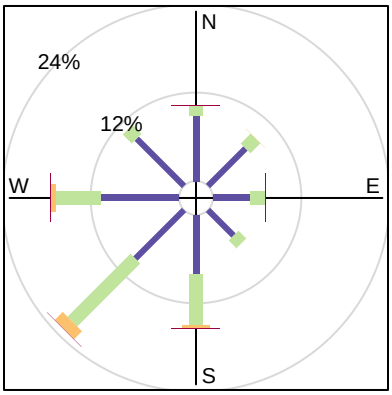

Figure 4.4: Wind rose with wind percentages in eight directions, for a generic station

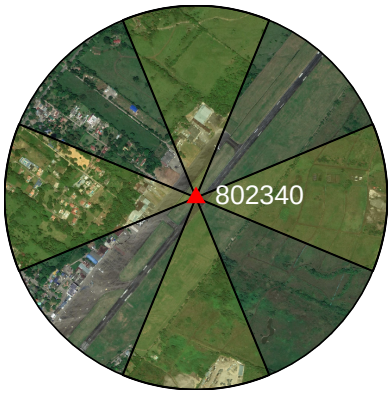

Figure 4.5: Digital imagery for 'Vanguardia' ISD station (USAF:802340), located in Villavicencio airport. with four (south, north, east, and west) 45 degree sectors highlighted. Radious of the circular zone is 800 meters

Figure ?? shows extreme conditions for roughness, open space in left image (ISD Station 804070), closed space in center image (ISD Station 803000), and a typical example where Lettau procedure is needed. Lettau equation need to be applied to each direction and then the final z_o value is the weighted average, using historical wind percentage. See figure 4.7 showing the strokes made to calculate the different areas for two Colombian stations. Information about wind percentage per direction at each station were obtained from IDEAM (1999).

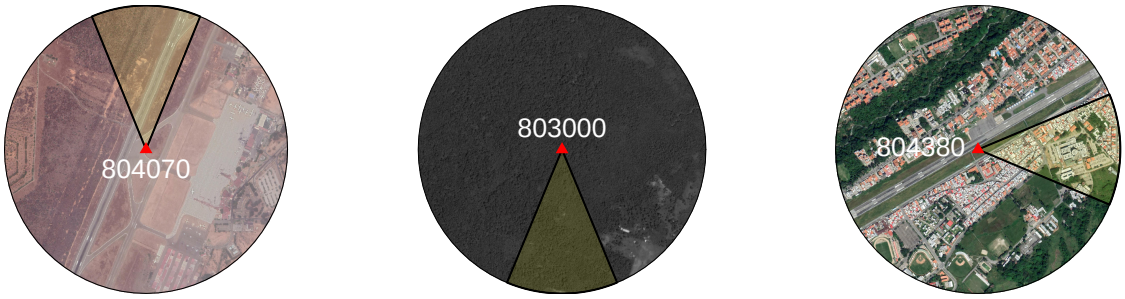

Figure 4.6: Roughness values: 0.03 for open space (left), 0.1 for closed space (center), and areas where Lettau equation is needed because roughness is different in each direction (right).

Figure 4.7: Lettau calculation. In red the area occupied by the obstacles, and in blue the perpendicular area. Source Triana (2019)

4.1.3 Averaging Time 3-s Gust

To transform hourly mean wind velocity V_{3600}, to 3-s gust velocity V_3, Engineers (2017) recommends to use C. S. Durst (1960). See Wind Loads Requirements. As the axis x represents duration t of the gust, what is done is to look there for the value 3 seconds, and read the corresponding gust factor $\frac{V_t}{V_{3600}}$, this is, the value in the axis y, then

$$V_t = V_{3\,seconds} = (gustfactor) * V_{3600\,seconds}$$

It is valid only for open terrain conditions. Durst curve shows in axis y the gust factor $\frac{V_t}{V_{3600}}$,

a ration 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.

4.2 Peaks Over Threshold - Poisson Process (POT-PP)

Similar to how the adjustment of statistical data to a normal distribution works in order to make inferences considering deviations from the mean, here only some part of the data (those that are extreme - over a high threshold - POT), need to be fitted to a PP considering extreme deviations from the mean. While in the first case (normal distribution) the inferences are for events similar to the samples, in this case, when working with extreme value theory, the inferences will be for more extreme events than any previously observed or measured. In the theoretical framework section are described the main elements of POT-PP.

In summary, POT means only to work with extreme values, and PP means to adjust data to a *pdf*, which depends on an intensity function $\lambda(t, y)$, where t is time, y is wind extreme velocity. As is shown in figure 3.8, in a POT-PP approach with domain D, all the observations follow a Poisson distribution with mean $\int_D \lambda(t, y)\, dt\, dy$. Main advantage of POT-PP is that it is designed to consider storm and not-storm events independently (for each disjoint sub-domain D_1 or D_2 inside D, the observations in D_1 or D_2 are independent random variables), but in the end use them both for the inferences,

$$\mathrm{pdf} = \mathrm{f(t, y|}\eta) = \frac{\lambda(\mathrm{t, y})}{\int_{\mathrm{D}} \lambda(\mathrm{t, y})\, \mathrm{dt\, dy}} \tag{4.1}$$

4.2.1 De-clustering

To make the assumptions of PP more justifiable, it is important to have only one sample per event, the highest one. For instance, if a hypothetical storm started at 11:30 in the morning and ended at 12:30 in the afternoon, and the time series for that event has thirty wind measurements (one each two minutes), it is necessary to leave only the stronger or maximum value, and this process is called de-clustering (see Figure 4.8). POT-PP defines that all the adjacent observations separated by six hours (6) or less in the case of thunderstorm events, and four (4) days or less, in the case of non-thunderstorm events belong to the same cluster.

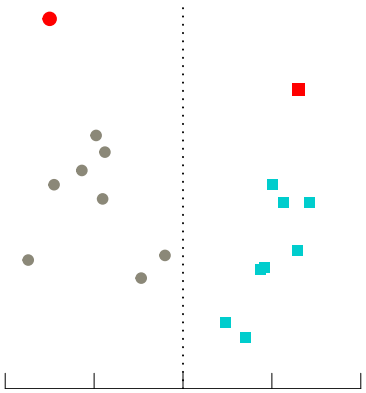

Figure 4.8: De-clustering in PP. Two thunderstorm clusters are shown. Separation between adjacent observations inside the clusters are always equal or less than six hours. Distance between the last event in the first cluster and the first event in the second cluster is larger than six hours. Only red samples are used to fit the PP, but in addition a POT (thresholding) process is needed

4.2.2 Thresholding

As the POT model requires to work only with the most extreme values in the time series, it is necessary to select a threshold to filter out small values. Selection of threshold value imply two effects in the model. Bias is high when a low threshold is selected (many exceedances) because the asymptotic support is weak. Opposite situation happens for high thresholds where variance is potentially high, so according to Davison & Smith (1990), it is needed to select a threshold value, consistent with model structure.

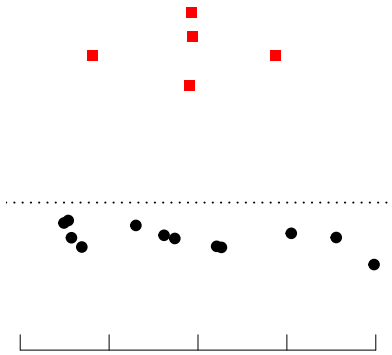

Figure 4.9: POT - Thresholding

Selection of the thresholds pairs, one for thunderstorm, and one for non-thunderstorm, is based in W transformation described in threshold selection section. W-statistic is done comparing the ordered empirical result of applying $W = -log(1 - U)$ to the data, axis y in figure 4.10, with the theoretical quantiles of an exponential variable with uniform distribution between 0 and 1, axis x in same figure. W-statistic is the highest vertical distance

between the 45º line and the points in the graphic. The best thresholds pairs returns the minimum value for W-statistics after testing, in an iterative process, with many threshold pairs combinations.

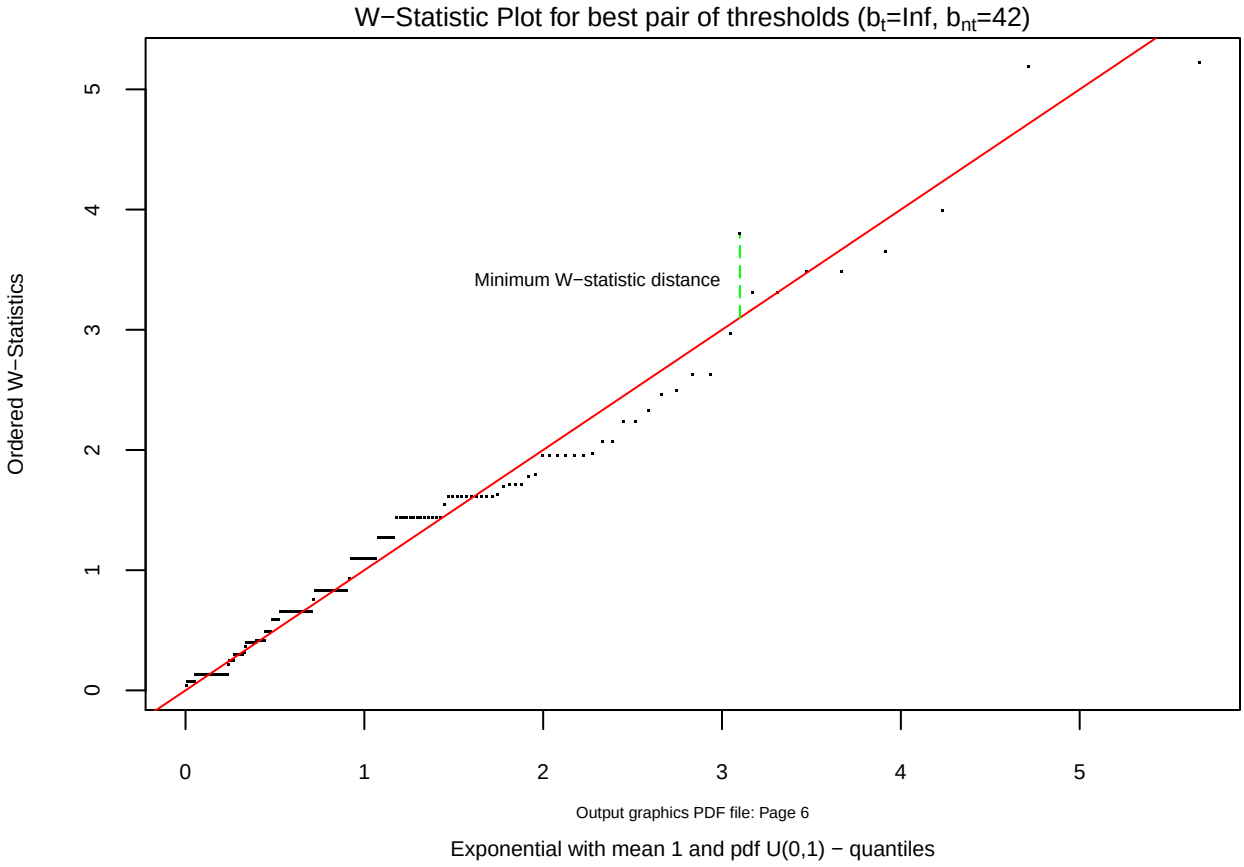

Figure 4.10: POT - Thresholding

4.2.3 Exclude no-data periods

PP requires to remove long periods of time when stations were not recording or failing. Proposed time in Pintar et al. (2015) is 180 days, namely, to remove all the gaps from the time series larger than six months.

4.2.4 Fit Intensity Function

Probability density function *pdf*, and cumulative distribution function *cdf*, of the PP, depend of the intensity function, and are shown in equation (4.1), and equation (3.9), respectively.

To facilitate the estimation of the parameters for the PP intensity function, parameter $shape = \zeta_t$ is taken to be zero in equation (3.6), then doing the limit, the resulting intensity

function is the same as the the GEV type I or Gumbel distribution,

$$\frac{1}{\psi_{\mathrm{t}}} \exp\left\{\frac{-(\mathrm{y}-\omega_{\mathrm{t}})}{\psi_{\mathrm{t}}}\right\} \tag{4.2}$$

In this study, the used intensity functions are shown in next equation (4.3).

$$\lambda\,(\mathrm{y},\mathrm{t}) \begin{cases} \dfrac{1}{\psi_s} \exp\left(\dfrac{-(y-\omega_s)}{\psi_s}\right), & for\ t\ in\ thunderstorm\ period \\ \dfrac{1}{\psi_{nt}} \exp\left(\dfrac{-(y-\omega_{nt})}{\psi_{nt}}\right), & for\ t\ in\ nonthunderstorm\ period \end{cases} \tag{4.3}$$

As is shown in 4.11, the fitting process involve finding the best group of parameters of the intensity function, in such a way that the red curve (*pdf* of the PP, based in intensity function) be as tight as possible to the shape of the data histogram. As is described in POT-PP, optimal parameters to do the fitting process of the intensity function are calculated using *maximum likelihood*.

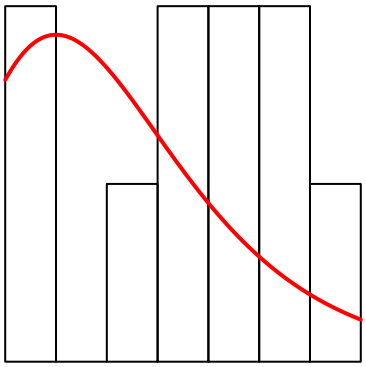

Figure 4.11: POT - PP intensity function fitting process

4.2.5 Hazard Curve - Return Levels - RL

If equation (3.8), Y_N is solved using estimated parameters of the intensity function, and a hazard curve is constructed as shown in figure 4.12, where axis x represents annual exceedance probability $P_e = \frac{1}{N}$, and axis y represents the RL Y_N for the corresponding N-years return period, then it will be possible to have the extreme return wind velocity level for any given return period going from axis x to axis y through the curve.

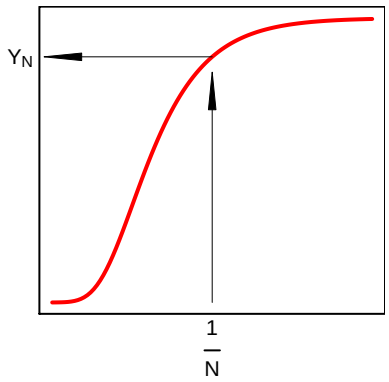

Figure 4.12: POT - PP fitting process

Two alternatives approaches for RL

There is an equation that allows direct calculations of return levels, and also it is possible to use the quantile function of Gumbel when shape parameter equals to zero, but it is important to emphasize that equation (4.4), and the use of Gumbel quantile function for RL calculations, is only valid when the analysis of POT-PP includes only one type of event (thunderstorm or non-thunderstorm), and the average estimated duration time of the event in a year is considered to be one (independent of the units in which time is processed), namely, the values for parameters A_t or A_{nt} of equation (3.8) are equal to one.

Instead of solving equation (3.8), next equation (4.4) can be used replacing directly the PP parameters and the N return periods to create the hazard curve and get RL.

$$\mathrm{Y_N} = \frac{\psi}{\zeta}\left[-\log\left(\frac{\mathrm{N}-1}{\mathrm{N}}\right)\right]^{-\zeta} - \frac{\psi}{\zeta} + \omega \tag{4.4}$$

As for this research $\zeta = 0$, return levels Y_N can be calculated using the Gumbel quantile function, but using $(1 - \frac{1}{N})$ as probability.

4.3 Spatial Interpolation

Probabilistic (Kriging) and deterministic (IDW, local polynomials) techniques are used to create maps for return levels with same return period. Interpolation with Kriging requires verification of minimum procedures to ensure proper use of the method, for instance,

- Structural analysis, which includes data normality check, for example, with Kolmogorov Smirnov or Shapiro Wilk goodness of fit tests, and if needed, data transformation to ensure data normality, e.g. using Box-Cox, and in addition, trend analysis to verify the need for trend modeling, in subsequent steps
- Semivariance Analysis: Use of available tools like cloud semivariogram, experimental semivariogram, directional semivariograms to verify isotropy or anisotropy, and different theoretical semivariograms, to ensure the best model of spatial autocorrelation, as a preliminary step to interpolation.

- Kriging Predictions: Use of different types of Kriging predictors, like simple, ordinary, universal, based on the results of the structural analysis.
- Cross Validation: Use of statistics like root mean square, average standard error, mean standardized, and root mean square standardized, that allow to measure the quality of the predictions and the magnitude of the errors.

Possible advantage of deterministic methods, is a better assessment of the local variability of spatial autocorrelation. It can also be considered with IDW or local polynomials a detailed assessment of structural analysis and cross validation. At the end of the spatial interpolation analysis all the predictions can be compared to select the most suitable result.

4.4 Integration with Non-Hurricane data

Engineers (2017) propose the equation C26.5-2 for combination of statistically independent events, of non-hurricane and hurricane wind speed data.

$$\mathrm{P_e(y > Y_N) = 1 - P_{NH}(y < Y_N) * P_H(y < Y_N)} \tag{4.5}$$

Where $P_e(y > Y_N)$ is the annual exceedance probability for the combined wind hazards, $P_{NH}(y < Y_N)$ is the annual non-exceedance probability for non-hurricane winds, and $P_H(y < Y_N)$ is the annual non-exceedance probability for hurricane winds.

To understand equation (4.5), it is important to remember that to calculate return level Y_N, for a given N-year return period, the exceedance probability $\frac{1}{N}$ of Y_N is calculated. Then, the non-exceedance probability for Y_N is $\left(1 - \frac{1}{N}\right)$. The procedure consist in the creation of a new hazard curve, calculating all $P_e(y > Y_N)$ values for different Y_N return levels, combining hazard curves from non-thunderstorm and thunderstorm data.

After the combined hazard curve is created, a new process of spatial interpolation need to be accomplished. In case of absence of hazard curves for stations, but availability of return levels maps, it becomes necessary to recreate hazard curves cell by cell, to apply equation (4.5). In this case are required as many maps as possible for different return periods, in order to estimate detailed enough hazard curves from return level values (cell values).

Chapter 5

Results

In this section, will be shown first, the data source comparison to face the downscaling issue by using ERA5 and ISD database, then, the resulting process of fitting a POT-PP in the station 801120, which includes revision of intensity function parameters, goodness of fit, hazard curve, return levels, and comparison with POT-GPD results, next, non-hurricane maps outputs, which includes results for ISD and ERA5 stations, using POT-PP and POT-GPD approaches, and finally, output maps combining hurricane and non-hurricane results will be displayed.

5.1 Data Standardization and Downscaling Support

Looking for a statistical justification in the use of ISD (model) and ERA5 databases (forecast), as input data for this study, and considering the *downscaling approach* described in the standartization process section of methodology, data sources ISD and IDEAM were standardized to enable comparison. Standardization consisted of transforming the data to be equivalent to 3-s gust V_3, 10 meters anemometer height, and open space roughness. In the comparison process, for coincident stations by spatial location, it was checked if the velocity values (standardized) in the three sources, for equal dates, were similar in magnitude.

5.1.1 Data Standardization

None of the sources required anemometer height standardization. Lettau (1969) was used for roughness standardization of ISD and IDEAM, applying the method station by station. Gust velocities standardization was done using Durst cuve, and in order to obtain V_3 from Durst curve, it is required to start from V_{3600} (average hourly speed), or from a different wind gust speed, for instance V_5.

For ERA5:

- Variable *10m wind gust - 10fg* of ERA5 data source does not need any standardization,

because it comes standardized from the source

For ISD:

- Wind velocity from ISD comes from source as V_5, that is, five seconds gust wind velocity. To standardize from V_5 to V_3, using Durst curve, the correction factor is 1.03.
- Wind velocity V_5 from 74 ISD stations, was standardized station by station, using procedure described in Surface Roughness at Open Space section, and Averaging Time 3-s Gust section

For IDEAM:

- As the original variables obtained from IDEAM, do not represent gust speeds, it is necessary to start from *average hourly speed* V_{3600}, to obtain 3-s gust V_3. To standardize from V_{3600} to V_3, using Durst curve, the correction factor is 1.51.
- It was not possible to obtain the *average hourly speed* V_{3600} from IDEAM, see table 2.2, but from *instantaneous wind velocity each 2 minutes - VV_AUT_2* it is possible to obtain a **good** estimator of V_{3600}, and from *instantateous wind velocity each 10 minutes - VV_AUT_10* it is possible to obtain a **poor** estimator of V_{3600}.

5.1.2 Data Comparison

The available IDEAM data allowed two comparison processes, with quality data for a few stations, and with low quality data, but available for all stations.

In both cases, to make the use of ISD and ERA5 viable, its time series are expected to be as similar as possible to IDEAM. To verify this, two types of graphics were constructed:

- Time series overlay for the three sources. Not very effective method due to the large amount of data that makes the graphics unreadable.
- Scatter plot maps comparing two different sources. Matching values by time, were sorted in ascending order and put together on a scatter plot. The expected behavior in case of similarity in the data, is that all the points fall in a 45° line

Quality data available in some IDEAM stations

IDEAM VV_AUT_2 was available for twenty (20) stations, of which only twelve (12) were *perfectly equivalent* to ISD stations (see table 5.1). VV_AUT_2 dataset was transformed to V_{3600} (average hourly speed), averaging all 20 values available per hour. For twelve matching stations, wind velocity V_{3600} (transformation of VV_AUT_2), was standardized station by station, using procedure described in Surface Roughness at Open Space section and Averaging Time 3-s Gust, and finally, for the same twelve ISD and IDEAM standardized stations, a comparison was done against matching ERA5 stations (the corresponding cell in ERA5 that has within ISD and IDEAM locations).

Table 5.1: Equivalent stations from ISD and IDEAM, representing the same weather station in the field

ISD_ID	IDEAM_ID
803980	48015050
803700	52055230
802110	26125061
802100	26125710
801120	23085270
801100	27015330
800970	16015501
800940	23195502
800630	13035501
800360	28025502
800350	15065180
800280	29045190

Stations 28025502 from IDEAM, 800360 from ISD, and 416 (cell with center point in $-73.25°$ longitude, and $10.5°$ latitude) from ERA5, showed high correspondence. See figures 5.2. Unfortunately, in the other eleven stations, there was no high equivalence between sources.

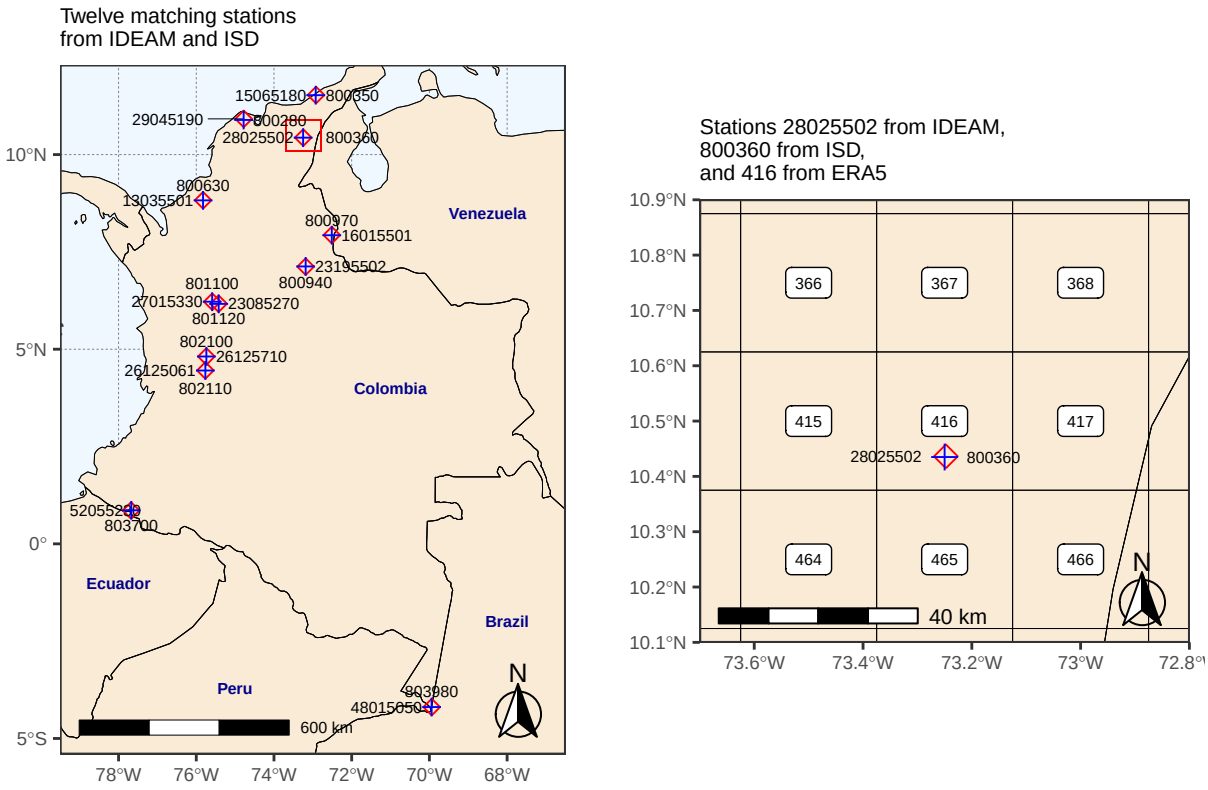

Figure 5.1: Left: Twelve matching stations from IDEAM and ISD. Right: Stations 28025502 from IDEAM, 800360 from ISD, and 416 from ERA5

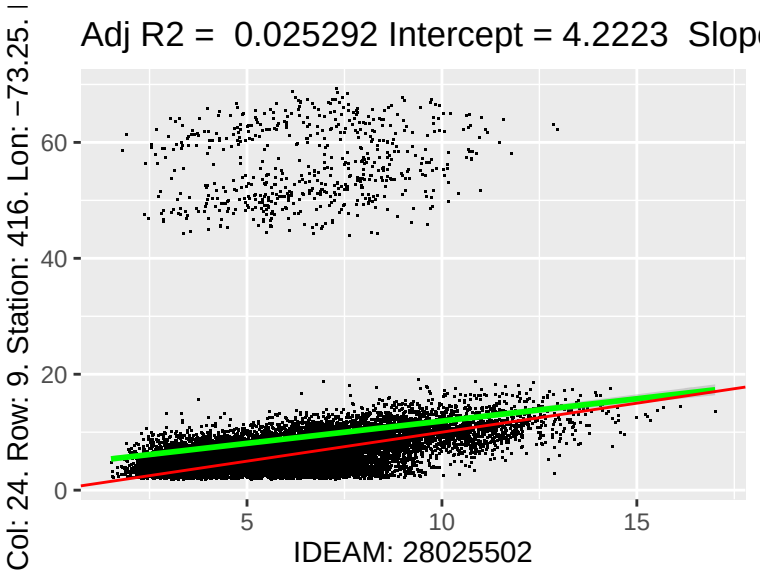

Figure 5.2: Scatter plot of IDEAM vs ERA5. Stations comparison: 28025502 (IDEAM), and 416 (ERA5)

Poor data available in all IDEAM stations

VV_AUT_10 was available for 204 stations, and despide that V_{3600} calculated from this source, is not an accurate or quality estimator, the standarization process was done to allow a comparison process.

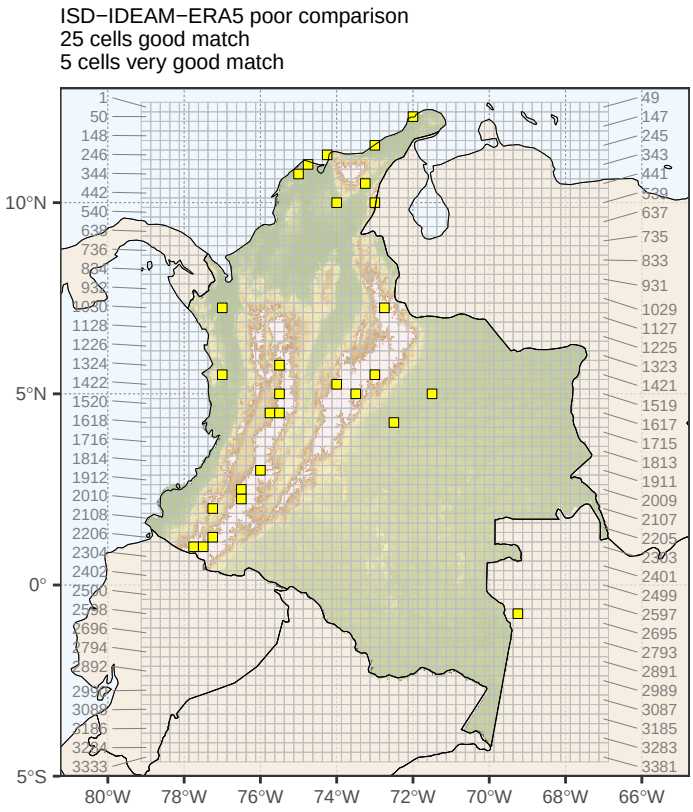

Figure 5.3: Left: Twelve matching stations from IDEAM and ISD. Right: Stations 28025502 from IDEAM, 800360 from ISD, and 416 from ERA5

5.2 POT-PP in ISD Station 801120

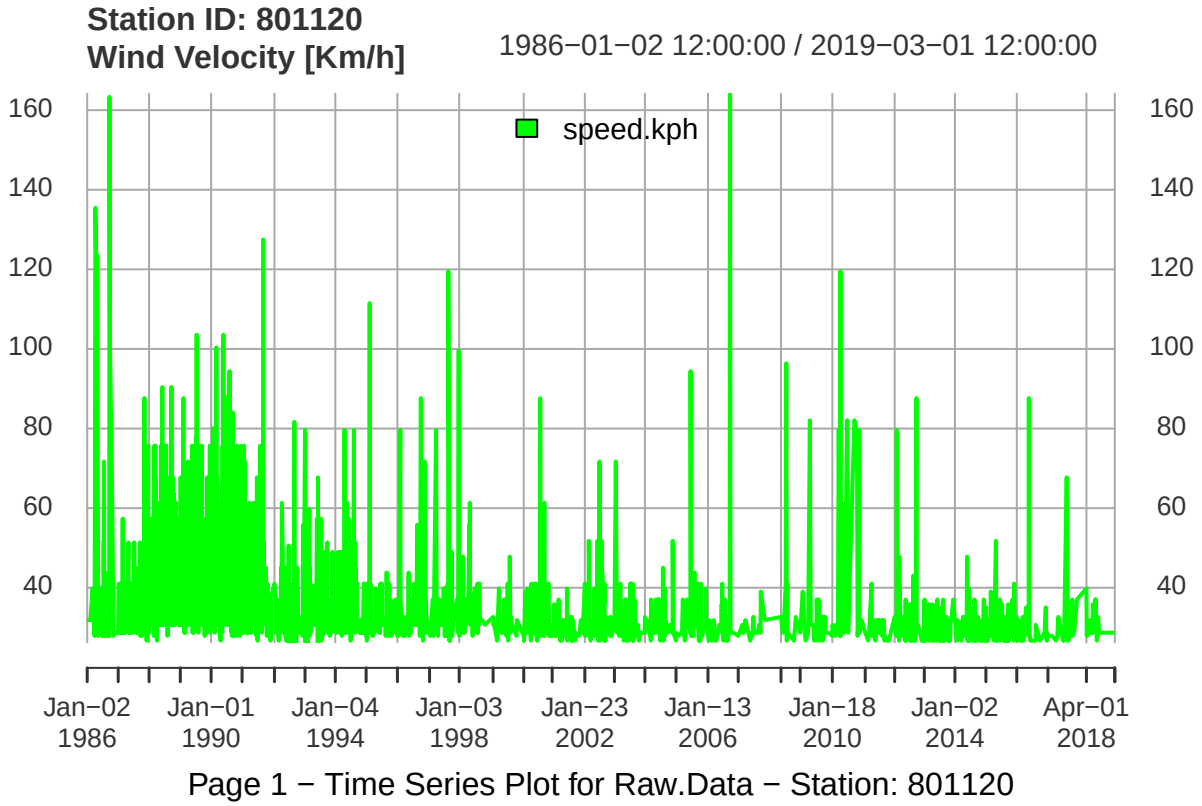

Figure 5.4: Time Series ISD. Station 801120

5.2.1 W-statiscis Plot

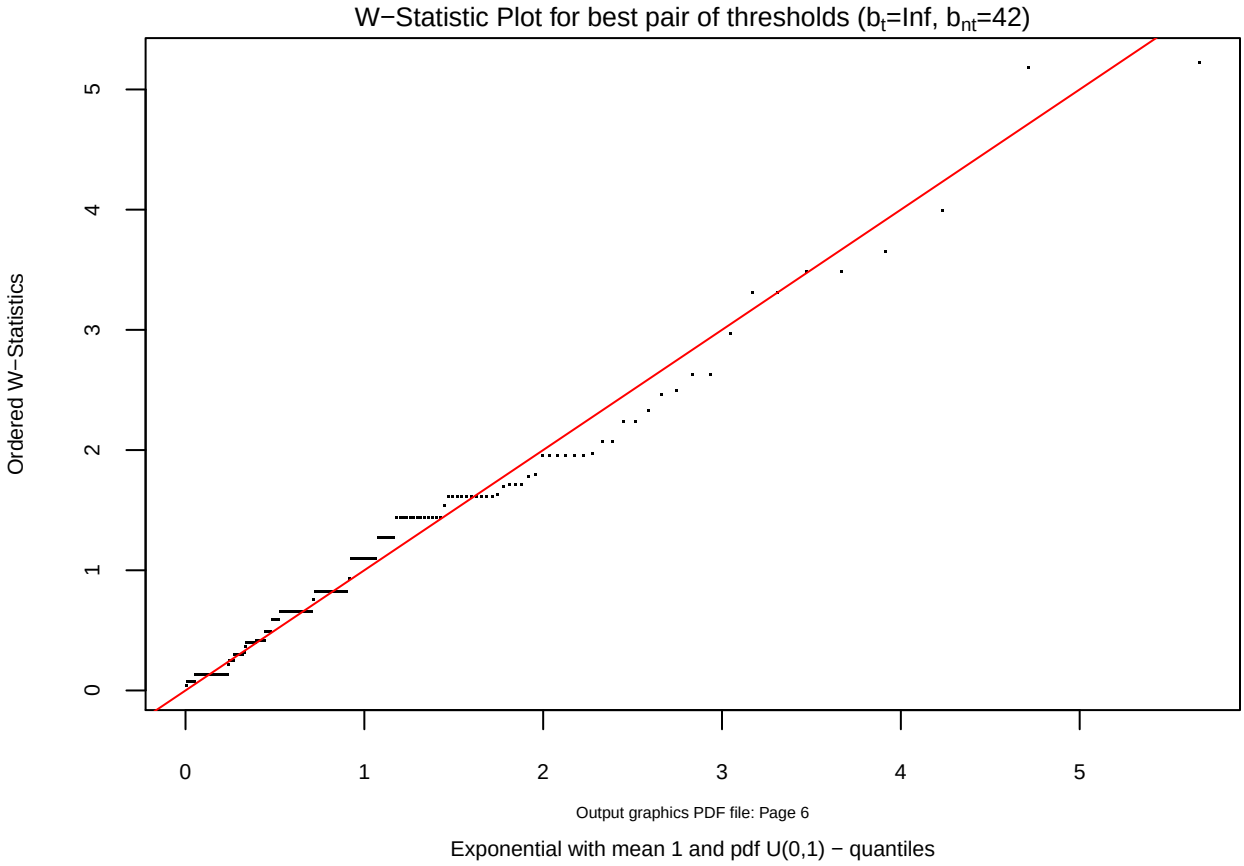

Figure 5.5: W-Statiscics Plot. Best Threshold Pair. Station 801120

5.2.2 Parameters

5.2.3 Fitted pdf and cdf

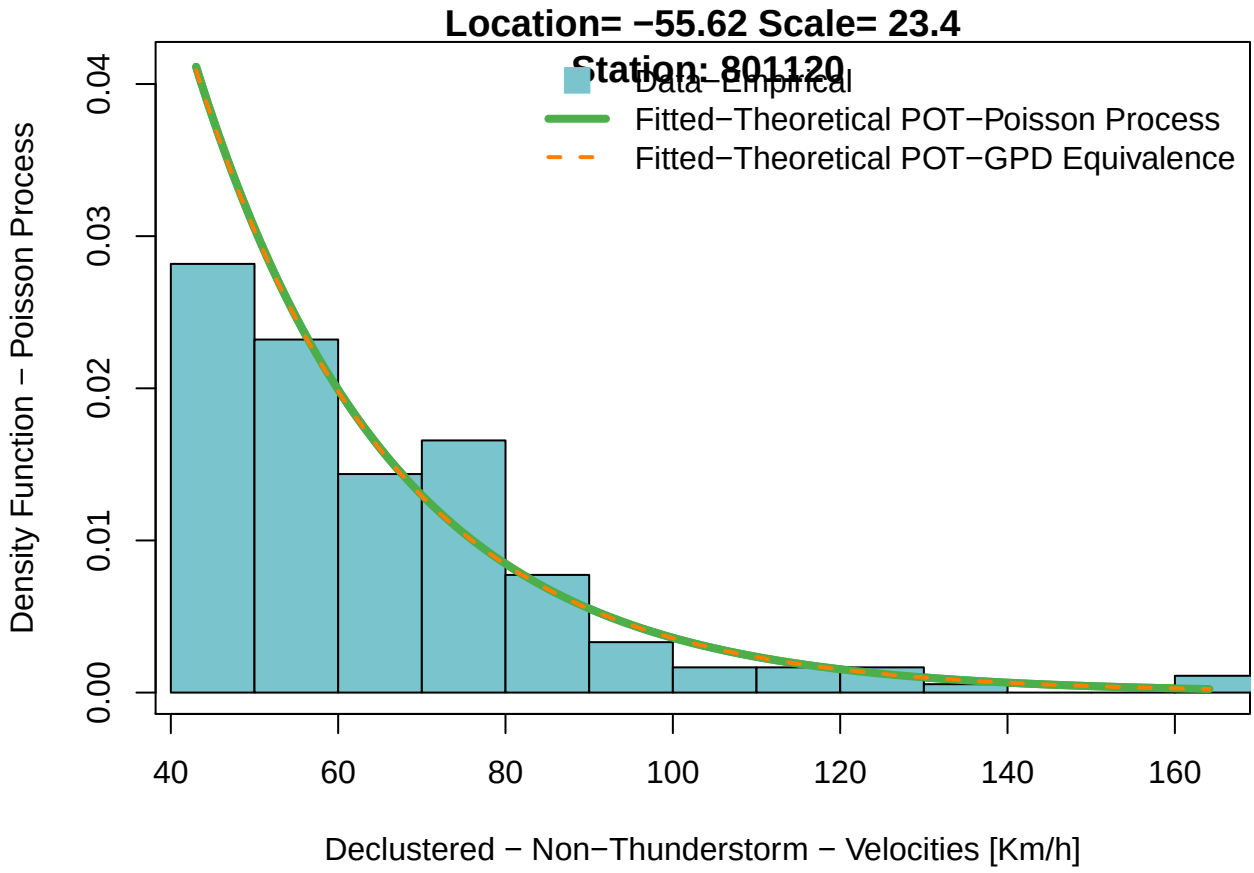

Figure 5.6: pdf POT-PP. Station 801120

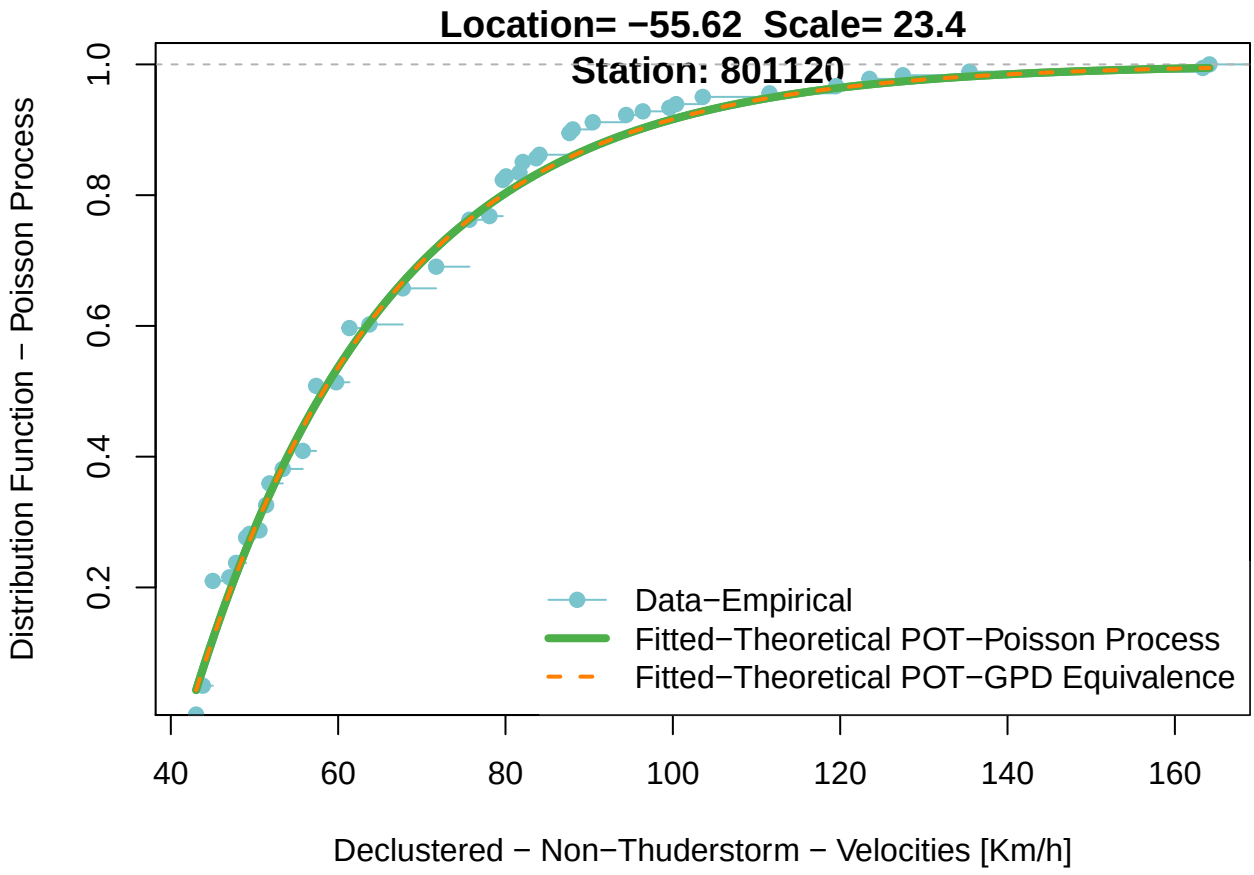

Figure 5.7: cdf POT-PP. Station 801120

5.2.4 Goodness of Fit

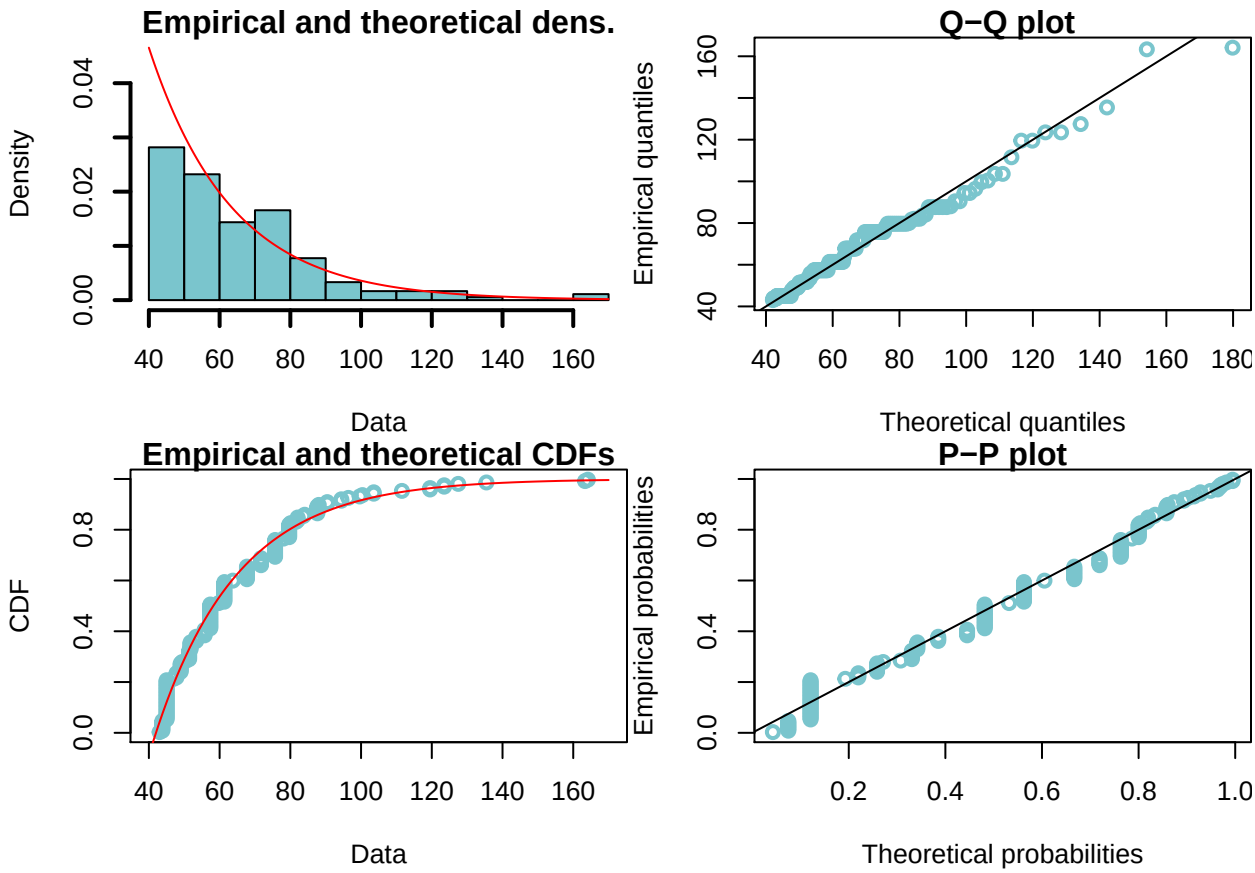

Figure 5.8: Graphic Diagnosis Of Goodness of Fit. Station 801120

5.2.5 Hazard Curve and Return Levels

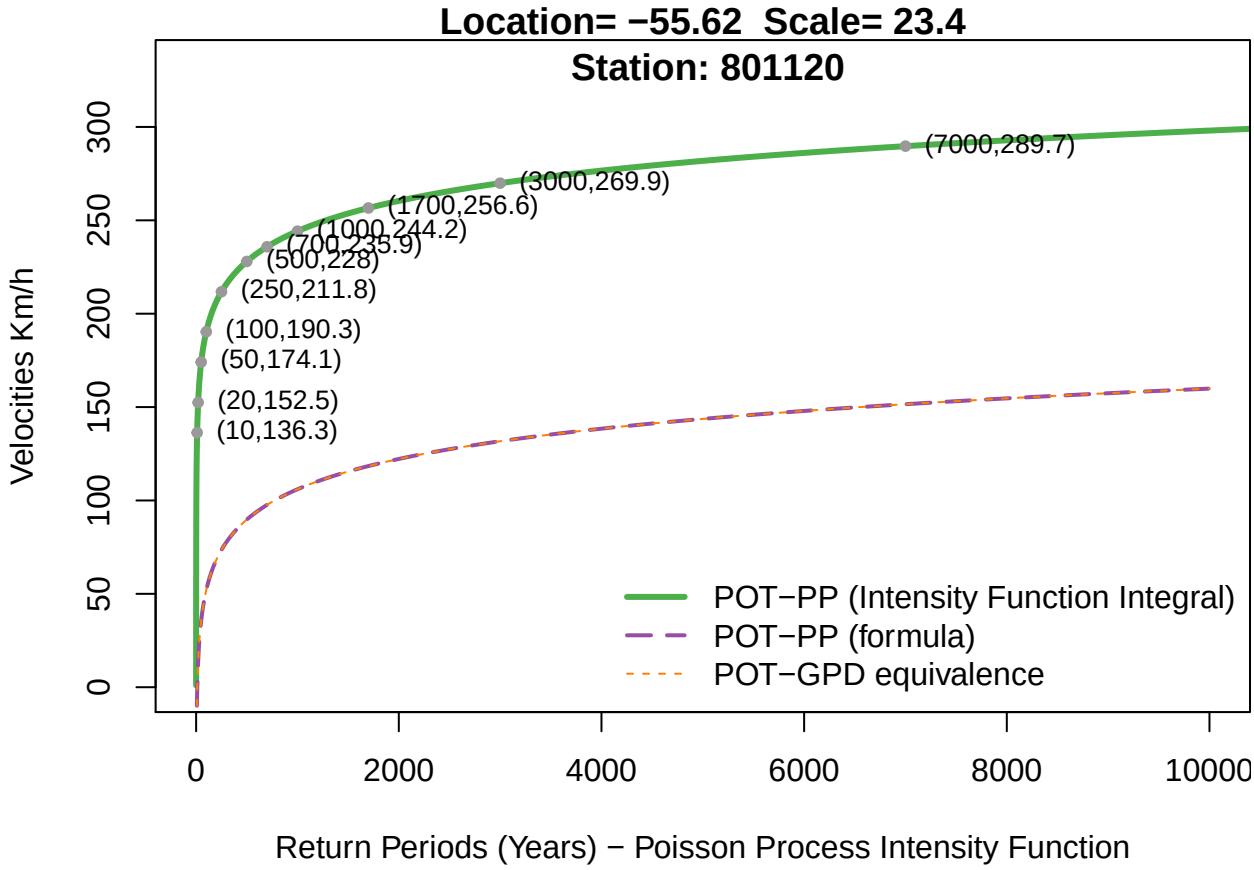

Figure 5.9: Hazard Curve. Station 801120

5.2.6 Comparison with POT-GPD

5.3 Non-Hurricane Maps

5.3.1 ISD

POT-PP

POT-GPD

5.3.2 ERA5

POT-PP

POT-GPD

5.4 Hurricane and Non-Hurricane Maps

5.4.1 ISD

POT-PP

POT-GPD

5.4.2 ERA5

POT-PP

POT-GPD

Chapter 6

Discussion

Conclusion

Appendix A

R Code

References

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer London. http://doi.org/10.1007/978-1-4471-3675-0

C. S. Durst, B. A., O. B.E. (1960). Wind speeds over short periods of time. *The Meteorological Magazine*, *89*(1056), 181–187. Retrieved from https://www.depts.ttu.edu/nwi/Pubs/ReportsJournals/ReportsJournals/Windspeeds.pdf

Davison, A. C., & Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, *52*(3), 393–442. Retrieved from http://www.jstor.org/stable/2345667

Engineers, A. S. O. C. (2017). *Minimum design loads and associated criteria for buildings and other structures (asce7-16)*. American Society of Civil Engineers. Retrieved from https://www.ebook.de/de/product/35017614/american_society_of_civil_engineers_minimum_design_loads_and_associated_criteria_for_buildings_and_other_structures_7_16.html

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

IDEAM. (1999, June). Aeronautical information. Annual wind regime. Web Page. Retrieved from http://bart.ideam.gov.co/cliciu/rosas/viento.htm

IDEAM. (2005). *Protocolo toma de datos de campo y emplazamiento de estaciones meteorológicas*.

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

Lettau, H. (1969). Note on aerodynamic roughness-parameter estimation on the basis of roughness-element description. *Journal of Applied Meteorology*, *8*(5), 828–832. http://doi.org/10.1175/1520-0450(1969)008%3C0828:NOARPE%3E2.0.CO;2

Masters, F. J., Vickery, P. J., Bacon, P., & Rappaport, E. N. (2010). Toward objec-

tive, standardized intensity estimates from surface wind speed observations. *Bulletin of the American Meteorological Society*, *91*(12), 1665–1682. http://doi.org/10.1175/2010bams2942.1

NIST. (2012, February). Standardized extreme wind speed database for the united states. Web Page. Retrieved from https://www.itl.nist.gov/div898/winds/NIST_TN/nist_tn.htm

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