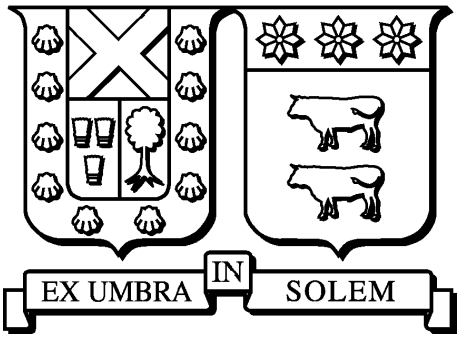


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE MATEMÁTICAS

MEMORIA DE INGENIERÍA:

METODOLOGÍA DE DOWNSCALING APLICADO A UN MODELO DE MESOESCALA EN LA REGIÓN DE VALPARAÍSO

Francisco Javier Benavides Lorca

Profesor Guía:

Juan G. Peypouquet Urbaneja.
Lisandro J. Fermin.

Universidad Técnica Federico Santa María

Departamento de Matemática
Valparaíso-Chile

Metodologia de Downscaling Aplicado a un Modelo de Mesoescala en la Región de Valparaíso

Memoria presentada por:
Francisco Javier Benavides Lorca

Como requisito parcial
para optar al título profesional Ingeniero Civil Matemático

Profesores Guías:
Juan G. Peypouquet
Lisandro J. Fermin

Examinadores:
Juan G. Peypouquet
Lisandro J. Fermin
Felipe Osorio

Diciembre, 2017

Material de referencia, su uso no involucra responsabilidad del autor o de la Institución.

TÍTULO DE LA MEMORIA:
Metodología de Downscaling Aplicado a un Modelo de Mesoescala en la Región de Valparaíso.

AUTOR: Francisco Javier Benavides Lorca.

TRABAJO DE MEMORIA, presentado como requisito parcial para optar al título profesional Ingeniero Civil Matemático de la Universidad Técnica Federico Santa María.

COMISIÓN EVALUADORA:

Integrantes	Firma
Juan G. Peypouquet Universidad Técnica Federico Santa María, Chile.	_____
Felipe Osorio Universidad Técnica Federico Santa María, Chile.	_____
Lisandro J. Fermin Universidad de Valparaíso, Chile.	_____

Valparaíso, Diciembre 2017.

Agradecimientos

Hay tantas personas a las cuales debo agradecer su apoyo durante todos estos años que es difícil no temer olvidar alguna. En primer lugar agradecer profundamente a mi familia, en especial a mi madre por siempre estar allí, a pesar de todo; A mis amigos de la vida y colegio, Danilo, Kathy y a todos de la expedición Darwin que me hacen rabiar (va para ti Danilo) y reír por partes iguales, pero no imagino mi vida sin ellos; A mis amistades de la época del pregrado que, a pesar de no tener contacto con algunos de ellos actualmente, los llevo a todos en mi corazón, especialmente a Eric, Pamela y Alexis que fueron pilares fundamentales para superar todas las barreras que puso la carrera; A esos amigos que he hecho en el último tiempo, Esteban y Chicoria, dos grandes personas que he podido conocer más a fondo este último año; A mi actual polola, Lorena, que a pesar de todo siempre esta allí para apoyarme; A mis profesores del colegio, especialmente a la Profe Eva y Paola que ayudaron a definir mi camino en la vida, como también a mis profesores de la universidad, sobre todo al profesor Lisandro, quien a trabajado conmigo codo a codo durante todo este proceso y a su familia; A toda la gente del CIMFAV y a la profe Soledad que hasta el día de hoy me brindan apoyo y confían en mis capacidades. Probablemente se me queden muchas otras personas por mencionar, pero todos los que se han cruzado en mi camino han ayudado a ser quien soy actualmente, por eso, MUCHAS GRACIAS.

"Lucha, nunca claudiques, recuerda que tu puedes conseguir lo que sea que quieras".

-Ikki (Caballero del Fénix)

Índice general

Agradecimientos v

Índice general vi

1. Introducción 1

2. Fundamentos 5
2.1. Conceptos Básicos . . . 5
2.2. Nociones sobre series temporales . . . 8
2.3. Clasificación de series temporales: Distancia Telescópica . . . 13
2.4. Estimación de funciones de densidad . . . 19

3. Downscaling 23
3.1. Downscaling Dinámico . . . 27
3.1.1. Weather Research and Forescasting(WRF) Model . . . 29
3.2. Downscaling Estadístico . . . 31
3.2.1. Métodos Lineales . . . 33
3.2.2. Métodos de clasificación del clima . . . 37
3.2.3. Simulación de Clima . . . 38
3.3. Modelo de Downscaling Geométrico . . . 41
3.3.1. Estimación de la función de interpolación: Metodología de Regresión Lineal Local . . . 44

4. Resultados 53
4.1. Caso de Estudio y metodología de trabajo . . . 53
4.2. Análisis Exploratorio . . . 57
4.3. Análisis de resultados. . . . 59
4.3.1. Campo de viento a 1.5[Km]: Interpolación Bicúbica . . . 59
4.3.1.1. Análisis Espacial . . . 62
4.3.1.2. Análisis temporal . . . 63
4.3.2. Regresión Lineal: 3km vs 1.5km . . . 69
4.3.2.1. Análisis exploratorio . . . 69
4.3.2.2. Análisis espacial . . . 73
4.3.2.3. Análisis temporal . . . 80

4.3.3. Downscaling: 0.5 [Km] 90
4.3.3.1. Análisis espacial 90
4.3.3.2. Análisis temporal 94
4.4. Clasificación 100
4.4.1. Regresión Lineal por clases. 105
4.4.2. Bandas de confianza. 106

5. Conclusión **110**

Capítulo 1

Introducción

En los últimos años la necesidad de cuidar el medio ambiente debido a los efectos del cambio climático y la búsqueda de energías renovables, han impulsado el estudio de fenómenos meteorológicos y el impacto de ellos en nuestro día a día. Olas de calor y bajas temperaturas, huracanes de gran magnitud y tormentas de gran intensidad son cada vez más comunes, por lo que gran parte de los esfuerzos están en la predicción de estados futuros del clima, creación de mecanismos de contención de catástrofes, variabilidad del clima, entre otras. Además, la necesidad de implementar energías más limpias como la eólica motivan a estudiar variaciones en el comportamiento de los campos de viento.

Para entender completamente un fenómeno climático se debe tener en cuenta a que escala se observa, pues el comportamiento de este puede ser completamente diferente según la altura y la superficie que abarque. Por ejemplo, los campos de viento medidos a 150 metros de altura presentan un comportamiento más regular que las mediciones tomadas a 10 metros, debido a que la influencia de la rugosidad del terreno es mayor para estas últimas. En general, se definen escalas verticales y horizontales, en donde las primeras están asociadas a diferentes niveles de presión, temperatura, altura, etc; y las horizontales hacen referencia al tamaño de la zona de interés. Las escalas verticales se eligen según la característica que se desee estudiar, siendo una opción el analizar el comportamiento del fenómeno a 10 metros sobre la superficie terrestre; mientras que para la escala horizontal se suelen identificar tres niveles principales de resolución: Escala Sinóptica o Global ($\geq$20000 [Km]), Meso Escala (entre 20000[Km] y 0.1[Km])

y Micro Escala (≤ 0.1[Km]).

Los modelos definidos sobre fenómenos climáticos que ocurren a nivel global, es decir, en la escala Sinóptica, poseen gran precisión en su capacidad de ajuste y predicción, pues se construyen mediante las leyes físicas que gobiernan a los fenómenos, existiendo dos tipos principales de estos modelos: los de circulación global o general (GCMs por su sigla en inglés) y los climáticos regionales (RCMs por su sigla en inglés) los cuales se diferencian en la escala en la que actúan, siendo los GCM usados a nivel Sinóptico y los RCM a nivel mesoescala. Estos modelos son ampliamente aceptados pues suelen poseer una gran exactitud. Sin embargo, al momento de adaptarlos a resoluciones espaciales más finas, estos pierden precisión en sus proyecciones, por ejemplo, en [13] comparan el desempeño de diferentes GCMs a baja resolución en la cuenca Mediterránea y en [26] analizan el desempeño del RCM llamado "Climate Comunity Model" CC2 a resoluciones de 310[Km] y 125[Km] en la cuenca del río Sacramento en California, mostrando en ambos casos la perdida de eficiencia y precisión de estos modelos.

La pérdida de precisión suele ser provocada por problemas de parametrización, pues para que el modelo funcione en resoluciones más finas, la información de entrada debe ser ajustada a este cambio, pero este proceso es acompañado de un gran aumento en el costo computacional al implementar el modelo. Esto motiva el desarrollo de técnicas que sean capaces de usar la información proveniente de estos modelos a gran escala, en donde presentan gran capacidad de predicción y ajuste, para así conocer a nivel local el comportamiento del fenómeno. En resumen, mediante una malla más gruesa (escala Sinóptica) de información se desea obtener una malla más fina (meso o micro escala) y de mayor información. Estas técnicas reciben el nombre de "*downscaling*".

El downscaling puede realizarse de forma espacial (obtención de mallas espaciales más finas) y/o temporal (escala de tiempo más fina, como por ejemplo pasar de una información mensual a una semanal) y dependiendo de su naturaleza suelen clasificarse en dos tipos: Dinámico (o modelos anidados) los cuales al igual que los GCM buscan resolver ecuaciones físicas que gobiernan a los fenómenos; y Estadísticos, que establecen relaciones empíricas entre el comportamiento local y el global.

En este trabajo estamos interesados en establecer una técnica estadística de downscaling espacial para la intensidad de viento en la región de Valparaíso en Chile, abarcando la zona entre 32°0′19,8″ y 34°3′9,72″ latitud sur; 70°21′46,08′ y 71°57′1,23″ longitud oeste, motivados por dos razones principales. La primera es la relevancia que ha cobrado en el último tiempo este fenómeno, debido a la alta tasa anual de incendios, emergencias energéticas, explotación de energía eólica, entre otros; la segunda razón es el efecto de su accidentada geografía en el comportamiento de este fenómeno, provocando cambios bruscos en el comportamiento del viento a cortas distancias, lo cual se traduce en trayectorias irregulares del viento. Debido a lo anterior es necesario implementar modelos que proporcionen buenos ajustes en escalas horizontales finas y que disminuyan los costos computacionales que requieren los GCM o RCM en baja resolución espacial.

Para nuestro estudio de los campos de viento en la zona de interés, utilizamos información de un RCM llamado *Weather and Research Forecasting* (WRF). Este modelo es ampliamente conocido por su precisión a la hora de simular diferentes fenómenos climáticos y la capacidad de asimilar datos reales a nivel global [5], además es capaz de entregar en un tiempo razonable (uno o dos días) las mallas de intensidad de viento de resolución de hasta 1[Km] para la región de Valparaíso en un periodo de tiempo de un mes. Sin embargo, el WRF puede tardar hasta un mes en entregar resultados a una resolución espacial de 0.5[Km]. Nosotros nos planteamos utilizar la información obtenida desde el WRF para el campo de viento a resolución espacial de 3[Km] y 1[Km], medido a una altura de 10 metros para estimar el campo de intensidad de viento a 0.5[Km]. El modelo de downscaling propuesto en este trabajo consiste en un esquema de tres pasos: El primer paso es un aumento de escala o *upscaling*, el cual consiste en estimar una malla intermedia a 1.5[Km] mediante la malla 1[Km], utilizando métodos de interpolación determinista; el segundo paso consiste en establecer una relación empírica entre las mallas a 3[Km] y 1.5[Km], basándose en las hipótesis de regularidad local, la cual permite ajustar un modelo lineal localmente, y autosimilaridad espacial, la que permite establecer una relación geométrica entre mallas de diferente resolución, captando el efecto de cambio de escala; el tercer paso, correspondiente al downscaling, consiste en usar la relación empírica establecida en el paso anterior para obtener una estimación del campo de intensidad de viento a 0.5[Km] mediante el campo de intensidad de viento a 1[Km]. Esta metodología emplea una combinación de upscaling y downscaling, y un enfoque geométrico basado en la estructura de los datos provenientes del WRF, escapándose de la estructura usual de los

métodos estadísticos. La elección de este método combinado, a través del uso de una malla intermedia a 1.5[Km] permite que el ajuste lineal local este bien determinado, evitando problemas de identificabilidad y colinealidad.

Esta memoria esta estructura de la siguiente manera: en el capítulo 2 presentamos los fundamentos teóricos en los que se sustenta nuestro trabajo, dando algunas definiciones básicas y resultados que ayudan a comprender el porqué del modelo desarrollado. En el capítulo 3 extendemos el concepto de downscaling, presentamos algunas de las técnicas básicas usadas y exponemos nuestro modelo. En el capítulo 4 exponemos el caso de estudio, describiendo los resultados obtenidos para el mes de septiembre del año 2014, con el fin de visualizar la calidad del método propuesto a nivel espacial y temporal. Finalmente, en el capítulo 5, daremos nuestras conclusiones y recomendaciones para trabajos futuros.

Capítulo 2

Fundamentos

En esta sección introduciremos los fundamentos teóricos sobre los cuales se sustenta nuestro trabajo y que permiten entender el desarrollo de las ideas planteadas.

2.1. Conceptos Básicos

Definición 1. Un espacio medible es un par $(\Omega, \mathcal{F})$, donde Ω es un conjunto y $\mathcal{F}$ es una σ-álgebra; es decir, una colección (no vacía) de subconjuntos de Ω tal que:

i) $\emptyset \in \mathcal{F}$.

ii) Si $A \in \mathcal{F}$ entonces $A^c = \Omega \setminus A \in \mathcal{F}$.

iii) Si $(A_k)_{k\in\mathbb{N}} \subset \mathcal{F}$, entonces $\bigcup_{k\in\mathbb{N}} A_k \in \mathcal{F}$.

Además la tripleta $(\Omega, \mathcal{F}, P)$ se denomina espacio de probabilidad, donde la función $P : \mathcal{F} \to [0, 1]$ es una medida de probabilidad, es decir, satisface las siguientes propiedades:

i) $P(\Omega) = 1$.

ii) $P(A^c) = 1 - P(A), \forall A \in \mathcal{F}$.

iii) Si $(A_k)_{k\in\mathbb{N}} \subset \mathcal{F}$ tal que $A_i \bigcap A_j = \emptyset$, $\forall i \neq j$, entonces:

$$P\left(\bigcup_{k\in\mathbb{N}} A_k\right) = \sum_{k\in\mathbb{N}} P(A_k).$$

Un ejemplo de espacio de probabilidad es $\big([0,1]^d, \mathcal{B}([0,1]^d), \lambda\big)$, donde $[0,1]^d$ corresponde al hipercubo unitario de dimensión d; $\mathcal{B}\left([0,1]^d\right)$ es la σ-álgebra de Borel para $[0,1]^d$, la cual corresponde a la menor σ-álgebra que contiene los conjuntos abiertos del hipercubo $[0,1]^d$ y λ corresponde a la medida de Lebesque o medida de probabilidad uniforme en $[0,1]^d$.

En adelante consideramos un espacio de probabilidad $(\Omega, \mathcal{F}, P)$, sobre el cual se definen los siguientes tipos de funciones a valores en $\mathbb{R}$.

Definición 2. Una variable aleatoria (v.a.) a valores reales es una función $X : \Omega \to \mathbb{R}$ tal que es $\mathcal{F} - \mathcal{B}(\mathbb{R})$ medible, es decir:

$$X^{-1}(B) \in \mathcal{F}, \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

Una característica de las variables aleatorias reales es la frecuencia con la cual sus valores se mantienen bajo cierto umbral, para ello se introduce la *función de distribución de* X.

Definición 3. Para una variable aleatoria X a valores reales, definida sobre el espacio $(\Omega, \mathcal{F}, P)$, su función de distribución $F_X : \mathbb{R} \to [0,1]$ está dada por:

$$F_X(x) = P\{w \in \Omega : X(w) \leq x\} = P(X \leq x).$$

La cual posee las siguientes propiedades:

i) $0 \leq F_X(x) \leq 1$.

ii) F_X es no decreciente.

iii) $\lim_{x \to -\infty} F_X(x) = 0$.

iv) $\lim_{x \to \infty} F_X(x) = 1$.

v) $F_X(b) - F_X(a) = P(a < X \leq b)$.

Cuando la función de distribución $F_X(x)$ es de la forma:

$$F_X(x) = \int_{-\infty}^{x} f_X(s) ds.$$

para alguna función f_X positiva e integrable en $\mathbb{R}$, se dice que la variable aleatoria X tiene por *función de densidad* a $f_X(\cdot)$, en cuyo caso $F'_X(x) = f(x)$.

La función de densidad nos da una interpretación gráfica y permite caracterizar las variables aleatorias. Por ejemplo, se dice que una variable aleatoria $X : \mathbb{R} \to \mathbb{R}$ tiene distribución normal, de parámetros μ y σ, si su función de densidad está dada por

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

En este caso, μ es un parámetro de centrado de X y σ^2 es un parámetro de forma, en el que está representada la variabilidad de X con respecto a μ. En la Figura 2.1 se presenta el gráfico de la función de densidad de la distribución normal estándar, la cual posee parámetros $\mu = 0$ y $\sigma = 1$.

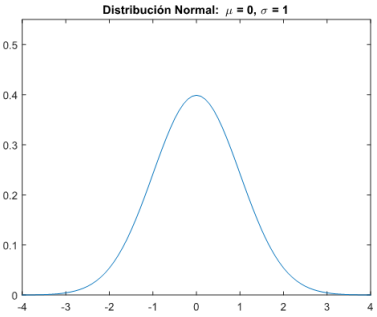


FIGURA 2.1: Gráfico de la función de densidad para la distribución normal estándar

Mediante la función de distribución es posible definir el valor esperado de una variable aleatoria X.

Definición 4. Para una variable aleatoria X definida sobre el espacio $(\Omega, \mathcal{F}, P)$, se define su valor esperado o esperanza de X como:

$$E[X] = \int_{-\infty}^{\infty} x dF_X(x). \tag{2.1}$$

Si existe la función de densidad, se tiene también:

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

Adicionalmente, $\forall g$ medible, se define

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) dF_X(x). \tag{2.2}$$

Las integrales (2.1) y (2.2) están definidas en el sentido de la integral de Lebesgue-Sterling (ver [14]). Esta noción de valor esperado nos permite cuantificar características de la variable aleatoria X, como por ejemplo su valor medio, la variabilidad respecto a la media, tasa de ocurrencia de eventos extremos, entre otros.

El concepto de variable aleatoria no es suficiente para estudiar fenómenos a través del tiempo. En particular la medición de intensidad del viento en un punto fijo en el espacio, para dos instantes de tiempo t_1 y t_2 distintos entre sí, se consideran como los valores que toman las variables aleatorias X_{t_1} y X_{t_2}, las cuales debido a factores endógenos y exógenos, como por ejemplo la temperatura y humedad del ambiente, suelen ser diferentes. Esto hace necesario considerar una sucesión de variables aleatorias $(X_t, t \in \mathcal{T})$, indexadas en el conjunto de índices $\mathcal{T}$, que consideraremos como el tiempo.

2.2. Nociones sobre series temporales

Procederemos definiendo el concepto de serie temporal, necesario para estudiar fenómenos que ocurren dentro de una ventana de tiempo determinada, además de mencionar características y propiedades necesarias en el desarrollo de este trabajo.

Definición 5. Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad y $\mathcal{T}$ un conjunto de índices. Un proceso estocástico o aleatorio $\mathbf{X}$ indexado en $\mathcal{T}$ es una familia de variables aleatorias $\mathbf{X} = (X_t, t \in \mathcal{T})$ definidas en $(\Omega, \mathcal{F}, P)$ con valores en el espacio $\left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\right)$, el cual es denominado espacio de estado.

Un proceso estocástico puede ser considerada de dos formas:

- Como una función definida sobre el espacio $\mathcal{T} \times \Omega$ que toma valores $X(t, \omega)$ en el espacio de estado $\left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\right)$, tal que para todo $t \in \mathcal{T}$ fijo, $X_t = X(t, \cdot)$ es una variable aleatoria sobre $(\Omega, \mathcal{F}, P)$. Así, $X(t, \omega)$ corresponde al valor que toma

la variable aleatoria X_t en el espacio de estado $\left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\right)$ al ser evaluada en $\omega \in \Omega$.

- Asociándolo a una variable aleatoria que toma valores en el espacio de funciones definidas sobre $\mathcal{T}$, a través de la aplicación $\omega \in \Omega \to \mathbf{X}(\omega) = \{X_t(\omega) : t \in \mathcal{T}\}$, que a cada $\omega \in \Omega$ asocia una trayectoria $\mathbf{X}(\omega)$ del proceso. La Figura 2.2 ilustra esta manera de interpretar a los procesos estocásticos, mediante un proceso AR(1) (ver Capítulo 3.2.3) considerando como espacio de estado $\mathbb{R}$.

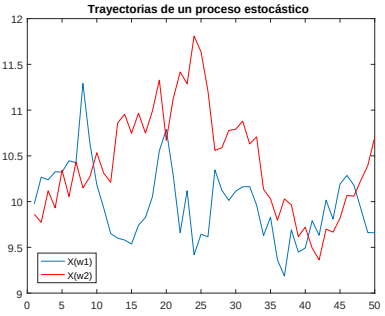


FIGURA 2.2: Ejemplo de trayectorias para un mismo proceso estocástico sobre el espacio de estado $\mathbb{R}$. Cada trayectoria corresponde a una función $X_t(\omega_0)$, donde ω_0 es fijo y $t \in \mathcal{T}$ varia.

Nuestro objetivo es encontrar una manera de caracterizar por completo el comportamiento de un proceso estocástico a través de su función de distribución. Para ello necesitamos los siguientes elementos:

Definición 6. Para un proceso estocástico $\mathbf{X} = (X_t, t \in \mathcal{T})$, se definen sus distribuciones finito dimensionales como:

$$F_{t_1,\ldots,t_n}(A) = P\left(\{X_{t_1}, \ldots, X_{t_n}\} \in A\right).$$

donde $t_1, \ldots, t_n \in \mathcal{T}$ y $A \in \mathcal{B}^n$, donde $\mathcal{B}^n$ es la σ-álgebra del espacio de estado $\mathbb{R}^{nd}$.

Las distribuciones finito dimensionales de los procesos estocásticos definen medidas de probabilidad en el espacio $(\Omega, \mathcal{F})$ y usaremos la notación $x_{1:k}$ para abreviar a una muestra $x_1, \ldots, x_k$.

Las distribuciones finito dimensionales de los procesos $\mathbf{X} = (X_t, t \in \mathcal{T})$ que estudiaremos satisfacen, $\forall A \in \mathcal{B}^n$ y $\forall t_1, \ldots, t_n \in \mathcal{T}$ las siguientes propiedades de consistencia:

- Simetría: $F_{t_1,...,t_n}(A) = F_{t_{\pi(1)},...,t_{\pi(n)}}(\pi(A))$, para cualquier permutación π del conjunto $\{1, ..., n\}$.
- Compatibilidad: $F_{t_1,...,t_n}(A) = F_{t_1,...,t_{n+1}}\left(A \times \mathbb{R}^d\right)$

Un teorema que garantiza la existencia de un proceso aleatorio con distribuciones finito dimensionales que satisfacen las propiedades de consistencia es el siguiente:

Teorema 1 (Teorema de Existencia de Kolmogorov). Si el sistema de distribuciones finito dimensionales $\{F_{t_1,...,t_n} : n \in \mathbb{N}, (t_1, \ldots, t_n) \in \mathcal{T}^n\}$ satisface las condiciones de simetría y compatibilidad, entonces existe un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ y un proceso $\mathbf{X}$, definido sobre $(\Omega, \mathcal{F}, P)$ tal que sus distribuciones finito dimensionales coinciden con las del sistema.

Las distribuciones finito dimensionales de un proceso aleatorio caracterizan muchas propiedades del proceso, pero no son suficientes. Por ejemplo, podemos definir dos procesos cuyas distribuciones finito-dimensionales sean las mismas y sin embargo uno posea trayectorias continuas y el otro discontinuidades de salto [8].

Para determinar la regularidad de las trayectorias de un proceso a partir de sus distribuciones finito dimensionales debemos considerar una hipótesis adicional, denominada separabilidad del proceso aleatorio.

Definición 7. Sea $\mathbf{X} = (X_t, t \in \mathcal{T})$ un proceso estocástico indexado en $\mathcal{T}$. Si $\mathcal{T}$ es un espacio métrico, entonces el proceso estocástico $\mathbf{X}$ se dice separable si existe un subconjunto de índices denso numerable $\{t_i,\ i \in \mathbb{N}\} \subset \mathcal{T}$ y un conjunto $\mathcal{N} \subset \Omega$ de medida de probabilidad cero, tales que para todo conjunto abierto $S \subset \mathcal{T}$ y todo conjunto cerrado $A \subset \mathbb{R}^d$, los subconjuntos

$$\{\omega :\ X_{t_i}(\omega) \in A,\ \ \forall t_i \in S\} \qquad \text{y} \qquad \{\omega :\ X_t(\omega) \in A,\ \forall t \in S\}$$

difieren el uno del otro solamente en un subconjunto de $\mathcal{N}$.

Definición 8. Dos procesos estocásticos $\mathbf{X}$ e $\mathbf{Y}$, definidos sobre el mismo espacio de probabilidad, se dicen estocásticamente equivalentes (o una versión el uno del otro) si

$$P\{\omega : X_t(\omega) = Y_t(\omega)\} = 1,\ \ \forall t \in \mathcal{T}.$$

Las trayectorias de dos procesos estocásticos equivalentes son iguales salvo en un conjunto $\mathcal{N} \subset \Omega$ tal que $P(\mathcal{N}) = 0$. La idea es, dado un proceso estocástico, siempre se puede trabajar con una versión equivalente a éste que sea separable.

Proposición 1. Si $\mathcal{T}$ es un espacio métrico, entonces todo proceso estocástico $\mathbf{X} = (X_t, t \in \mathcal{T})$ posee una versión equivalente separable.

De esta forma, para cada proceso estocástico $\mathbf{X}$ podemos trabajar con su versión equivalente separable, así al aplicar el Teorema 1 de Kolmogorov a las distribuciones finito dimensionales, obtenidas al considerar el conjunto de índices numerable, este proceso queda completamente determinados por dichas distribuciones finito dimensionales.

Estamos interesados en cierto tipo de procesos estocásticos, denominados *procesos de segundo orden*. Para ello consideramos el espacio $L^2_{\mathbb{R}^d}(\Omega, \mathcal{F}, P)$, el cual consiste de todas las variables aleatorias X definidas sobre $(\Omega, \mathcal{F}, P)$ que toman valores en el espacio de estado $\mathbb{R}^d$ y cumplen que:

$$E[X^\top X] = \|X\|_{L^2_{\mathbb{R}^d}} < \infty$$

donde $X = X(\omega) = (X_1(\omega), ..., X_d(\omega))^\top$ es denominado *vector aleatorio* y $\top$ denota al operador de transposición.

Definición 9. La esperanza o valor esperado de un vector aleatorio se define coordenada a coordenada. Si X es un vector aleatorio de dimensión d en $L^2_{\mathbb{R}^d}(\Omega, \mathcal{F}, P)$ su esperanza $\mu(X) = E[X]$ es un vector en $\mathbb{R}^d$ y la matriz de covarianza $\Gamma(X)$ es una matriz de dimensión $d \times d$ definida por

$$\Gamma(X) = E[(X - E[X])(X - E[X])^\top].$$

Definición 10. Si X, Y son vectores aleatorios pertenecientes al espacio $L^2_{\mathbb{R}^d}(\Omega, \mathcal{F}, P)$, la matriz de covarianza de X e Y se define como

$$\Gamma(X, Y) = E[(X - E[X])(Y - E[Y])^\top].$$

Definición 11. Un proceso estocástico $\mathbf{X} = \{X_t, t \in \mathcal{T}\}$ es de segundo orden si $X_t \in L^2(\Omega, \mathcal{F}, P)$, para todo $t \in \mathcal{T}$.

Las series temporales son procesos de segundo orden indexados en un conjunto $\mathcal{T}$ numerable ($\mathcal{T} \subset \mathbb{Z}$). De aquí en adelante hablaremos solamente de series temporales.

Definición 12. Consideremos la serie temporal $\mathbf{X} = (X_t, t \in \mathcal{T})$, definimos:

i) La función de media o esperanza como:

$$\mu_X(t) = E[X_t], \quad \forall t \in \mathcal{T}.$$

ii) La función de varianza como:

$$\begin{aligned} \sigma_X^2(t) &= \Gamma(X_t, X_t), \\ &= E[(X_t - \mu(X_t))^\top (X_t - \mu(X_t))], \quad \forall t \in \mathcal{T}. \end{aligned}$$

iii) La función de covarianza como:

$$\Gamma_X(s,t) = \Gamma(X_s, X_t), \quad \forall s,t \in \mathcal{T}.$$

iv) La función de autocorrelación como:

$$R_X(s,t) = \frac{\Gamma_X(s,t)}{\sigma_X(s)\sigma_X(t)}.$$

Algunas propiedades importantes sobre las funciones de distribución de series temporales son las siguientes:

Definición 13. Sea $\mathbf{X} = (X_t, t \in \mathcal{T})$ una serie temporal. Diremos que es estacionaria en el sentido

- Débil, si su función de media μ_X y su función de covarianza Γ_X satisfacen

  $$\mu_X(t+s) = \mu_X(t) \quad \wedge \quad \Gamma_X(s+h, t+h) = \Gamma_X(s,t), \quad \forall t, s, t+h, s+h \in \mathcal{T}.$$

  Como consecuencia inmediata tenemos que para esta clase de procesos $\mu_X(t) = \mu_X(0), \forall t \in \mathcal{T}$, además que $\Gamma_X(s,t) = \gamma(t-s)$, donde $\gamma(\cdot)$ es una función de una variable, es decir, la función de media es constante y la función de covarianza solo depende de la distancia $\tau = t - s$ con $s \leq t$.

- Fuerte, si sus distribuciones finito dimensionales satisfacen:

  $$F_{t_1,t_2,\ldots,t_n}(A) = F_{t_1+k,t_2+k,\ldots,t_n+k}(A), \quad \forall A \in \mathcal{F}, n \in \mathbb{N},$$
  $$\forall t_1, t_2, \ldots, t_n \in \mathcal{T}, \text{y} \quad k \in \mathbb{N} \quad \text{tal que} \quad t_1 + k, t_2 + k, \ldots, t_n + k \in \mathcal{T}.$$

En nuestro contexto la hipótesis de estacionaridad fuerte es equivalente a suponer que el comportamiento del fenómeno es estable o que no varía con el tiempo. Se puede verificar que la estacionaridad fuerte implica estacionaridad débil (ver [32]).

Definición 14. Una distribución F de una serie temporal definida sobre el espacio $(\Omega, \mathcal{F}, P)$ se denomina (estacionaria) ergódica si para todo conjunto $A \in \mathcal{F}$ tal que:

$$F^{-1}(A) = A \tag{2.3}$$

se tiene que $P(A) \in \{0, 1\}$. Los elementos en $\mathcal{F}$ que satisfacen la ecuación (2.3) se denominan invariantes.

La propiedad de ergodicidad es de gran importancia en la práctica, pues nos permite, como veremos más adelante, aproximar las funciones de distribución usando una única observación de la serie temporal.

2.3. Clasificación de series temporales: Distancia Telescópica

A continuación daremos las nociones básicas sobre el problema de clasificación en general, para luego hablar sobre el caso particular de la clasificación de distribuciones y la métrica que usaremos para ello, denominada *distancia telescópica*. El objetivo es, dado un conjunto de series temporales, poder clasificarlas de acuerdo a sus funciones de distribución, en donde dos series de tiempo pertenecen a un mismo grupo si y solo si sus funciones de distribución son similares con respecto a una medida adecuada. Lo anterior motiva a introducir una noción de distancia entre distribuciones de series temporales que permita realizar dicha clasificación, a través de muestras finitas de las series temporales

El problema de clasificación general es el siguiente: Dado un conjunto discreto de elementos $O = \{o_1, \ldots, o_N\}$, deseamos agrupar sus elementos en m subconjuntos o clases, denotados por $\{C_1, \ldots, C_m\}$, donde $1 \leq m \leq N$ y deben satisfacer las siguientes

propiedades:

- $C_i \bigcap C_j = \emptyset, \quad \forall i \neq j.$
- $\bigcup_{i=1}^{m} C_i = O.$

Cada subconjunto C_i está definido por una propiedad o característica común para sus elementos. En general, para resolver un problema de clasificación se necesitan dos ingredientes principales: un algoritmo consistente, es decir, que converja a la clasificación verdadera u objetivo (*target clustering*) y una métrica que nos permita decir que tan parecidos entre sí son los elementos a clasificar, la cual suele denominarse *medida de similaridad*.

Un algoritmo clásico corresponde al de clasificación del punto más lejano (*farthest point clustering*). La idea principal de este algoritmo es que cada clase C_j posee un elemento representativo c_j y el resto de los elementos es asignado a la clase cuyo representante este más cercano. Este algoritmo define a los elementos representativos de la siguiente manera:

i) El primer elemento se asigna como el primer representante, es decir $c_1 := o_1$.

ii) El i-ésimo representante, con $i \in \{2, ..., m\}$, se asigna al elemento o_j con $j \in \{1, \ldots, N\}$ que maximiza la función:

$$\min_{k=1,\ldots,i-1} d(o_j, c_k).$$

esto es $c_i \in \arg\max_j \{\min_{k=1,\ldots,i-1} d(o_j, c_k)\}$, en donde $d(\cdot, \cdot)$ denota la medida de similaridad usada para comparar los objetos de O.

Nuestro problema de clasificación particular consiste en clasificar series temporales a través de sus funciones de distribución, el cual corresponde a una subclase de los denominados clasificación de procesos (*clustering processes*). Consideraremos N muestras de series temporales $\{X^1 = (X_1^1, \ldots, X_{n_1}^1), \ldots, X^N = (X_1^N, \ldots, X_{n_N}^N)\}$, las cuales queremos clasificar en $m \leq N$ grupos, con m un valor conocido. Para esto usamos el enfoque mencionado en [28]: Asumimos que cada una de las muestras fue generada por una de las m distribuciones de series temporales $F_1, \ldots, F_m$, las cuales pueden ser

desconocidas, así asignaremos dos series temporales en un mismo conjunto si y solo si sus funciones de distribución son las mismas, por lo cual necesitaremos definir una métrica sobre el espacio de las distribuciones de series temporales.

Una dificultad al momento de definir una medida sobre este espacio es calcularla explícitamente, puesto que si las distribuciones son desconocidas o están definidas de manera no paramétrica resulta imposible. Sin embargo, nos bastará con una estimación consistente de la medida de similaridad. Además, el algoritmo de clasificación que deseamos usar debe ser *asintóticamente consistente*, es decir, con probabilidad 1, a partir de cierto n, los conjuntos entregados por el algoritmo coinciden con los de la *clasificación objetivo*. La clasificación objetivo se define de acuerdo a si las muestras son generadas por la misma distribución o no, es decir, dos muestras pertenecen a un mismo conjunto si y sólo si fueron generadas por la misma distribución. El valor de n corresponde al largo de la muestra con menos elementos, es decir:

$$n := \min_{i=1,\dots,N} \{n_i\} \tag{2.4}$$

El *algoritmo del punto más lejano* es asintóticamente consistente para la clasificación de distribuciones de series temporales si la medida de similaridad considerada puede ser estimada consistentemente. En general existen diversas medidas de similaridad de distribuciones que cumplen con esta característica. Nosotros usamos la llamada *distancia telescópica* por sus propiedades e hipótesis con respecto a las distribuciones de las series temporales. Esta distancia fue introducida en [29], siendo su ventaja frente a las distancias usuales es que solo necesitamos que las distribuciones sean estacionarias ergódicas para poder estimarla de manera consistente, mientras que otras distancias necesitan usualmente que las distribuciones sean independientes e idénticamente distribuidas o markovianas.

Expondremos a continuación la construcción de la distancia telescópica de la misma forma presentada en [29].

Definición 15. Para dos distribuciones de probabilidad P y Q de variables aleatorias definidas sobre el espacio $(\Omega, \mathcal{F})$ y un conjunto $\mathcal{H}$ de funciones medibles en $(\Omega, \mathcal{F})$ se define la función:

$$d_{\mathcal{H}}(P, Q) = \sup_{h \in \mathcal{H}} |E_P(h) - E_Q(h)|$$

La cual es medible si el conjunto $\mathcal{H}$ es separable, es decir: existe un conjunto numerable $\mathcal{H}'$ de funciones tal que cualquier función en $\mathcal{H}$ es el límite puntual de una sucesión de elementos en $\mathcal{H}'$.

Esta función fue introducida por primera vez en [42] y satisface las propiedades de métrica bajo el supuesto que $d_{\mathcal{H}}(P,Q) = 0$ implica $P = Q$, el cual es nuestro caso de interés. Además, consideramos al conjunto $\mathcal{H}$ como el de las funciones indicadoras sobre Ω, en cuyo caso podemos identificar a cada función $h \in \mathcal{H}$ con el conjunto indicador $\{\omega : h(\omega) = 1\} \subset \Omega$ y, con abuso de notación, escribir:

$$d_{\mathcal{H}}(P,Q) := \sup_{h\in\mathcal{H}} |P(h) - Q(h)|$$

Bajo estas últimas condiciones se tiene el siguiente lema:

Lema 1. $d_{\mathcal{H}}$ es una métrica en el espacio de distribuciones de probabilidad sobre $(\Omega, \mathcal{F})$ si y solo si $\mathcal{H}$ genera a $\mathcal{F}$, es decir, $\mathcal{F}$ es la σ-álgebra más pequeña que hace a las funciones en $\mathcal{H}$ medibles.

De momento hemos definido una distancia para distribuciones de variables aleatorias, por lo que basándose en $d_{\mathcal{H}}$ (ver [29]), se construye la distancia telescópica definida para distribuciones de series temporales, de la siguiente manera: Para dos distribuciones de series temporales F_1, F_2 tomaremos la distancia $d_{\mathcal{H}}$ entre sus distribuciones finito dimensionales de tamaño k para cada $k \in \mathbb{N}$ y realizaremos la suma ponderada en k por pesos decrecientes de dichos valores.

Definición 16. (**Distancia Telescópica**) Para dos distribuciones de series temporales F_1 y F_2 sobre el espacio $(\Omega, \mathcal{F})$ y una secuencia de conjuntos de funciones $\mathbf{H} = (\mathcal{H}_1, \mathcal{H}_2, ...)$ se define la distancia telescópica como:

$$D_{\mathbf{H}}(F_1, F_2) = \sum_{k=1}^{\infty} W_k \sup_{h\in\mathcal{H}_k} |E_{F_1} h(X_1, ..., X_k) - E_{F_2} h(X_1, ..., X_k)|$$

donde W_k, $k \in \mathbb{N}$, es una secuencia sumable positiva de pesos decrecientes (por ejemplo $W_k = 1/k^2$ o $W_k = 2^{-k}$)

Notemos que la distancia telescópica pondera la distancia existente entre las distribuciones finito dimensionales de las series temporales. Es por esto que necesitamos el teorema de existencia de Kolmogorov (Teorema 1) y los resultados previos que caracterizan la distribución de la serie mediante las distribuciones de la serie truncada. El siguiente lema nos da una condición para que $D_{\mathbf{H}}(\cdot,\cdot)$ sea una métrica. Esto resulta

intuitivo considerando que dos distribuciones de series temporales son iguales si y solo si sus distribuciones finito dimensionales coinciden.

Lema 2. $D_{\mathbf{H}}$ es una métrica si y sólo si $d_{\mathcal{H}_k}$ es una métrica $\forall k \in \mathbb{N}$.

En la práctica no contamos con las distribuciones de las series temporales, solo con una observación muestral del proceso. Por lo cual es necesario definir una versión empírica de la distancia telescópica.

Definición 17. (**Distancia telescópica empírica**) Para un par de muestras $X_{1:n_x}$ e $Y_{1:n_y}$ se define la distancia telescópica empírica como:

$$\hat{D}_{\mathbf{H}}(X_{1:n_x}, Y_{1:n_y}) = \sum_{k=1}^{\min(n_x,n_y)} W_k \sup_{h\in\mathcal{H}_k} \left| \frac{1}{n_x-k+1} \sum_{i=1}^{n_x-k+1} h\left(X_{i:(i+k-1)}\right) - \frac{1}{n_y-k+1} \sum_{i=1}^{n_y-k+1} h\left(Y_{i:(i+k-1)}\right) \right|$$

Para analizar la consistencia asintótica de la distancia telescópica empírica debemos introducir los siguientes conceptos:

Definición 18. Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad y $\mathcal{C} \subset \mathcal{F}$. Se dice que la familia $\mathcal{C}$ es capaz de fragmentar a un conjunto $D \subset \Omega$ si la colección de subconjuntos:

$$\{C \cap D : C \in \mathcal{C}\}$$

contiene a todos los posibles subconjuntos de D, es decir:

$$|C \cap D : C \in \mathcal{C}| = 2^{|D|}$$

en donde $|D|$ denota la cardinalidad del conjunto.

Definición 19. Para una familia $\mathcal{C} \subset \mathcal{F}$ se define su dimensión de Vapnik-Chernovenkis o dimensión (VC) como la mayor cardinalidad que puede poseer un conjunto $D \subset \Omega$ tal que $\mathcal{C}$ es capaz de fragmentarlo.

La Figura 2.3 ejemplifica la fragmentación de un conjunto de tres elementos mediante hiperplanos. Se busca poder separar estos 3 puntos, suponiendo que cada uno puede pertenecer a dos grupos, en cada combinación posible mediante hiperplanos. En caso de que esto sea posible, el conjunto de hiperplanos posee, al menos, dimensión (VC) 3.

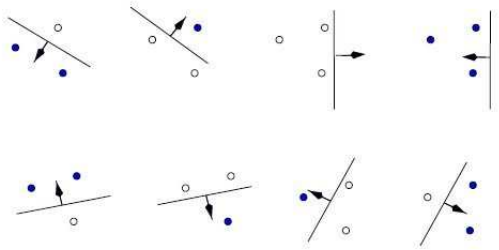


FIGURA 2.3: Fragmentación de tres puntos mediante hiperplanos, suponiendo que los puntos se distribuyen en dos clases diferentes.

Los resultados que se presentan a continuación son sobre la distancia empírica $\hat{D}_{\mathbf{H}}$. El siguiente teorema plantea que este es un estimador asintóticamente consistente de la distancia telescópica, lo que nos permitirá usarlo para obtener un algoritmo de clasificación asintóticamente consistente.

Teorema 2. Sea $\mathbf{H} = (\mathcal{H}_k)_{k\in\mathbb{N}}$ una secuencia de conjuntos separables de funciones indicatrices, donde cada $\mathcal{H}_k$ está definido sobre el espacio de medida $(\Omega, \mathcal{F})$. Supongamos que cada $\mathcal{H}_k$ genera a $\mathcal{F}$ y posee dimensión Vapnik Chervonenkis (VC) finita, es decir, es capaz de fragmentar un conjunto de cardinalidad finita. Entonces, para dos distribuciones de series temporales estacionarias ergódicas F_x, F_y que generan las muestras $X_{1\dots n_x}$ e $Y_{1\dots n_y}$ respectivamente se tiene:

$$\lim_{n_x,n_y\to\infty} \hat{D}_{\mathbf{H}}(X_{1\dots n_x}, Y_{1\dots n_y}) = D_{\mathbf{H}}(F_x, F_y) \quad c.s.$$

La condición de que los conjuntos $(\mathcal{H}_k)_{k\in\mathbb{N}}$ sean de funciones indicadoras y posean dimensión (VC) finita viene de un resultado enunciado por [1]. Ellos usan estas condiciones y la siguiente caracterización de distribuciones ergódicas enunciadas en [17]:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n-k+1} \mathbb{I}_{X_{i:i+k}\in A} = F(A); \quad F - cs \quad \forall A \in \mathcal{F}. \tag{2.5}$$

para demostrar que, para cualquier distribución estacionaria ergódica F, la expresión

$$\sup_{h\in\mathcal{H}_k} \frac{1}{n-k+1} \sum_{i=1}^{n-k+1} h(X_{i:(i+k-1)}) \tag{2.6}$$

denota a un estimador asintóticamente consistente de $\sup_{h \in \mathcal{H}_k} E_F h(X_{1:k})$. Esto nos dice que $d_{\mathcal{H}}$ y $D_{\mathbf{H}}$ pueden ser estimadas de forma consistente.

Además, en [29] se hace uso de la propiedad de *separabilidad estricta*, el cual es un concepto de los problemas de clasificación definido por [4]. Esta propiedad plantea que dos puntos de una misma clase están más cerca entre sí que de cualquier otro punto perteneciente a una clase diferente. Bajo esta hipótesis se plantean los siguientes dos teoremas, los cuales entregan el fundamento teórico para usar el algoritmo de clasificación del punto más lejano, en conjunto con la distancia telescópica para clasificar las series temporales a través de sus funciones de distribución.

Teorema 3. Sean los conjuntos $\mathcal{H}_k$, $k \in \mathbb{N}$ conjuntos separables de funciones indicadoras sobre $(\Omega, \mathcal{F})$. Si cada $\mathcal{H}_k$ tiene dimensión VC finita y genera a $\mathcal{F}$, además de que las distribuciones $F_1, ..., F_m$ que generan las muestras $X^1, ..., X^N$ sean estacionarias ergódicas. Entonces con probabilidad 1, para $n = \text{mín}_{i=1,...,N}\{n_i\}$, la clasificación objetivo posee la propiedad de separación estricta con respecto a la métrica $\hat{D}_{\mathbf{H}}$.

Teorema 4. Bajo las condiciones del teorema 3, el algoritmo del punto más lejano es asintóticamente consistente si se conoce el número m de clases.

2.4. Estimación de funciones de densidad

A continuación, plantearemos los fundamentos para realizar estimaciones de funciones de densidad. El objetivo de esto es, una vez clasificadas las series temporales, obtener información complementaria sobre su comportamiento a través de dicha densidad.

Sea F una función de distribución con función de densidad f. Supongamos que se tiene una muestra proveniente de F:

$$X_1, ..., X_n \sim F.$$

Queremos realizar una estimación no paramétrica de la función f, que denotamos por $\hat{f}_n$. En general el objetivo de las estimaciones no paramétricas es utilizar la menor cantidad de supuestos posibles, cumpliendo los siguientes tres requerimientos:

A) $\hat{f}_n$ es una función de densidad.

B) $\hat{f}_n$ satisface condiciones de regularidad sobre f, como continuidad, derivabilidad, etc.

C) $\hat{f}_n$ aproxima bien a f, según alguna distancia.

El histograma es uno de los métodos más simples para estimar una función de densidad, pero nos da una idea de los elementos importantes a la hora de construir mejores estimaciones. La construcción de un histograma consiste en cortar el espacio en el cual viven los datos en clases, por ejemplo, si los datos están en los reales, la recta numérica se corta en intervalos; luego se cuenta la cantidad de observaciones que están contenidas en cada clase, así la altura de cada una de las barras en el histograma es proporcional a la cantidad de elementos en cada clase. El largo h de cada intervalo se denomina *parámetro de suavizamiento* y dependiendo de su tamaño puede producir un efecto de sobre suavizamiento o de poco suavizamiento. Una desventaja del histograma es que no es una estimación suave, continua o diferenciable de la función de densidad, independiente de las hipótesis que se tengan sobre ella o las propiedades que se deseen.

Las estimaciones mediante funciones de kernel satisfacen estos requerimientos y no demandan muchos supuestos. Comenzamos enunciando los conceptos principales utilizados en esta técnica para luego definir el estimador de kernel.

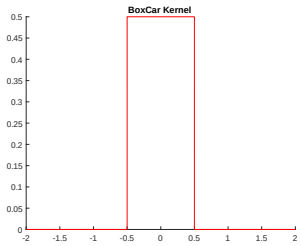
Definición 20. Sea $K : \mathbb{R} \to \mathbb{R}$ una función suave. Diremos que K es un kernel si:

i) $\int K(x)dx = 1$.

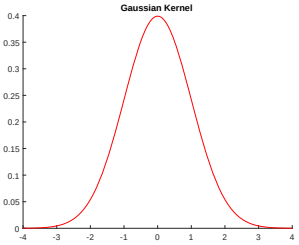
ii) $\int xK(x)dx = 0$.

iii) $\sigma_K^2 = \int x^2 K(x)dx > 0$.

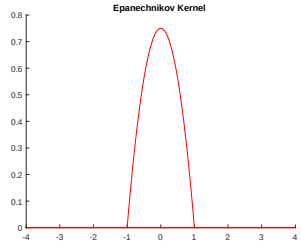
Algunas funciones de kernel clásicas en la literatura son BoxCar, Gaussiano, Epanechnikov y tricúbico, los cuales están ilustrados en la Figura 2.4.



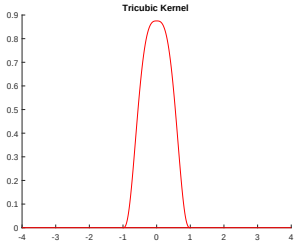
$$K_b(x) = \frac{i(x)}{2}$$



$$K_g(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$



$$K_e(x) = \frac{3}{4}(1 - x^2)i(x)$$



$$K_t(x) = \frac{70}{80}(1 - |x|^3)^3 i(x)$$

FIGURA 2.4: Gráfico de diferentes funciones de kernel, donde $i(x)$ es la función indicatriz del intervalo $[-1, 1]$

En general los kernel son usados para tomar promedios locales. Por ejemplo, supongamos que tenemos muestras de la forma $(x_1, y_1), ..., (x_n, y_n)$ y deseamos tomar un promedio sobre los y_i tales que los x_i asociados estén a lo más a distancia h de un punto x fijo, considerando un kernel K este promedio se puede escribir como:

$$\sum_{i=1}^{n} y_i \frac{K(\frac{x_i - x}{h})}{\sum_{i=1}^{n} K(\frac{x_i - x}{h})} \tag{2.7}$$

Definición 21. El estimador de kernel, con función de kernel K asociado a la muestra $X_1, \ldots, X_n$ y ventana de ancho h está definido por:

$$\hat{f}_h = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right); \quad x \in \mathbb{R}. \tag{2.8}$$

Este estimador satisface las condiciones deseadas para los estimadores de funciones de densidad. Sus propiedades generales son:

- $\hat{f}_h$ es una densidad.
- El estimador posee la misma regularidad que la función K, es decir, si K es continuo o diferenciable, también lo será $\hat{f}_h$.

La ventana (o bandwidth) h del estimador influye sobre la cantidad de observaciones a tomar en cuenta para estimar la función en un punto x, al igual que en el histograma. Si h es demasiado grande, el estimador es muy plano y demasiado suave (oversmoothing), en cambio sí es demasiado pequeño, el estimador es muy irregular y no suficientemente suave (undersmoothing).

Para los estimadores de kernel se tiene el siguiente resultado de consistencia:

Teorema 5. Sea una muestra iid $X_1, \ldots, X_n$ con función de densidad común f.Si f es continua en x y $h_n \to 0$, $nh_n \to \infty$ cuando $n \to \infty$, entonces:

$$\hat{f}_h(x) - f(x) \xrightarrow[n\to\infty]{} 0, \quad cs.$$

Bajo las mismas condiciones de regularidad, si $X_1, \ldots, X_n$ es una sucesión estacionaria, con una estructura de dependiente débil, es decir, la covarianza entre X_i y X_{i+k} converge rápidamente a 0 cuando $k \to \infty$, se obtiene el mismo resultado. Como ejemplo se pueden tomar modelos ARMA, GARCH, modelos markovianos (ver Capítulo 3.2.3).

En el caso d-dimensional, basta considerar una función de kernel d-dimensional. En particular, podemos considerar un kernel unidimensional cambiando la variable x por alguna normal d-dimensional.

Capítulo 3

Downscaling

En este capítulo discutimos las ideas básicas detrás de los métodos de downscaling, comenzando por la idea general de estas técnicas, continuando con un repaso de los modelos estadísticos usuales. Finalmente proponemos nuestro modelo de downscaling estadístico que consta de tres pasos, con el fin de aprovechar la estructura de dependencia entre datos a diferentes escalas de resolución, diferenciándose de las técnicas estadísticas clásicas.

En la actualidad el estudio de los fenómenos meteorológicos ha sido de gran importancia debido al comportamiento errático del clima en los últimos años, en donde el interés en la predicción de estados futuros, prevención de catástrofes relacionadas con fenómenos climatológicos, uso de energía renovable, entre otras, ha potenciado la investigación sobre los efectos que ejercen las distintas componentes climáticas globales y sus variaciones en localidades específicas como zonas urbanas, ganaderas, costas, etc; en donde cada una de ellas presenta sus particularidades geográficas.

Un factor importante a la hora de estudiar el comportamiento de un fenómeno climático es la escala sobre la que actúa. Por ejemplo, las corrientes de viento a alturas de 150[Km] parecen tener un comportamiento suave, mientras en zonas montañosa presentan fluctuaciones bruscas. Usualmente se definen escalas verticales y horizontales, en donde las primeras están asociadas a diferentes niveles de presión, temperatura, altura, etc. en la cual ocurre el fenómeno; y las horizontales hacen referencia al tamaño

de la zona afectada (latitud y longitud). En la figura 3.1 se puede apreciar un ejemplo de malla para estas escalas. Se suelen identificar tres grandes niveles principales de escalas horizontales: Escala Sinóptica o Global (≥20000[Km]), Meso Escala (entre 20000[Km] y 0.1[Km]) y Micro Escala (≤ 0.1[Km]) [24].

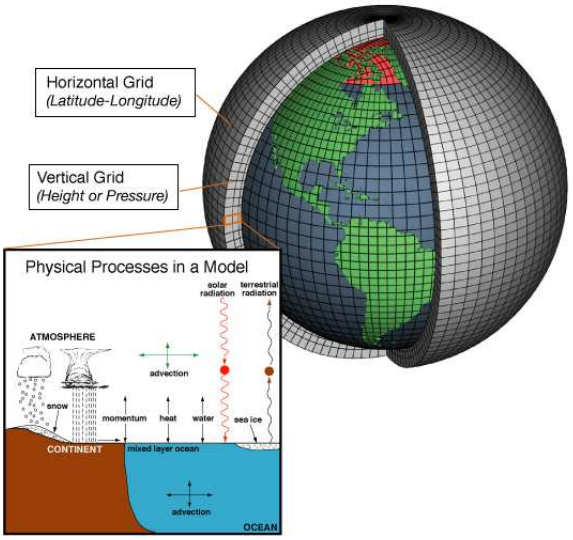


FIGURA 3.1: Visualización de las escalas verticales y horizontales[1]

Los procesos atmosféricos sobre los que se tiene mayor conocimiento y precisión a la hora de construir modelos acordes a las leyes físicas que los gobiernan son los que ocurren a nivel global; es decir, si pensamos sólo en escalas horizontales nos referimos a los que actúan sobre la escala Sinóptica. Los modelos de circulación global o general (GCM por su sigla en inglés) son los que buscan explicar estos fenómenos, caracterizándose por su gran precisión al momento de simular la atmósfera, los océanos, procesos bióticos y las diferentes interacciones y retroalimentación entre ellas. Los GCM consideran una malla, que representa las posiciones horizontales y verticales de la tierra, con una resolución que va desde los 100[Km] a los 500[Km] para su modelamiento. Cada componente de la malla posee información de interés para el modelo tales como: dirección o intensidad de viento, vapor de agua e interacción atmosférica entre nubes, efectos directos e indirectos de los aerosoles en la radiación y precipitación, cambios en las capas de nieve y hielo oceánico, el almacenamiento de calor en el suelo y océanos, flujos de calor y humedad, transporte de calor y agua por la atmósfera y océanos a gran escala [39]. Este tipo de modelo consiste, en general, en resolver ecuaciones en derivadas parciales del tipo Navier-Stokes con condiciones de borde expresadas en funciones simples. Por su naturaleza, no contemplan información

[1]National Oceanic and Atmospheric Administration (NOAA),2012

local o geográfica que afecta dicho flujo a escalas menores, pues el objetivo es entregar información y predicciones que son válidas a nivel Sinóptico, el cual puede contemplar países enteros, tal como se puede ver en la figura 3.2, lo que facilita la obtención de soluciones numéricas para estos modelos en términos de tiempo computacional y precisión.

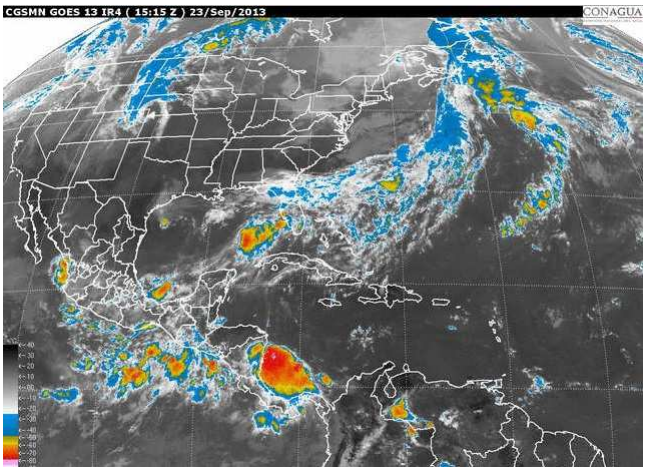


FIGURA 3.2: Ejemplo de fenómenos en escala Sinóptica [2]

A pesar de que el uso de los GCM para fenómenos globales es de gran aceptación, estos modelos en resoluciones espaciales más finas, donde las celdas de la malla son mucho más pequeñas, poseen una menor precisión en sus proyecciones [18]. Muchos ejemplos han ilustrado esta deficiencia en los GCM al simular fenómenos climáticos a nivel local, por ejemplo: [13] compara el desempeño de diferentes GCM a baja resolución en la cuenca Mediterránea y en [26] analizan el desempeño del modelo "Climate Comunity Model" (CC2) a resoluciones de 310[Km] y 125[Km] en la cuenca del río Sacramento en California, notándose en ambos casos la pérdida de eficiencia y precisión de estos modelos.

El hecho de que los GCM funcionen bien a nivel global, pero fallen a nivel local puede parecer una contradicción o poco intuitivo en un primer acercamiento, sin embargo hay que considerar que el comportamiento climático global es una gran respuesta extendida de las diferentes fuerzas/fenómenos solares, rotación terrestre y la estructura a gran escala de la superficie terrestre (topografía, distribución de los océanos, etc.), mientras que el comportamiento local es una respuesta del clima global con respecto

[2]Satélite Conagua, Septiembre 23 de 2013

a los detalles locales o regionales, los cuales agregan una gran cantidad de información.

Los principales motivos del mal comportamiento de los GCM a nivel local son la resolución espacial y la parametrización de las condiciones de borde para los fenómenos. Usualmente la resolución espacial en la que se trabajan estos modelos no proveen una descripción adecuada de la estructura de la superficie terrestre [21]. Un ejemplo de esto es el ciclo anual de precipitación de los Alpes: en el verano, en la zona noreste se presentan las mayores precipitaciones, mientras que en invierno, las mayores precipitaciones se presentan al suroeste [43], siendo los GCM ciegos ante esta diferencia. Además, la representación en los procesos de las submallas o fenómenos regionales como la formación de nubes, lluvias, infiltraciones, evaporación, entre otros, debe ser planteada con parámetros que se ajusten correctamente al modelo y a la nueva resolución, pues pueden ser una fuente de error importante para los estos modelos, pudiendo ser aún más influyentes que la falta de información debido a la resolución [26].

Debido a limitaciones en la capacitad de cómputo e inestabilidad de los métodos numéricos empleados en la aproximación de las soluciones, las resoluciones de las mallas en las que actúan los GCM no son lo suficientemente finas como para captar de manera precisa el comportamiento del fenómeno de forma local (ciudad, región, entre otros). Esto motiva el desarrollo de técnicas que sean capaces de traducir el comportamiento global del fenómeno en información sobre su comportamiento local, es decir, mediante una malla más gruesa (escala Sinóptica) de información se desea obtener una malla más fina (meso o micro escala) y de mayor información. Estas técnicas reciben el nombre de *downscaling*.

Al momento de realizar alguna técnica de downscaling hay que considerar que los fenómenos a gran escala varían de forma lenta y no brusca (suave). Si en el fenómeno se producen cambios a cortas distancias, entonces ocurre a nivel local o regional. La importancia de esta observación es que, al hablar de nivel local o regional no nos referimos necesariamente a considerar una zona geográfica más pequeña, sino que al mirar una zona considerando un mayor detalle geográfico. Por ejemplo, la cantidad de agua que cae por lluvia en un punto está determinada por la presencia local de nubes y sus características, más que por la formación a gran escala de nubosidad [6].

El downscaling puede realizarse de forma espacial (obtención de mallas espaciales más finas) y/o temporal (escala de tiempo más fina, como por ejemplo pasar de una información mensual a una semanal) y suelen clasificarse en dos tipos dependiendo de su naturaleza: Dinámico (o modelos anidados) los cuales al igual que los GCM buscan resolver las ecuaciones físicas que gobiernan a los fenómenos; y Estadísticos, que establecen relaciones empíricas entre el comportamiento local y el global.

3.1. Downscaling Dinámico

Estos métodos no consideran información histórica entre el comportamiento local y el global, si no que buscan, al igual que los GCM, representar las leyes físicas que gobiernan el comportamiento local mediante ecuaciones en derivadas parciales, pero considerando una malla más fina e información extra en sus celdas, como por ejemplo factores topográficos, que ayudan a obtener escenarios o proyecciones más precisas y detalladas a nivel local. Los modelos usados en el desarrollo de esta metodología reciben el nombre de climatológicos regionales (RCM) o modelos anidados, pues la malla sobre la que actúan esta encajonada en la de algún GCM, lo cual permite considerar la información proveniente del modelo a gran escala como condiciones de frontera al momento de resolver para el comportamiento local, tal como se muestra en la figura 3.3.

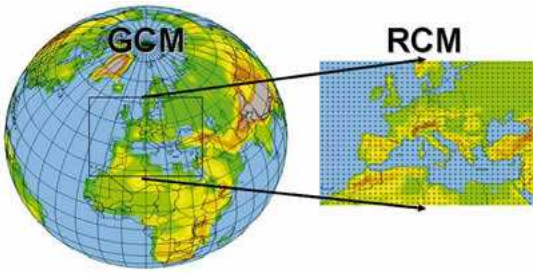


Figura 3.3: Descripción esquemática de un RCM o modelo anillado [3]

A pesar de que los RCM son desarrollados para trabajar en mallas de mayor resolución que los GCM, es importante realizar una reparametrización y no heredar la proveniente del GCM que proporciona las condiciones de frontera para el RCM, ya que en caso contario este hereda, además de los errores de precisión, errores sistemáticos que

[3]Regionalization of climate change information for impact assessment and adaptation, Filippo Giorgi

generan mayores sesgos en los resultados [30]. Otras consideraciones generales para los RCM son:

- Los parámetros del modelo son ajustados al estado climático actual, los cuales pueden dejar de ser válidos en situaciones atípicas o fenómenos naturales aleatorios que alteren el comportamiento climático, como por ejemplo erupciones volcánicas o terremotos. Esta característica es compartida con los GCM.
- Su naturaleza los hace computacionalmente costosos, lo cual dificulta comprobar hipótesis mediante simulaciones o generar diferentes escenarios.

Al momento de implementar un RCM, es recomendado considerar diferentes GCM para indicar las condiciones de borde, pues los resultados, como es de esperarse, pueden variar completamente. Esto genera una variedad de simulaciones para un mismo modelo, por lo que generalmente se considera un modelo mixto que tome en cuenta estos diferentes escenarios en busca de mejorar la aproximación y validar el ajuste con datos reales [25]. En la figura 3.4 se muestra resultados para un RCM en diferentes escenarios, donde cada escenario es representado por un GCM diferente. Se observa que diferentes parametrizaciones pueden arrojar resultados muy diferentes y por ello resulta importante realizar una validación con datos reales.

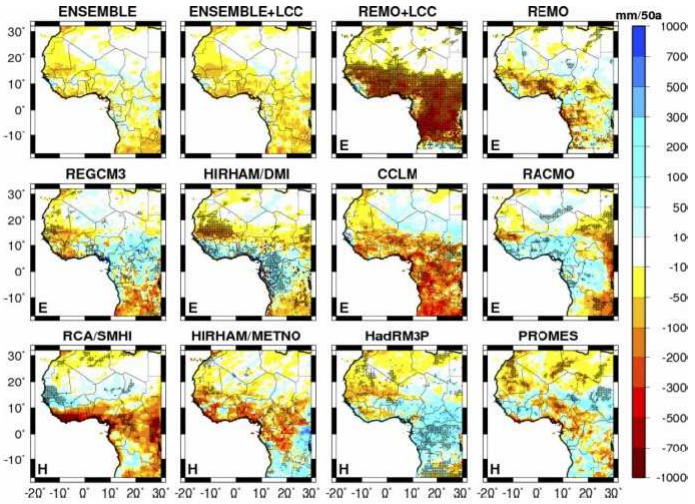


FIGURA 3.4: Resultados de un RCM para diferentes escenarios [4]

Usualmente estos modelos generan escenarios y predicciones para resoluciones que alcanzan el rango entre 50[Km] y 10[Km] de manera precisa y eficiente en ámbitos

[4]Progress in regional downscaling of west african precipitation, Paeth et al.

computacionales. Modelos más avanzados como el Weather Research and Forecasting Model (WRF) logran llegar hasta mallas de 1[Km], aumentando el costo computacional.

3.1.1. Weather Research and Forescasting(WRF) Model

El "Weather Research and Forescasting (WRF) Model" es desarrollado como un esfuerzo multidisciplinario para entregar un modelo de predicción a mesoescala y un sistema de asimilación de datos que permita avanzar hacia un mejor entendimiento y predicciones del clima a nivel de mesoescala y acelerar la comunicación entre los avances investigativos y su implementación en problemas prácticos.

Este modelo comenzó a ser desarrollado a finales de 1990 como una colaboración entre la división meteorológica a mesoescala y microescala del NCAR (NCAR Mesoscale and Microscale Meteorology (MMM)), la administración oceánica y atmosférica nacional de centros para la predicción ambientan y laboratorio de sistemas de estimación (National Oceanic and Atmospheric Administration's (NOAA) National Centers for Environmental Prediction (NCEP) and Forecast System Laboratory (FSL)), el departamento de la fuerza aérea de agentes climáticos y laboratorio de investigación naval (Department of Defense's Air Force Weather Agency (AFWA) and Naval Research Laboratory (NRL)), el centro de análisis y predicción de tormentas de la universidad de Oklahoma (Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma) y la administración federal de aviación (Federal Aviation Administration (FAA)), junto con investigadores de variadas universidades en el mundo.

El WRF está diseñado en su programación para ser eficiente en el uso de procesos en paralelo y tomar ventaja de computadores de alto rendimiento. Esto permite que el modelo pueda ser configurado tanto para fines investigativos o aplicaciones, ofreciendo diversas opciones para la física de diferentes fenómenos, siendo posible su uso en escalas que van desde algunos metros hasta cientos de kilómetros, dependiendo del lugar geográfico sobre el cual se implementa. El WRF permite a los investigadores producir simulaciones que reflejen datos reales obtenidos a través de análisis u observación o condiciones atmosféricas ideales. Actualmente el WRF posee una gran comunidad de usuarios con más de 30.000 usuarios registrados a través de 150 países, con workshops

y tutoriales que se realizan al menos una vez al año en el NCAR. Este modelo es altamente usado en estimaciones a tiempo real a nivel mundial.

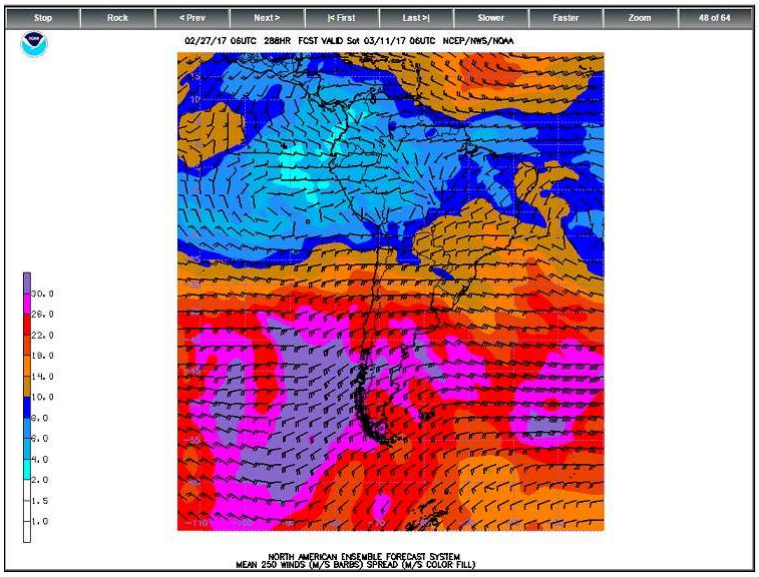


FIGURA 3.5: Predicción para América del sur y el mar del caribe de la velocidad del viento [5]

El WRF ofrece dos métodos dinámicos para calcular las ecuaciones que gobiernan los fenómenos atmosféricos, además de variantes del modelo conocidas como WRF-ARW (Advanced Research WRF) y WRF-NMM (Nonhydrostatic Mesoscale Model). El primero es apoyado a través de la comunidad de la división meteorológica a mesoescala y microescala del NCAR, mientras que el WRF-NMM está basado en el Modelo Eta y en el "Nonhydrostatic Mesoscale Model"desarrollado por el NCEP.

Los datos que se obtienen mediante este modelo poseen diferentes niveles de resolución espacial, que sirven para visualizar mayor detalle de las zonas estudiadas, tal como se muestra en la Figura 3.6. La relación existente entre las resoluciones de cada nivel es de uno a tres, por ejemplo, si el nivel mayor posee una resolución espacial de 9 [Km], el siguiente tiene una resolución de 3 [Km] y el más fino una de 1 [Km], dotando a los datos de diferentes niveles características de dependencia geométrica.

[5] http://mag.ncep.noaa.gov/model-guidance-model-area.php

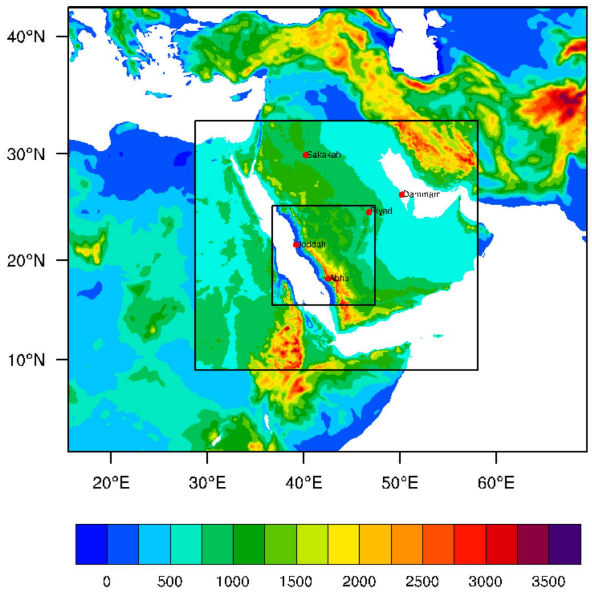


FIGURA 3.6: Estructura de las simulaciones del WRF [6]

3.2. Downscaling Estadístico

Estas técnicas consisten en establecer una relación empírica (observada) entre los fenómenos que se producen a gran escala y las características locales climáticas, es decir, a diferencia de los GCMs y las técnicas de downscaling dinámico, aprovechan la información histórica existente. Una vez que esta relación es validada, la información otorgada por el GCM para el fenómeno a nivel global es usada para obtener información de un fenómeno a nivel local, es decir, consideramos resultados provenientes de un GCM o RCM como variables predictoras y los valores locales, los cuales serán determinados mediante la relación de los fenómenos, como las variables dependientes o predictantes.

El downscaling estadístico, debido a su naturaleza, permite obtener mallas de mayor resolución que las otorgadas por los GCM en un tiempo menor que los RCM, pues suelen ser computacionalmente económicos y fáciles de implementar, pues la relación empírica entre los fenómenos suele ser simple de modelar o resolver. Generalmente se consideran técnicas estadísticas para este tipo de downscaling, las cuales varían dependiendo del objetivo.

[6] https://nar.ucar.edu/

Los supuestos transversales para el downscaling estadístico son los siguientes [33]:

1. La relación estadística establecida no cambia a través del tiempo (condición de estacionaridad), esta es necesaria para poder estimar el modelo a partir de una sola observación temporal.

2. El predictor capta los cambios climáticos, es decir, es un buen representante de las variaciones en los fenómenos climatológicos. Por ejemplo, usar como predictor un fenómeno que es estable ante las alteraciones climáticas no es válido.

3. La relación entre predictor y predictante es lo suficientemente fuerte como para ser usada, la dependencia se supone estadísticamente significativa.

4. El modelo de entrada, que puede ser un GCM o RCM, se ajusta con precisión al predictor.

El punto débil de estas técnicas es la condición de estacionaridad, pues ésta en la realidad no está garantizada y suele no satisfacerse. Sin embargo, si las series temporales que sirven como datos de entrenamiento del modelo son lo suficientemente largas, se puede asumir que el modelo contempla una cantidad considerable de escenarios estacionarios posibles para el fenómeno. Lo anterior no asegura que el modelo pueda ser adecuado para situaciones atípicas o si la relación empírica cambia en el tiempo, siendo este problema el análogo al de la parametrización presente en los GCMs y RCMs.

Gráficamente se puede pensar que los GCM, RCM y los modelos de downscaling estadísticos son diferentes niveles de amplificación para mirar características meteorológicas de una zona geográfica, como se muestra en la Figura 3.7. Generalmente se suele aplicar un RCM como primer filtro de información para luego visualizar en mayor resolución una zona particular mediante la aplicación de una técnica estadística. Estos modelos que contemplan estas combinaciones se denominan downscaling empírico - estadístico (empirical-statistical downscaling). La Figura 3.7 representa el uso conjunto de estas técnicas, en donde primero bajamos de una malla global a una intermedia, hasta obtener un nivel de detalle en el cual es posible distinguir las características geográficas.

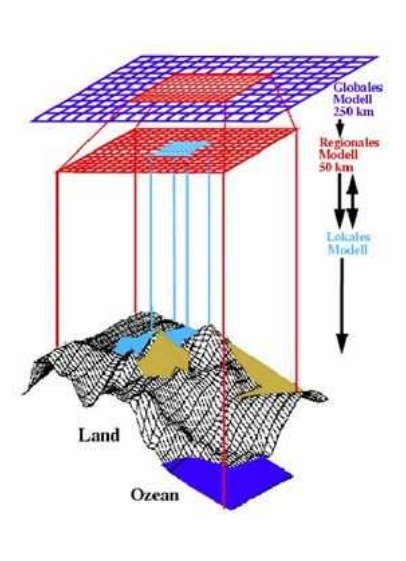


FIGURA 3.7: Relación entre GCM, RCM y técnicas de downscaling estadístico [7]

Las técnicas clásicas de downscaling estadístico suelen clasificarse en tres grandes grupos, los cuales detallaremos a continuación con algunos de sus métodos estándar.

3.2.1. Métodos Lineales

Estos métodos establecen una relación lineal o proporción entre el fenómeno climático sobre el cual se desea obtener información y las variables con las cuales se quiere explicar éste (información topográfica, fenómenos subyacentes, entre otros). Suelen estar asociadas al dowscaling espacial siendo las técnicas más comúnmente usadas:

a) **Métodos Delta:** Estos métodos buscan relacionar el estado futuro del fenómeno mediante un estado pasado conocido del mismo, el cual suele ser el actual. Consideran que, a cualquier nivel de escala, un valor futuro puede ser obtenido mediante la relación:

$$FC^f = \delta \cdot FC^p, \tag{3.1}$$

donde FC^f es el estado futuro del fenómeno que se desea estimar, FC^p es el estado pasado o actual y δ es el factor de cambio, siento todos valores unidimensionales. La hipótesis fundamental del método delta es suponer que el factor de cambio es invariante ante los cambios de escala, por lo que usan información empírica o de algún modelo para calcularlo y usarlo de manera local. Esquemáticamente el método se describe como:

[7] http://wiki.bildungsserver.de/klimawandel/index.php/Regionale_Klimamodelle

1. Mediante un estado pasado del fenómeno global (FC_g^p) y la predicción del comportamiento futuro dada algún GCM o RCM (FC_g^f) usando dicho estado pasado, calculamos δ como:

$$\delta = FC_g^f / FC_g^p. \tag{3.2}$$

2. Luego con una medición del comportamiento actual del fenómeno local (FC_l^p) y δ calculado en el paso anterior, obtenemos el valor del comportamiento futuro local (FC_l^f) por medio de la ecuación (3.1).

Este modelo posee las siguientes limitantes:

- El fenómeno de interés debe ser el mismo en ambas escalas.
- El factor de cambio δ se considera homogéneo en toda la zona de aplicación (homogeneidad local).
- El GCM o RCM usado para predecir el comportamiento futuro del fenómeno a nivel global debe captar de mejor manera los cambios relativos (δ) que los valores absolutos [19].

b) **Regresión Lineal:** Estas técnicas suponen que la relación entre los valores de los fenómenos atmosféricos a nivel global $X_1, ..., X_p$ (predictores) y el valor a nivel local Y(predictante) es de la forma:

$$Y = b_0 + b_1 X_1 + ... + b_p X_p + \epsilon, \tag{3.3}$$

donde ϵ es un error aleatorio asociado a las mediciones realizadas. El objetivo es estimar los coeficientes $b_0, ..., b_p$ mediante un método estadístico, donde usualmente es el método de los mínimos cuadrados.

Esta técnica usa N observaciones del valor local y de los fenómenos a nivel global, denotados por $Y^{(i)}$, $X_j^{(i)}$ con $i \in \{1, ..., N\}$, $j \in \{1, ..., p\}$ respectivamente. De esta forma podemos escribir estimadores para el error de medición en la forma:

$$e^{(i)} = Y^{(i)} - (b_0 + b_1 X_1^{(1)} + ... + b_p X_p^{(i)}). \tag{3.4}$$

Así los estimadores $\hat{b_0}, ..., \hat{b_p}$ para los coeficientes de la regresión se obtienen al resolver el siguiente problema de minimización:

$$\min_{b_0,\ldots,b_p} \sum_{i=1}^{n} (e^{(i)})^2, \tag{3.5}$$

lo que permite obtener una aproximación para los valores a nivel local dada por $\hat{Y}^{(i)} = \hat{b}_0 + \hat{b}_1 X_{(i)} + \ldots + \hat{b}_p X_p^{(i)}$.

La desventaja principal de este método es su sensibilidad a observaciones atípicas (fuera de la masa en la cual se concentra la mayor cantidad de datos), es decir, si en los datos para obtener los estimadores de los coeficientes hay valores correspondientes a eventos climatológicos poco usuales que alteren de forma brusca el comportamiento del fenómeno de interés, la estimación se verá afectada notablemente.

Sin embargo este método es útil y ampliamente usado cuando se puede suponer que la relación entre Y y las variables $X_1, \ldots, X_p$ es suficientemente regular a escalas locales como para ser aproximada linealmente.

Existen diversos indicadores para evaluar y comparar la calidad entre diferentes ajustes de regresión lineal, siendo el error cuadrático medio (ECM) y coeficiente de determinación o R^2 unos de los más reconocidos. El ECM mide el error cuadrático promedio producido por el modelo o que tan lejos están los datos en promedio del ajuste de regresión. El ECM se calcula como:

$$ECM = \frac{1}{N} \sum_{i=1}^{N} \left(Y^{(i)} - \hat{Y}^{(i)} \right)^2,$$

mientras que el coeficiente de determinación indica que porcentaje de la varianza presente en los datos es capaz de explicar el ajuste de regresión lineal, el cual se calcula como:

$$R^2 = \frac{\sum_{i=1}^{N} \left(\hat{Y}^{(i)} - \overline{Y} \right)^2}{\sum_{i=1}^{N} \left(Y^{(i)} - \overline{Y} \right)^2}.$$

donde $\overline{Y}$ es el promedio de los valores a nivel locales originales. Estos dos indicadores permiten contrastar el desempeño del modelo en diferentes conjuntos de datos.

c) **Análisis de correlación canónico:** El análisis de correlación canónico (CCA por su sigla en inglés) es un método para medir la relación lineal entre dos variables multidimensionales que llamaremos X e Y. La idea es encontrar combinaciones lineales de X e Y, tal que la correlación lineal con respecto al tiempo entre ellas sea máxima [9], mediante la búsqueda de patrones espaciales coherentes. El resultado de este proceso es un campo de correlación que indica como varia el comportamiento del fenómeno local según los cambios producidos a nivel global.

Para ello consideramos las variables $x = X^\top w_x$, $y = Y^\top w_y$, donde w_x, w_y son vectores escalares por determinar y x, y se denominan *variables canónicas*. Así la función de correlación ρ debe ser maximizada con respecto a w_x y w_y, asumiendo la siguiente forma:

$$\begin{aligned}
\rho(w_x, w_y) = \frac{\mathbb{E}[x^\top y]}{\sqrt{\mathbb{E}[|x|^2]\mathbb{E}[|y|^2]}} &= \frac{\mathbb{E}[w_x^\top X Y^\top w_y]}{\sqrt{\mathbb{E}[w_x^\top X X^\top w_x]\mathbb{E}[w_y^\top Y Y^\top w_y]}}, \\
&= \frac{w_x^\top \mathbb{E}[X Y^\top] w_y}{\sqrt{w_x^\top \mathbb{E}[X X^\top] w_x w_y^\top \mathbb{E}[Y Y^\top] w_y}}, \\
&= \frac{w_x^\top C_{xy} w_y}{\sqrt{w_x^\top C_{xx} w_x w_y^\top C_{yy} w_y}}.
\end{aligned}$$

Los vectores $(\hat{w}_x, \hat{w}_y)$ para los cuales ρ es máximo determinan la combinación lineal en la cual la correlación entre los fenómenos es máxima. La estimación $\hat{\rho} = \rho(\hat{w}_x, \hat{w}_y)$ se denomina *correlación canónica*.

En resumen, este método encuentra un par de variables multidimensionales, tal que al escribir X e Y en función de ellos, la información que se puede obtener de una a partir de la otra es máxima. Esto permite, si se considera X como el fenómeno a nivel global e Y como el fenómeno a nivel local, obtener la mayor cantidad de información posible para Y desde X por medio de sus combinaciones lineales, generando un mapa espacial que determina la dependencia temporal.

Algunas de las técnicas estadísticas utilizadas en un CAA son: Análisis de componentes principales, descomposición en valores singulares (SVD), análisis de discriminante y variaciones de estos métodos.

3.2.2. Métodos de clasificación del clima

En estos métodos, los datos empíricos o los obtenidos de un GCM se clasifican según características comparables, la cual puede ser geográfica o propia del fenómeno, en diferentes clases o estados, en donde dos elementos pertenecen a una misma clase siempre y cuando sus características sean las mismas bajo cierto criterio de similaridad. Esto permite estimar el comportamiento de la variable local según la clase que se le asocie bajo dicho criterio.

a) **Método Análogo:** Este método consiste en comparar los datos actuales con los datos históricos con el fin de encontrar el estado más cercano a él (estado análogo) bajo cierto criterio que puede ser dado por una distancia entre los dos objetos o alguna medida de similaridad. Luego, el comportamiento de la variable local según el estado análogo es la predicción para la situación local actual [43].

   Las desventajas de este método son, en primer lugar, la incapacidad de obtener estimaciones para valores futuros que estén fuera del rango histórico, además de requerir de una gran cantidad de observaciones historias, pues en caso contrario la implementación de este método no es válida.

b) **Análisis de Clúster:** El Clustering o análisis de clústers se basa en intentar responder cómo un conjunto de objetos pertenecen a un cierto número de clases o grupos, de manera que dos objetos que pertenecen a una misma clase comparten ciertas características comunes. Esta división usualmente está oculta, por lo que el objetivo principal es descubrir cuáles son estas clases y qué objetos las componen.

Un clustering de un conjunto $O = \{O_1, ..., O_N\}$ no vacío es una partición de este en $1 \leq m \leq N$ subconjuntos $\{C_1, ..., C_m\}$ tales que:

- $C_i \bigcap C_j = \emptyset, \quad \forall i \neq j.$
- $\bigcup_{i=1}^{m} C_i = O.$

A la clasificación verdadera de los elementos de O se le denomina *target clustering* (ver sección 2.3 para más información). Un algoritmo de clasificación bastante usado, además del punto más lejano mencionado en el Capítulo 2, es el llamado *enlace promedio*, el cual utiliza la distancia entre clases, definida como el promedio de las distancias entre todos sus puntos, para realizar el clustering y está definido por los siguientes pasos:

i) Colocar a cada elemento O_i, $i \in \{1, ..., N\}$ en una clase diferente.

ii) Unir las dos clases más cercanas en una sola, de forma iterativa, hasta obtener m.

c) **Otros métodos:** Existen diversos métodos de clasificación como análisis de redes neuronales ([12],[35]), machine learning ([34], [41]) variantes de kriging ([36],[37]), entre otros. Cualquiera de estos métodos puede ser utilizado, dependiendo su elección del objetivo y las características particulares del problema.

3.2.3. Simulación de Clima

Estos métodos buscan simular secuencias temporales del fenómeno mediante estimaciones de su media y varianza obtenidas a partir de información empírica o dada por algún GCM o RCM. Existen variados métodos adaptados para simular diversos fenómenos climáticos, donde una técnica estadística usual es el modelo de Markov oculto.

a) **Modelo de Markov Oculto (HMM):** Esta técnica busca representar la distribución de probabilidad del proceso $\mathbf{X} = (X_t, t \in \mathcal{T})$ con $\mathcal{T} \subset \mathbb{Z}$ que caracteriza al fenómeno climático bajo las siguientes tres propiedades que definen a este tipo de modelo:

i) La observación en tiempo t es generada por un proceso $\mathbf{S}$ cuyo estado en dicho instante, denotado por S_t es desconocido u oculto para el observador.

ii) Dados los estados anteriores del proceso oculto $\{S_{t_i}\}_{i=1}^{n-1}$, el estado actual S_{t_n} solo depende del estado anterior $S_{t_{n-1}}$. Esta es la denominada *propiedad de Markov*.

iii) La observación en el tiempo X_t solo depende del estado del proceso en tiempo t, es decir, depende solo de S_t.

Lo anterior nos permite tener la siguiente forma de la distribución del proceso:

$$P(S_{t1:T}; X_{t1:T}) = P(S_{t_1})P(X_{t_1}|S_{t_1}) \prod_{i=2}^{T} P(S_{t_i}|S_{t_{i-1}})P(X_{t_i}|S_{t_i}).$$

En este contexto, la información empírica o proveniente GCM o RCM corresponde a la serie temporal $\mathbf{X} = (X_t, t \in \mathcal{T})$ y los estados del proceso oculto $\mathbf{S} = (S_t, t \in \mathcal{T})$ corresponden al comportamiento del fenómeno climatológico que se desea simular, los cuales se asumen markovianos [27].

b) **Modelos de Series temporales:** Estos modelos describen el estado actual o futuro del fenómeno mediante información pasada, la cual puede ser obtenida mediante observación o estimaciones. Tres modelos básicos de series temporales son los autoregresivos (AR), los de media movil (MA) y los autoregresivos integrados de media móvil (ARIMA).

Un modelo es autoregresivo si la variable endógena al tiempo t es explicada por las observaciones de ella misma correspondientes a periodos anteriores más un término de error. Estos modelos se abrevian como AR(p), donde p corresponde al *orden del modelo*, el cual indica la cantidad de observaciones pasadas usadas para calcular la actual. Un ejemplo de estos modelos es el AR(2), el cual está dado por la relación:

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t.$$

Un modelo de media móvil es aquel que explica el valor de una variable determinada en un instante de tiempo t en función de un término independiente y una sucesión de errores correspondientes a periodos precedentes, ponderados convenientemente. Estos modelos se denotan por MA(q), donde q es el orden

del modelo. Por ejemplo un modelo MA(1) es de la forma:

$$X_t = \phi_0 + \phi_1 \epsilon_{t-1}.$$

En general se prefieren modelos de ordenes bajos o bien coincidentes con la periodicidad del fenómeno estudiado o de los datos de la serie (si es trimestral usamos orden 4, semestral usamos orden 6).

Un modelo autoregresivo integrado de media móvil contiene una parte autoregresiva y una parte de media móvil, donde además se incluye un parámetro d, el cual muestra la cantidad de diferenciaciones que se deben aplicar para que la serie analizada sea estacionaria. Estos modelos se abrevian como ARIMA(p,d,q), donde p indica el orden de la parte autoregresiva, q el orden de la parte media móvil y d que indica la cantidad de diferenciaciones o aplicaciones del operador de diferencias $\triangle$, descrito a continuación junto con la ecuación general de estos modelos:

$$X_t = -\left(\triangle^d X_t - X_t\right) + \phi_0 + \sum_{i=1}^{p} \phi_i \triangle^d Y_{t-1} - \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t.$$

donde $\triangle Y_t = Y_t - Y_{t-1}$.

El modelo ARIMA es uno de los modelos más usados, puesto que tiene en cuenta la hipótesis de estacionalidad para realizar downscaling.

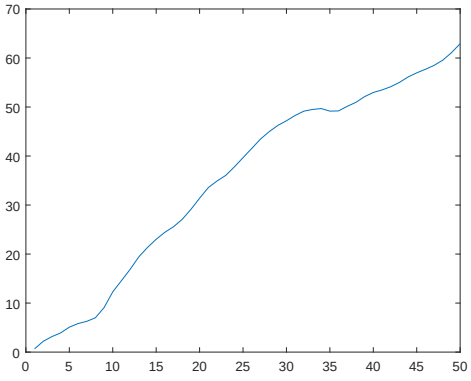


FIGURA 3.8: Trayectoria de un modelo ARIMA(1,1,1)

iii) **Otros Modelos:** Otras técnicas usadas para realizar simulaciones de clima son los campos gaussianos [20], los cuales son ampliamente usados en geoestadística,

los modelos autoregresivos con cambio de régimen markoviano o MSAR [23] y las ecuaciones diferenciales estocásticas [7].

3.3. Modelo de Downscaling Geométrico

En esta sección describiremos el modelo de downscaling que se propone, el cual consiste en el uso sucesivo de algunas de las técnicas descritas, buscando obtener la mayor cantidad de información de los datos utilizados. Comenzamos con aspectos formales como algunas definiciones previas, luego presentaremos el modelo desde un punto de vista global, describiendo las ideas detrás de cada uno de los pasos, finalizando con las técnicas de estimación que consideramos más adecuadas para cada uno de ellos. Denotamos por:

- $M^{(b)}$ a la malla definida sobre el espacio de resolución b [Km], es decir, cada punto de la malla está a b [Km] de distancia con los puntos verticales y horizontales contiguos, como se muestra en la Figura 3.9.

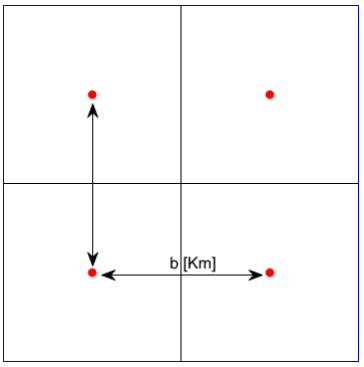


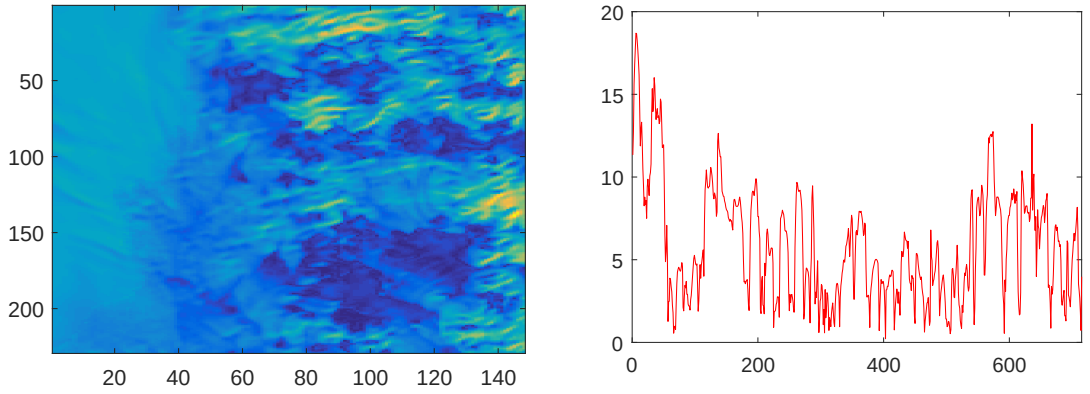
FIGURA 3.9: Separación entre elementos de una malla

- $X^{(b)}$ al campo de intensidad de viento observado sobre $M^{(b)}$ durante un horizonte de tiempo discreto $\mathcal{T} = \{1, ..., T\}$, es decir;

$$X^{(b)} = \left\{ x^{(b)}_{(s,t)} \in \mathbb{R}^{+}_{0} : s \in M^{(b)}, t \in \mathcal{T} \right\}, \tag{3.6}$$

donde $x^{(b)}_{(s,t)}$ corresponde a una observación de la intensidad del viento en la posición $s \in M^{(b)}$ en el tiempo $t \in \mathcal{T}$. Podemos interpretar s como un par ordenado (i, j) el cual denota las coordenadas horizontales y verticales en el espacio.

Cada elemento $x^{(b)}_{(\cdot,t)}$ corresponde a una observación del campo de intensidad del viento en el instante t, y $x^{(b)}_{(s,\cdot)}$ corresponde a una serie temporal en la posición s. La Figura 3.10 muestra ejemplos de estos elementos.



(a)Realización del campo de intensidad del viento $x^{(b)}_{(\cdot,t_0)}$ para un tiempo fijo t_0.

(b)Trayectoria de la intensidad del viento $x^{(b)}_{(s_0,\cdot)}$ en la posición s_0.

FIGURA 3.10: Ejemplos de realizaciones para $x^{(b)}_{(\cdot,t_0)}$ y $x^{(b)}_{(s_0,\cdot)}$.

Nuestro objetivo es desarrollar un método que permita obtener de forma rápida, mediante $X^{(1)}$ y $X^{(3)}$, una buena estimación para $X^{(0,5)}$ la cual denotamos $\hat{X}^{(0,5)}$, donde:

$$\hat{X}^{(0,5)} = \left\{ \hat{x}^{(0,5)}_{(s,t)} : s \in M^{(0,5)}, t \in \mathcal{T} \right\},$$

siendo $\hat{x}^{(0,5)}_{(s,t)}$ una estimación para $x^{(0,5)}_{(s,t)}$. Los campos $X^{(1)}$ y $X^{(3)}$ son obtenidos mediante simulaciones del modelo WRF, el cual puede realizar estimaciones de las variables atmosféricas para instantes de tiempo pasados.

Proponemos un modelo para $X^{(0,5)}$ que considere la información proveniente desde $X^{(1)}$ (el uso de $X^{(3)}$ será explicado más adelante), pues es el campo de intensidad conocido que contiene mayor cantidad de información. Nuestro modelo se obtiene desde la siguiente descomposición:

$$\begin{aligned} X^{(0,5)} &= \mathbb{E}\left[X^{(0,5)}|X^{(1)}\right] + \left(X^{(0,5)} - \mathbb{E}\left[X^{(0,5)}|X^{(1)}\right]\right), \\ &= I\left(X^{(1)}\right) + R_{(0,5:1)}, \end{aligned}$$

donde:

- $I(\cdot)$ es llamada función de interpolación estocástica definida mediante la esperanza condicional
$$I\left(X^{(l')}\right) = \mathbb{E}\left[X^{(l)}|X^{(l')}\right],$$
la cual es interpretada como el mejor predictor lineal de $X^{(l)}$ dada la información de $X^{(l')}$. El mejor predictor lineal se define como la combinación lineal de elementos en $X^{(l')}$ que minimiza la distancia L_2 a la menor σ-álgebra que contiene a $X^{(l)}$. Este predictor representa la mayor cantidad de información que contiene $X^{(l')}$ de $X^{(l)}$.

- $R_{(l:l')}$ es el error o residuo producido al estimar $X^{(l)}$ mediante $I\left(X^{(l')}\right)$.

En general no se conoce una forma explícita para la función $I(\cdot)$. Por esta razón es necesario construir una estimación de la función de interpolación que denotaremos por $\hat{I}(\cdot)$ y así proponer como estimador de $X^{(l)}$ a $\hat{X} = \hat{I}\left(X^{(l')}\right)$.

Para estimar $I(\cdot)$ debemos imponer algunas condiciones de regularidad; por ejemplo al suponer que I es localmente derivable, es posible estimarla localmente mediante funciones lineales y usar el método de regresión lineal para obtener el estimador $\hat{I}$.

La estimación $\hat{I}$ produce un estimador para el residuo, dado por la descomposición

$$X^{(l)} = \hat{I}\left(X^{(l')}\right) + \hat{R}_{(l:l')},$$

donde $\hat{R}_{(l:l')}$ contiene información que nuestro modelo no es capaz de explicar y posee características aleatorias, las cuales caracterizamos mediante las distribuciones de las series temporales. Para ello supondremos las siguientes condiciones sobre el residuo: estacionaridad ergódica, la cual permite estimar las distribuciones mediante una sola observación (como se mencionó en el Capítulo 2.3) y homogeneidad espacial local, es decir, en vecindades espaciales suficientemente pequeñas las distribuciones de las series temporales son similares. Este último supuesto está ligado a que los campos de viento son globalmente continuos, pero no regulares o suaves.

Caracterizamos la similaridad entre las distribuciones de las series temporales mediante la distancia telescópica. A partir de la distancia telescópica definimos zonas o clases de homogeneidad en el espacio de las distribuciones de las series temporales de los

residuos, y en cada clase realizamos estimaciones consistentes de dichas distribuciones estacionarias. Lo anterior nos permite caracterizar las zonas de homogeneidad, evaluar la calidad de nuestra función de interpolación mediante, por ejemplo, intervalos de confianza, además de implementar métodos de validación para el modelo.

A continuación explicaremos nuestra propuesta para estimar la función de interpolación $I(\cdot)$.

3.3.1. Estimación de la función de interpolación: Metodología de Regresión Lineal Local

A continuación procederemos a describir el esquema con el cual estimamos la función de interpolación $I(\cdot)$, a través de los campos de intensidad de viento a 1 [Km] y 3 [Km]. Consideramos las siguientes hipótesis de trabajo, las que provienen de las características del problema y la naturaleza de los datos:

- Regularidad y homogeneidad local: Debido a las características topográficas no hay regularidad global en el comportamiento de la intensidad del viento, pero para vecindades espaciales lo suficientemente pequeñas, los campos de viento presentan la regularidad necesaria como para realizar interpolación espacial mediante funciones lineales, que permitan aproximar la intensidad del viento en un punto a través de las mediciones de la intensidad para sus vecinos más cercanos.

  Las vecindades se definen en cada malla, mediante cuatro puntos más cercanos, formando cuadrados disjuntos que solo coinciden en sus bordes, es decir, no se asolapan. La Figura 3.11 muestra una vecindad fija $V = \{s_1, s_2, s_3, s_4\}$ para la malla $M^{(3)}$, donde sus elementos se caracterizan mediante puntos rojos.

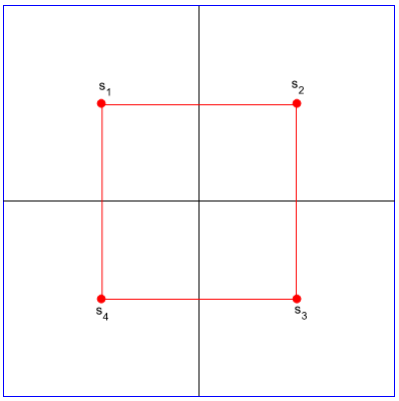


FIGURA 3.11: Representación de la vecindad $V = \{s_1, s_2, s_3, s_4\} \subset M^{(3)}$. Sus elementos son caracterizados mediante puntos rojos.

La Figura 3.12 muestra los diagramas de dispersión de una de vecindad de la malla a 3 [Km]. Cada gráfico fuera de la diagonal muestra la relación existente entre dos puntos, mostrando una dependencia lineal entre las trayectorias. En la diagonal se muestran los histogramas de cada serie temporal, los que resultan similares, justificando la hipótesis de que la distribución de los datos espaciales es homogénea localmente.

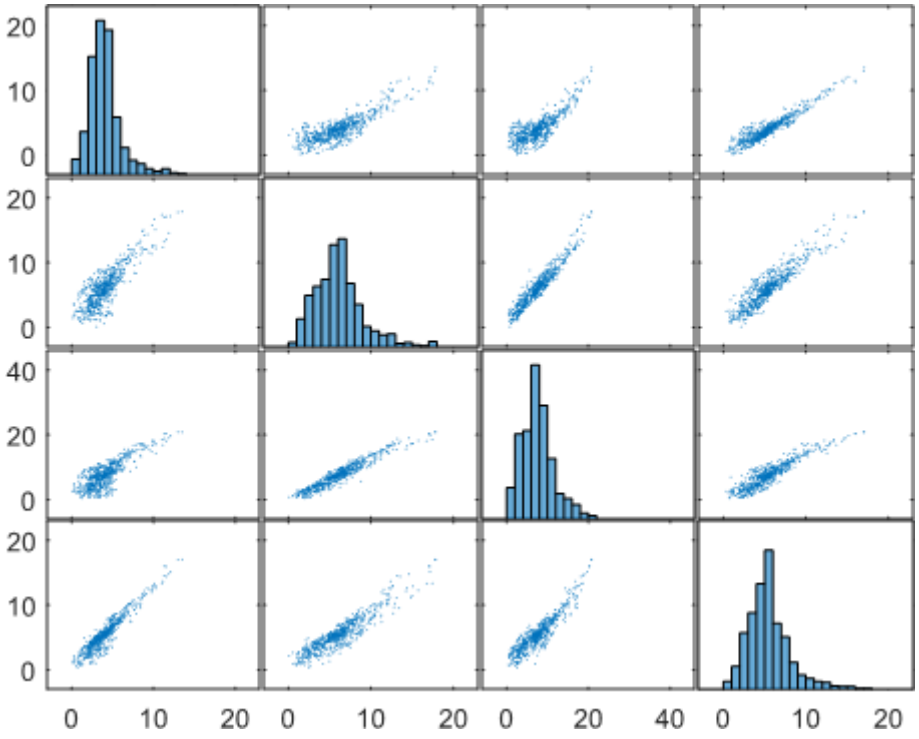
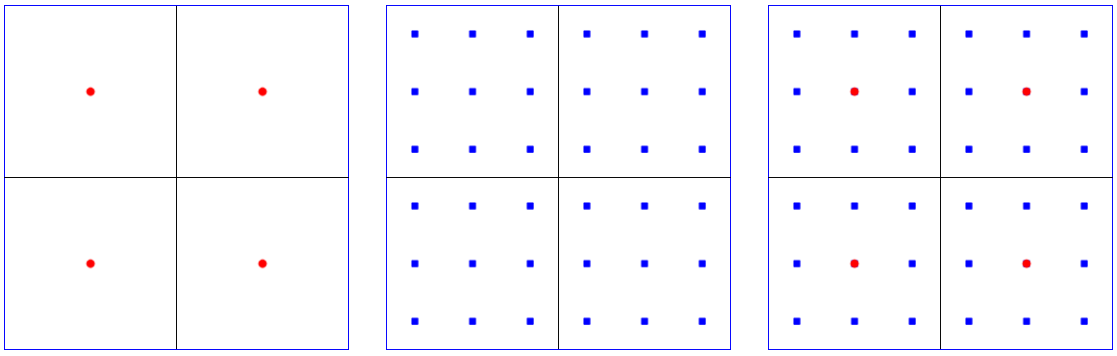


FIGURA 3.12: Diagrama de dispersión para las trayectorias de intensidades de viento en una vecindad $V \subset M^{(3)}$. En la diagonal se presentan los histogramas correspondientes.

- Auto-similaridad espacial: Debido a que el WRF es un modelo de downscaling dinámico, los datos que se obtienen a partir de él están anidados, lo cual preserva la estructura de la malla más gruesa en las mallas más finas. En nuestro caso, lo anterior se traduce a que la malla $M^{(3)}$ está contenida en la malla $M^{(1)}$ en el sentido de posiciones geográficas, es decir, todo punto en $M^{(3)}$ es también un punto en $M^{(1)}$, como se aprecia en la Figura 3.13 (este fenómeno ocurre en cualquier par mallas cuya relación en sus escalas sea un factor de tres, como por ejemplo $M^{(1,5)}$ con $M^{(0,5)}$). Sin embargo, si consideramos un horizonte de tiempo común $\mathcal{T}$, las series de tiempo en los puntos comunes de las mallas $M^{(1)}$ y $M^{(3)}$ no son necesariamente iguales debido al efecto de escala. A pesar de lo anterior, suponemos que son realizaciones del mismo proceso estocástico, es decir, sus funciones de distribución son iguales.



(a)$M^{(3)}$ (b)$M^{(1)}$ (c)$M^{(3)}$ y $M^{(1)}$

FIGURA 3.13: Relación entre los elementos de $M^{(3)}$ y $M^{(1)}$.

Las hipótesis de regularidad y homogeneidad local garantizan la suavidad necesaria para estimar la función de interpolación de manera local mediante el método de regresión lineal. Para ello definimos vecindades de puntos y sobre cada una de ellas estimamos el interpolador $I(\cdot)$.

Las vecindades deben definirse de manera única, de manera que no se asolapan, permitiéndoles coincidir solo en sus bordes. Al definir las vecindades de esta forma se garantiza la identificabilidad del modelo, en el sentido de que bajo el esquema que queremos describir, la estimación local de la función de interpolación debe ser unívocamente definida. Cabe añadir que localmente, en las zonas de homogeneidad espacial, las series de tiempo del residuo R se asumen estadísticamente similares, esto quiere decir que la distancia telescópica entre las distribuciones correspondientes es pequeña.

En particular presentan similaridad en sus componentes medias, variabilidad y comportamiento de dependencia temporal o autocovarianza.

Una observación importante sobre nuestro modelo es que, dada la hipótesis de autosimilaridad y la información proveniente de los campos de viento sobre las mallas $M^{(3)}$ y $M^{(1)}$, lo natural es realizar un downscaling desde $M^{(1)}$ para obtener información sobre la malla $M^{(1/3)}$, pues es ella la que preserva la estructura de $M^{(1)}$. El esquema que permitiría realizar dicha estimación se basa en establecer una relación empírica entre las mallas $M^{(3)}$ y $M^{(1)}$, la cual, mediante la hipótesis de autosimilaridad, debe ser la misma que la existente entre $M^{(1)}$ y $M^{(1/3)}$, salvo un factor de escala. Este es un problema mal planteado pues al momento de establecer la relación entre $M^{(3)}$ y $M^{(1)}$ la variable explicativa es $M^{(3)}$ y la respuesta es $M^{(1)}$, lo que produce un problema subdeterminado. Acompañado a lo anterior, se presenta un problema de identificabilidad, puesto que es necesario considerar vecindades de, al menos, $2^3 = 8$ puntos y estás no se pueden definir de manera única, teniendo problemas de redundancia de información y por ende colinealidad en el modelo. Este problema es denominado como *maldición de la dimensión* o curse of dimensionality.

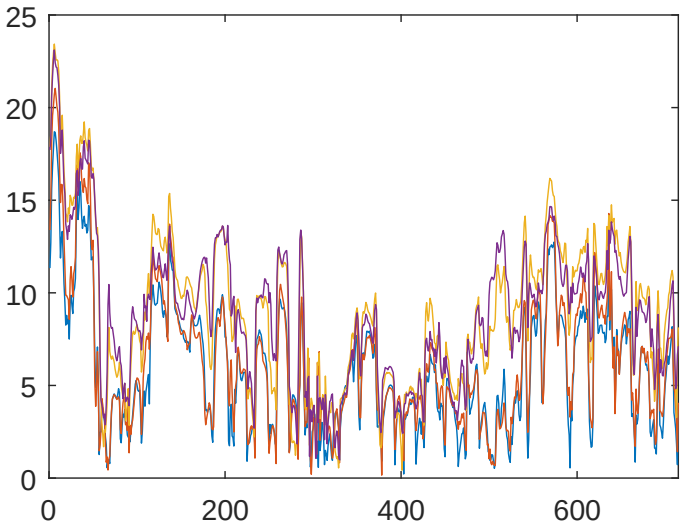


FIGURA 3.14: Series de tiempo para los elementos de la vecindad $V \subset M^{(3)}$.

Para lidiar con los problemas anteriores proponemos considerar vecindades de $2^2 = 4$ puntos, como se describió anteriormente (ver Figura 3.11), las cuales quedan definidas de manera única. La Figura 3.14 muestra las series temporales para una vecindad

$V \subset M^{(3)}$ a modo de ejemplificar el cumplimiento de las hipótesis de regularidad.

Para aprovechar estas vecindades realizamos un paso intermedio de upscaling determinista o incremento de escala, en donde a partir de una malla generamos otra menos fina. Comenzamos con un paso de subida de escala o upscaling, generando mediante el campo de intensidad de viento sobre $M^{(1)}$ el campo de intensidad de viento sobre la malla $M^{(1,5)}$. Luego establecemos la relación empírica que hay entre $M^{(3)}$ y $M^{(1,5)}$, la cual queda unívocamente definida en cada vecindad de cuatro puntos. Además dada la hipótesis de autosimilaridad dicha relación debe ser la misma que existe entre $M^{(1)}$ y $M^{(0,5)}$, salvo un factor de escala, siendo esta la motivación para estimar $X^{(0,5)}$ desde $X^{(1)}$ y $X^{(3)}$, usando la estructura geométrica de las mallas. La Figura 3.15 muestra los pasos a realizar.

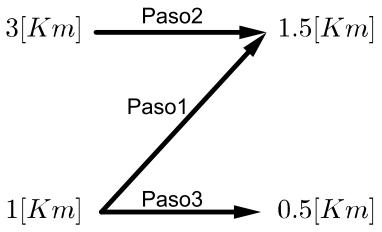


FIGURA 3.15: Esquema general del modelo de downscaling.

En cada vecindad $V = \{s_1, s_2, s_3, s_4\}$ de la malla $M^{(3)}$, representados en la Figura 3.16 por los puntos rojos, podemos describir nuestro método esquemáticamente como:

- **Paso 1 (upscaling):** Identificamos los únicos cuatro puntos $V' = \{s'_1, s'_2, s'_3, s'_4\}$ de $M^{(1,5)}$ que se encuentran al interior de la vecindad V. Los elementos de V' están representados en la Figura 3.16 por los triángulos verdes.

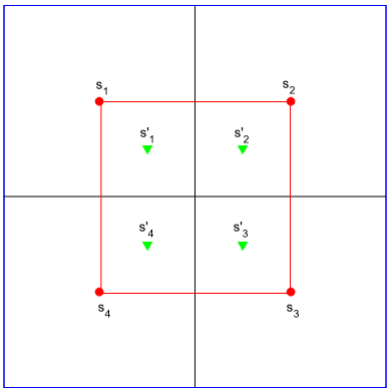


FIGURA 3.16: Representamos la vecindad $V = \{s_1, s_2, s_3, s_4\} \in M^{(3)}$ con puntos rojos. Al interior de V se encuentran los puntos de la vecindad $V' = \{s'_1, s'_2, s'_3, s'_4\} \in M^{(1,5)}$ representado por triángulos verdes.

Sobre V' estimamos las series temporales $x^{(1,5)}_{(s'_i,\cdot)}$ mediante una interpolación determinista que toma en cuenta los puntos en $M^{(1)}$ al interior de la vecindad V, es decir establecemos una relación determinista entre $\left(x^{(1,5)}_{(s'_1,\cdot)}, x^{(1,5)}_{(s'_2,\cdot)}, x^{(1,5)}_{(s'_3,\cdot)}, x^{(1,5)}_{(s'_4,\cdot)}\right)$ y los 16 elementos de $X^{(1)}$ contenidos en V, los cuales son representados por cuadrados azules en la Figura 3.17. Note que en esta figura los cuatro puntos de la vecindad son comunes de las mallas $M^{(3)}$ y $M^{(1)}$.

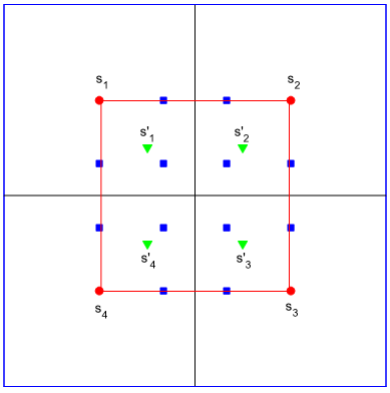


FIGURA 3.17: Se representan mediante cuadrados azules los puntos de la malla $M^{(1)}$ contenidos en la vecindad V. A su vez, con triángulos verdes, los puntos de la vecindad V'.

- **Paso 2 (ajuste de regresión lineal):** Una vez obtenidas las series temporales $x^{(1,5)}_{(s'_i,\cdot)}$ para $s'_i \in V'$, procedemos a establecer la relación empírica entre los elementos de V y V' (ver Figura 3.18). Esta relación se estima mediante técnicas de regresión lineal, generando la función de interpolación $\hat{I}(\cdot)$ definida para puntos al interior de la vecindad V.

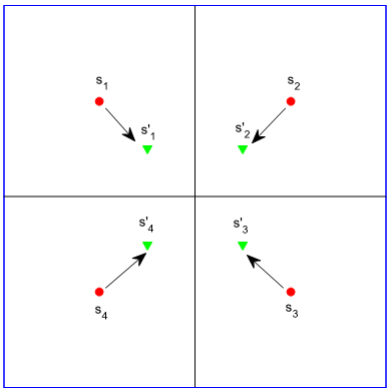


FIGURA 3.18: Relación de dependencia entre los elementos de V y V'.

- **Paso 3 (downscaling):** Finalmente aplicamos la función $\hat{I}(\cdot)$ a cada una de las nueve vecindades de cuatro puntos $W \subset M^{(1)}$ al interior de la vecindad V, las cuales están representadas en la Figura 3.19.

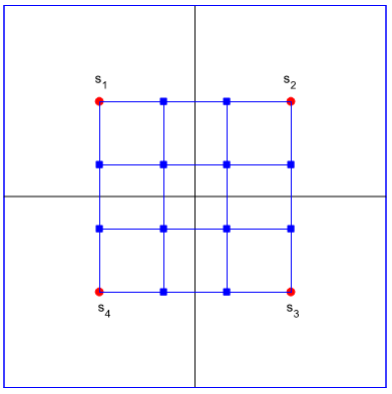


FIGURA 3.19: Vecindades $W \subset M^{(1)}$ al interior de la vecindad $V \subset M^{(3)}$. Los elementos pertenecientes a $M^{(1)}$ están representados por cuadrados azules.

de esta manera se genera la estimación $\hat{X}^{(0,5)}$ sobre cada una de las vecindades $W \subset M^{(0,5)}$ al interior de $V \subset M^{(3)}$ sobre la malla $M^{(0,5)}$. En la Figura 3.20 los diamantes negros representan a estos puntos de la malla $M^{(0,5)}$.

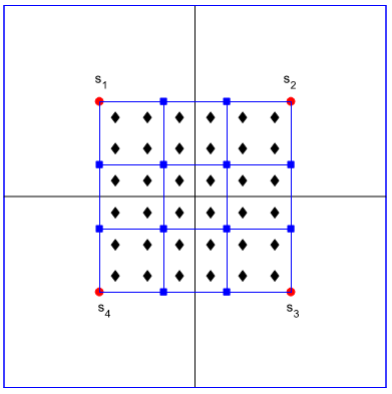


FIGURA 3.20: Elementos de la malla $M^{(0,5)}$ sobre los cuales se realiza el downscaling. Estos puntos están representados por los diamantes negros.

En cada uno de los pasos anteriores, consideramos las siguientes técnicas:

- **Paso 1:** La estimación sobre V' se realiza mediante una interpolación bicúbica, la cual es una generalización al caso bidimensional de las interpolaciones cubicas. Esta técnica consiste en dos interpolaciones cúbicas consecutivas, una horizontal y otra vertical, usando un total de 16 puntos en la estimación, como se muestra en la Figura 3.21.

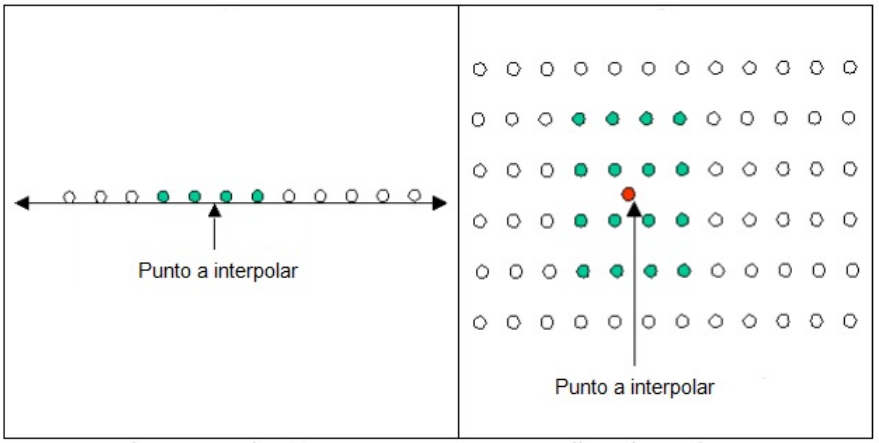


FIGURA 3.21: Puntos usados en la interpolación cúbica (izquierda) y en la interpolación bicúbica (derecha).

La función de usada para esta interpolación en una dimensión está dada por:

$$u(d) = \begin{cases} \frac{3}{2}3|d| - \frac{5}{2}2|d| + 1, \text{si} \quad 0 \leq |d| < 1 \\ -\frac{1}{2}3|d| + \frac{5}{2}2|d| - 4|d| + 2, \text{si} \quad 1 \leq |d| < 2 \\ 0, \text{si} \quad 2 < |d| \end{cases} \tag{3.7}$$

donde d es la distancia entre el punto a estimar y el punto en la malla considerado en la interpolación. De esta forma, en el caso bidimensional la función de interpolación está dada por:

$$g(x,y) = \sum_{n=-1}^{2} \sum_{m=-1}^{2} c_{j+n,k+m} u(\text{dist}_{x_{j+n}}) u(\text{dist}_{y_{k+m}}), \tag{3.8}$$

con (x,y) el punto que a estimar, (x_j, y_k) los puntos usados en la interpolación, dist_x y dist_y las distancias horizontales y verticales de los puntos de la interpolación al punto a estimar y las constantes $c_{j,k}$ son el valor de la malla en el punto (x_j, y_k).

- **Paso 2:** Asumimos que los elementos de V y los estimados sobre V' poseen la siguiente relación lineal:

$$\begin{aligned} \left[\hat{x}^{(1,5)}_{(s'_1,\cdot)}, \hat{x}^{(1,5)}_{(s'_2,\cdot)}, \hat{x}^{(1,5)}_{(s'_3,\cdot)}, \hat{x}^{(1,5)}_{(s'_4,\cdot)}\right] &= \left[x^{(3)}_{(s_1,\cdot)}, x^{(3)}_{(s_2,\cdot)}, x^{(3)}_{(s_3,\cdot)}, x^{(3)}_{(s_4,\cdot)}\right] \beta, \\ \mathbf{Y}_{V'} &= \mathbf{X}_V \beta. \end{aligned}$$

de esta forma, estimar la función $I(\cdot)$ es equivalente a estimar β. Usando el método de mínimos cuadrados obtenemos el estimador $\hat{\beta}$ mediante la ecuación:

$$\hat{\beta} = (\mathbf{X}_V^\top \mathbf{X}_V)^{-1} \mathbf{X}_V^\top \mathbf{Y}_{V'}.$$

Notemos que en nuestro caso β y $\hat{\beta}$ son matrices de dimensión 4×4, pues estamos realizando una regresión lineal multivariada.

- **Paso 3:** Finalmente aplicamos $\hat{\beta}$ a cada una de las vecindades en $M^{(1)}$ contenidas en la vecindad V. Así, obtenemos la malla (locamente) a 0.5 mediante la siguiente expresión:

$$\mathbf{Y}_{V'} = \mathbf{X}_V \hat{\beta}. \tag{3.9}$$

Capítulo 4

Resultados

4.1. Caso de Estudio y metodología de trabajo

En este capítulo realizamos la descripción de los datos usados y el análisis de los resultados obtenidos. Comenzamos señalando antecedentes e hitos geográficos de la zona de estudio, entendiendo la necesidad de desarrollar técnicas que ayuden a comprender el comportamiento de fenómenos meteorológicos.

Chile es un país que presenta una gran variabilidad en cuanto a sus climas y geografía, pues su parte sudamericana se extiende a lo largo de 39° de latitud, mayormente desde la ribera sudoriental del océano Pacífico hasta la cordillera de los Andes, entre los paralelos 17°29′57" S y 56°32′ S, a lo largo de 4270 [Km], presentándose hasta nueve tipos de climas diferentes, sin contar subclases. Su territorio abarca zonas altamente sísmicas y volcánicas, pertenecientes al Cinturón de fuego del Pacífico, facilitando la subducción o solapamiento de las placas de Nazca y Antártica en la placa Sudamericana (ver Figura 4.1), además de la ocurrencia de eventos atípicos como erupciones volcánicas, terremotos, tormentas, entre otros, que afectan el comportamiento del clima a corto y largo plazo.

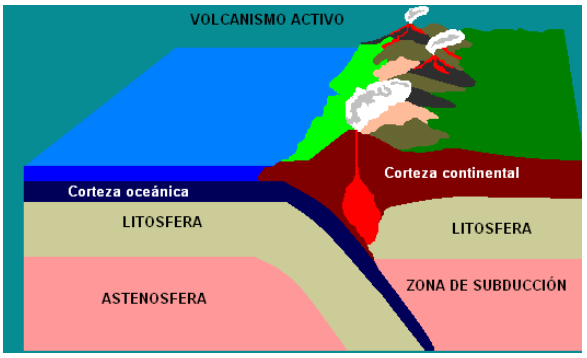


FIGURA 4.1: Fenómeno de subducción entre dos placas, en donde una de las placas (corteza oceánica) se coloca por debajo de la otra (corteza terrestre)[1].

Dentro de todas las regiones presentes en Chile, la región de Valparaíso presenta una de las mayores heterogeneidades geográficas, pues cuenta con la presencia de las dos cordilleras más importantes del país, la cordillera de los Andes y de la Costa, además de la presencia de diferentes hitos geográficos, tales como los denominados valles transversales y planicies litorales. Esta región es considerada como una zona de transición, pues los valles transversales son característicos del norte del país y además se presentan vestigios de valles longitudinales, característicos de zonas más al sur de Chile. A lo anterior se suma la presencia de cuatro tipos de clima en la región, por lo que es de esperarse que diferentes localidades posean comportamientos y características diferentes para algunos fenómenos climatológicos, tales como la cantidad de precipitación, humedad, intensidad y dirección de masas de aire o viento, etc.

Estas cualidades geográfica dificultan estudiar el comportamiento de la intensidad del viento, pues este cambia bruscamente debido a la presencia de las cordilleras y valles, las que provocan irregularidades de altura, cambios abruptos de pendientes y diferentes exposiciones a las corrientes de viento; por lo que en la práctica mediciones de esta característica que están geográficamente cercanas pueden ser muy diferentes entre sí, por ejemplo en la comuna de Valparaíso la costa está a unos metros de sus cerros, encontrándonos con comportamientos completamente diferentes a pocos metros de distancia. Sin embargo, el estudio del comportamiento del viento en la región, tanto su intensidad como su dirección, ha tomado relevancia en los últimos años, pues existe gran interés en la implementación de sistemas de energía renovable basado en granjas eólicas [16] como en la prevención y reacción temprana ante catástrofes relacionadas a eventos climáticos, como el incendio que afectó a la ciudad de Valparaíso el 12 de abril

[1] https://previa.uclm.es/profesorado/egcardenas/subduccion.htm

de 2014, marejadas en las costas que afectan la infraestructura, como la del invierno del 2015 y 2017, fallas del sistema eléctrico ocasionadas por intensidades de viento extremas, entre otros.



FIGURA 4.2: Mapa morfológico de la región de Valparaíso, en donde se aprecia la diversidad geográfica presente en la región[2].

Lo anterior motiva el desarrollo de técnicas de downscaling que permitan estimar cómo se comporta este fenómeno a la mayor resolución espacial posible de manera rápida y efectiva. Nuestro objetivo principal es implementar el modelo desarrollado en el capítulo 3.3, que considera mallas de viento con resoluciones espaciales de 3[Km] y 1[Km], obtenidas desde el modelo dinámico Weather and Research Forecasting o WRF (ver Capítulo 3.1.1) ubicadas geográficamente en la región de Valparaíso, abarcando la zona entre 32°0′19,8″ y 34°3′9,72″ latitud sur; 70°21′46,08′ y 71°57′1,23″ longitud oeste, la cual corresponde a la zona dentro del rectángulo rojo en la Figura 4.3. Con estas mallas buscamos obtener una estimación del viento sobre una malla de resolución 0.5[Km].

[2]http://www.educarchile.cl/ech/pro/app/detalle?id=132432

Estas simulaciones son realizadas a diez metros por sobre el nivel del mar y son tomadas cada una hora, en concreto. En nuestro estudio, se consideran simulaciones provenientes del WRF correspondientes al mes de Septiembre del 2014. Las simulaciones son realizadas cada una hora, a partir de las 01:00 horas del primero de Septiembre de 2014, finalizando las 20:00 horas del día treinta de Septiembre de 2014.

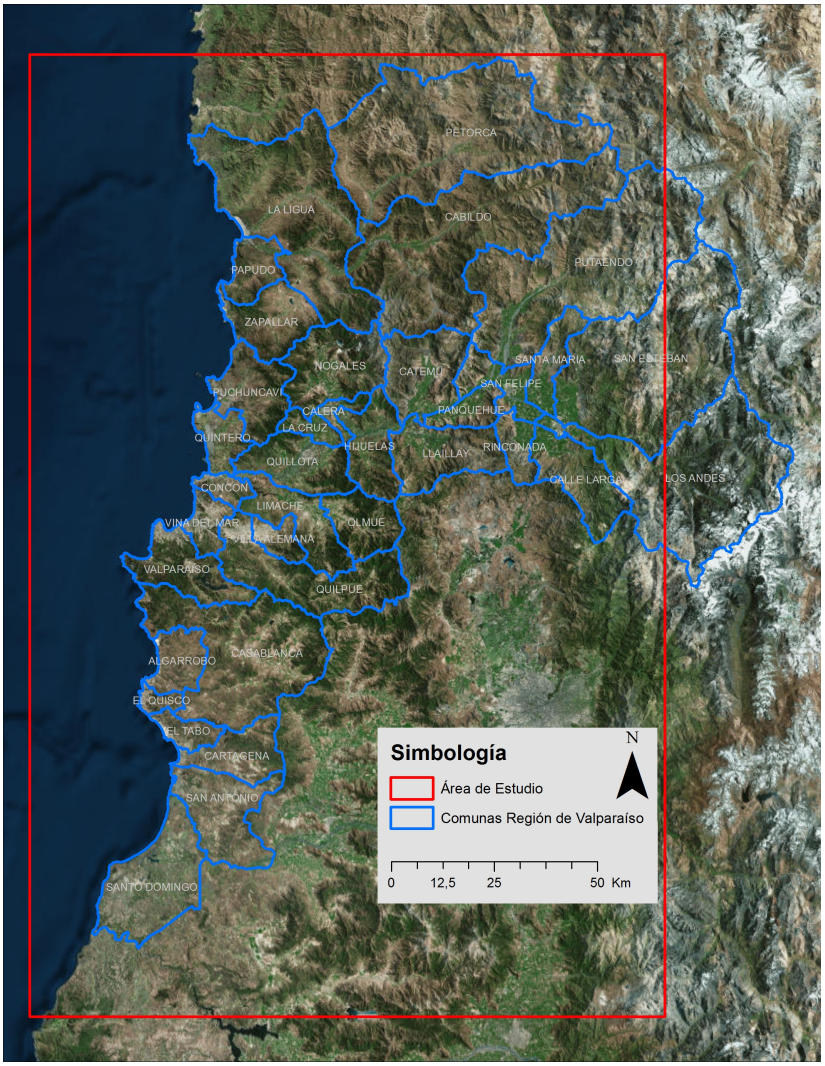


FIGURA 4.3: Zona de estudio para la intensidad de viento, correspondiente a la región de Valparaíso y parte de la región metropolitana, con coordenadas 32°0′19,8″ y 34°3′9,72″ latitud sur; 70°21′46,08′ y 71°57′1,23″ longitud oeste.

Comenzamos haciendo un pequeño análisis exploratorio de los datos, luego mostramos el resultado de la interpolación para obtener campos de viento con resolución de 1.5[Km]. Siguiendo el esquema geométrico de escalamiento presentado en la Figura 3.15, continuamos con el ajuste del modelo de regresión lineal entre los campos a 3 [Km] y 1.5 [Km], para finalmente, a partir de este ajuste, obtener una aproximación

del campo de viento con resolución de 0.5 [Km].

El análisis espacial de los resultados es realizado los días 1, 21 y 29 del mes de Septiembre de 2014, mostrando en cada uno de ellos dos horas diferentes, correspondientes a las 10:00 y 22:00 horas para el 01/09/2014 y las 11:00 y 23:00 horas para el 21/09/2014 y 29/09/2014. El análisis temporal se realiza en tres vecindades, ubicadas en zonas de características geográficas diferentes, las cuales son el mar, valle y montaña. Esto ayuda a tener una mejor comprensión de la heterogeneidad espacial de los datos.

4.2. Análisis Exploratorio

Para ilustrar el comportamiento aleatorio de las intensidades diarias de viento, comenzamos realizando los gráficos de caja (boxplots) de los datos a 3[Km] y 1[Km], donde cada caja se construye agrupando todas las mediciones en la zona de estudio realizadas en un día, es decir, tomamos veinticuatro campos de viento $X^{(l)}_{(\cdot,t_0:t_0+24)}$ para cada resolución $l = \{1, 3\}$, con t_0 fijo y agrupamos sus valores en un vector. Con esto, generamos un gráfico de caja para los primeros 29 días del mes de Septiembre de 2014, lo que permite visualizar la distribución diaria de los datos y su evolución en el tiempo.

Notamos que la distribución de la intensidad diaria del viento está fuertemente sesgada a derecha, por lo que no sigue una distribución normal. Además, se observa gran cantidad de datos atípicos con respecto a la mediana y al tercer cuartil, señalando que la distribución posee colas pesadas. Podemos notar que la distribución diaria del viento presenta una componente estacional y presenta una variabilidad no constante en el tiempo, por lo cual no es estacionaria. Este hecho no influye en nuestro método, pues necesitamos que la distribución del residuo $R_{(0,5:1)}$ sea estacionaria; asumimos que $I\left(X^{(1)}\right) = E\left[X^{(0,5)}|X^{(1)}\right]$ capta el comportamiento no estacionario en media y $R_{(0,5:1)}$, por la homogeneidad local de los datos, es localmente estacionario. Esta hipótesis es necesaria para realizar el paso de clasificación del método (ver Capítulo 2.3).

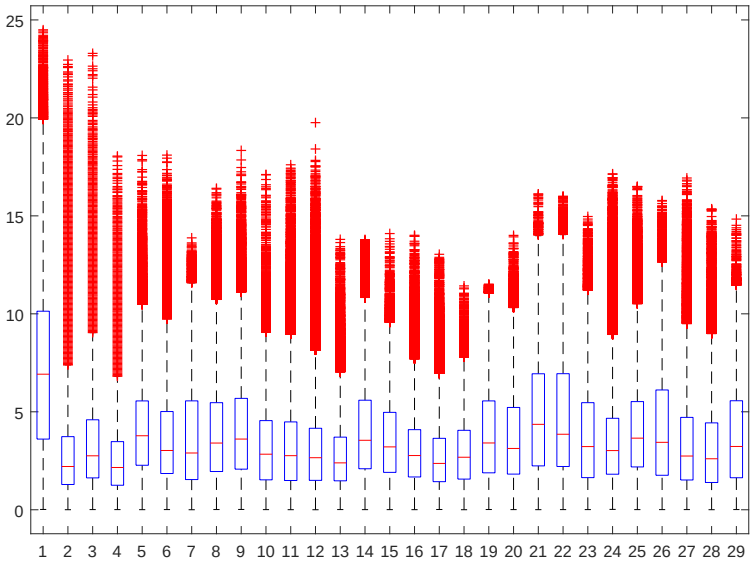


FIGURA 4.4: Boxplot de la intensidad del viento diaria para los datos a 3[Km] correspondientes a Septiempre de 2014

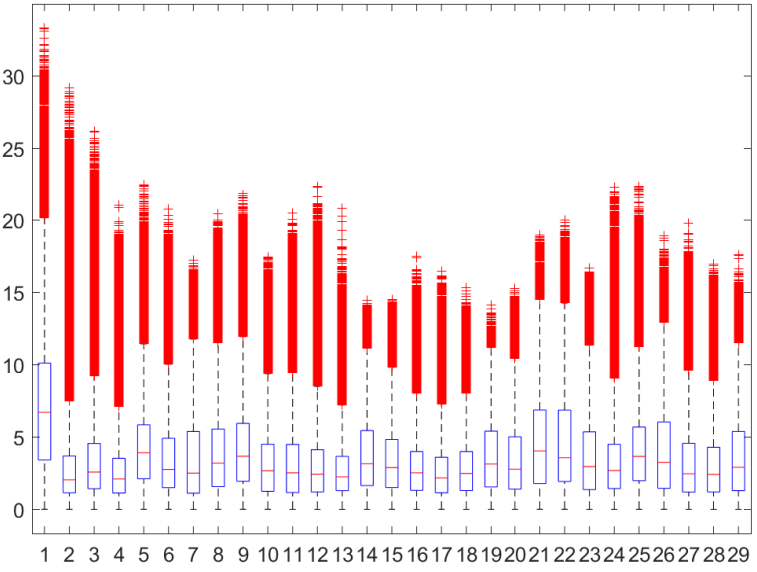


FIGURA 4.5: Boxplot de la intensidad del viento diaria para los datos a 1[Km] correspondientes a Septiempre de 2014

4.3. Análisis de resultados.

En esta sección procederemos a describir los resultados obtenidos mediante el modelo de downscaling desarrollado en el Capítulo 3.3. Para ello seguiremos el esquema de trabajo planteado por la Figura 3.15, en donde obtenemos el campo a 1.5[Km] mediante el interpolado bicúbico (ver ecuaciones (3.7) y (3.8)), luego ajustamos un modelo de regresión lineal entre los campos a 3[Km] y 1.5[Km], para cada vecindad de cuatro puntos de la malla a 3[Km] definidas como en la Figura 3.11. Finalmente obtenemos en cada una de estas vecindades el campo a 0.5[Km] a través del campo a 1[Km] y el ajuste de regresión.

4.3.1. Campo de viento a 1.5[Km]: Interpolación Bicúbica

El modelo general descrito en el Capítulo 3.3, establece en el paso de upscaling que el campo de viento a 1.5[Km] es una interpolación determinista del campo de viento a 1[Km]. Recordemos que nuestra elección de interpolación es el método bicúbico, descrito en el Capítulo 3.3.1. Es por esto que comenzamos discutiendo brevemente las ventajas de este interpolador frente a otros igualmente conocidos, tales como el interpolador bilineal y el del vecino más cercano.

El interpolador bicúbico consiste en realizar dos interpolaciones cúbicas consecutivas, una vertical y otra horizontal. El interpolador bilineal funciona similar al bicúbico, pues realiza dos interpolaciones lineales, suponiendo que la relación entre los puntos es lineal. El del vecino más cercano consiste en asociar la misma información entre elementos proximales. Una diferencia que poseen estos dos métodos con respecto al bicúbico es que usan una menor cantidad de puntos en la interpolación. En una y dos dimensiones, el bicúbico utiliza 4 y 16 puntos respectivamente, mientras que los otros mencionados usan 2 y 4, como se representa en la Figura 4.6.

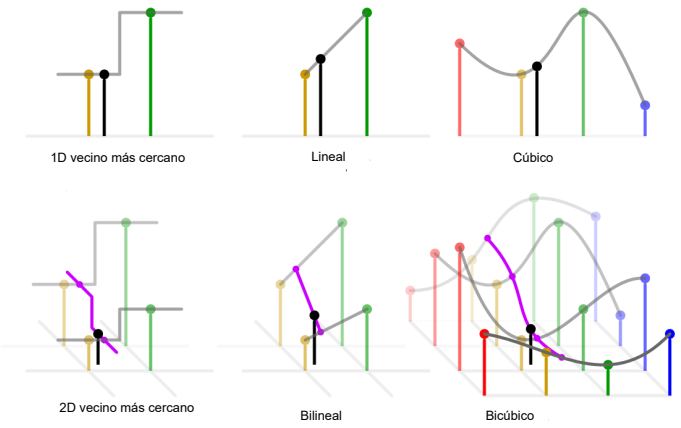


FIGURA 4.6: Comparación entre los interpoladores lineal, cúbico y vecino más cercano El punto negro corresponde al punto que se desea interpolar y los rojos, amarillos, verdes y azules corresponden a los puntos usados por los interpoladores. La altura representa los valores correspondientes a la interpolación[3].

Nuestra elección fue tomada considerando el efecto de la interpolación en las fronteras entre zonas de alto contraste en las imágenes de los campos de intensidad del viento. En nuestro modelo, al usar datos provenientes del modelo WRF que presentan fronteras suaves entre las zonas de contraste, es crucial en el paso de upscaling usar un interpolador que respete las fronteras y mantenga la suavidad en la transición entre estas zonas. Por un lado, el interpolador del vecino más cercado, dada su naturaleza, puede presentar un efecto de mezcla de pixeles, produciendo que la frontera de zonas de alto contraste resulte irregular. Por otro parte, el interpolador bilineal respeta las fronteras, la transición entre zonas de alto contraste es más suave en el bicúbico, ya que, al usar una mayor información del entorno, capta de mejor manera la transición entre diferentes colores y aproxima de manera suave los contornos. En la Figura 4.7 se muestra un ejemplo aplicado a un campo de viento al hacer un upscaling determinista de 1[Km] a 1.5[Km]. Se observa que, en comparación con el interpolador bicúbico, en una zona pequeña de la región de Valparaíso ubicada entre 32°21′25,2″ y 33°4′37,2″ latitud sur; 70°52′24,24″ y 71°18′10,8″ longitud oeste, el interpolador del vecino más cercano produce un efecto de mezcla de pixeles entre las zonas amarillas y celestes, mientras que las diferencias entre el bilinial y el bicúbico no son apreciables a esta resolución.

[3]https://en.wikipedia.org/wiki/Bicubic_interpolation

Además, en la Figura 4.8 se muestra un ejemplo aplicado a un campo de viento al realizar un downscaling determinista de 1[Km] a 0.5[Km], en el cual se aprecia que el interpolador bicúbico presenta fronteras más suaves.

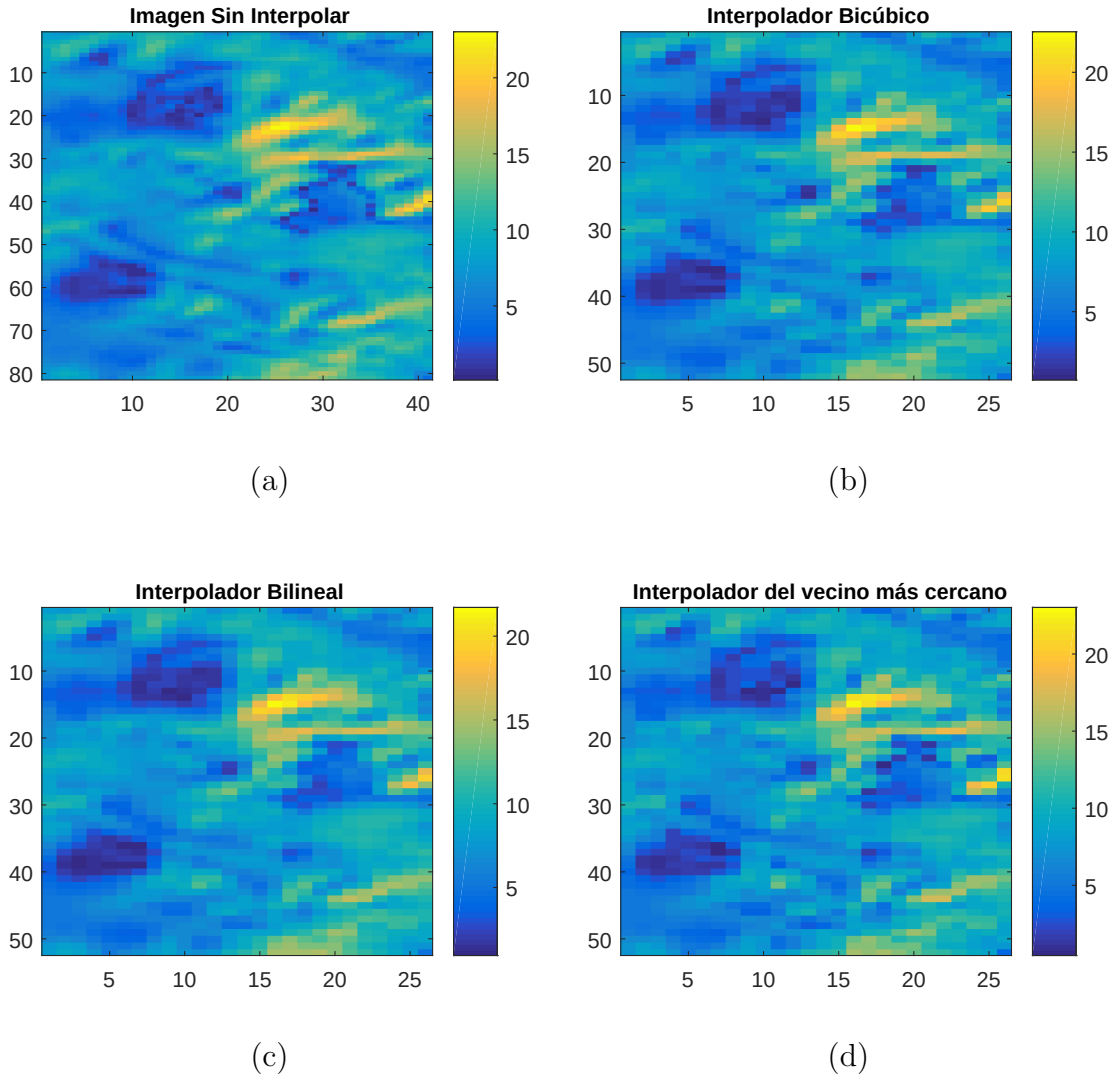


FIGURA 4.7: Comparativa entre diferentes métodos de interpolación para un upscaling de 1[Km] a 1.5[Km] para el campo de intensidad del viento: (a) corresponde a la imagen original, (b) corresponde al interpolador bicúbico, (c) corresponde al interpolador bilineal y (d) al interpolador del vecino más cercano.

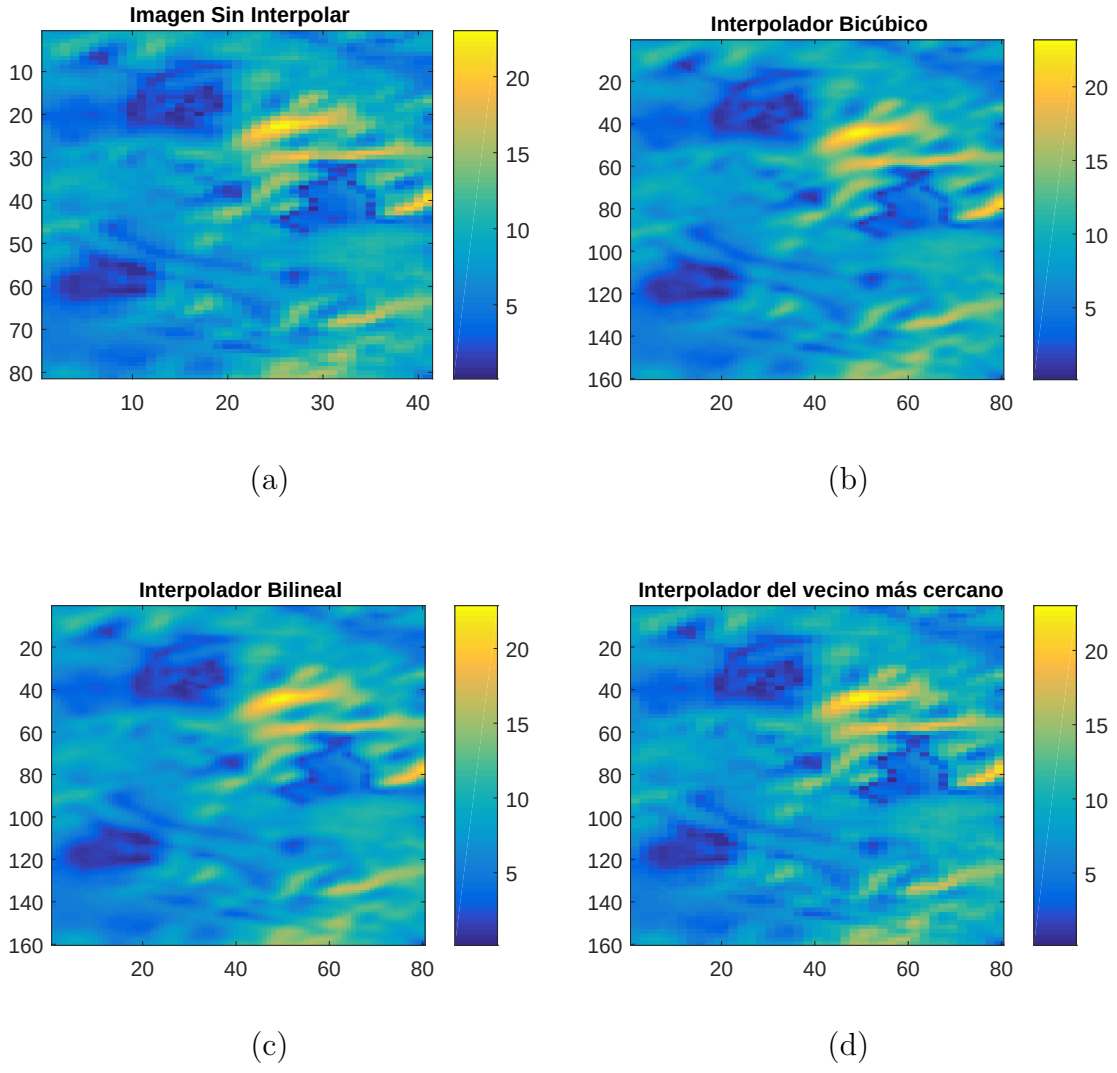


FIGURA 4.8: Comparativa entre diferentes métodos de interpolación para un downscaling de 1[Km] a 0.5[Km] para el campo de intensidad del viento: (a) corresponde a la imagen original, (b) corresponde al interpolador bicúbico, (c) corresponde al interpolador bilineal y (d) al interpolador del vecino más cercano.

4.3.1.1. Análisis Espacial

A continuación presentamos los resultados que se obtienen al aplicar el interpolador bicúbico al campo de viento a 1 [Km] para obtener el campo de viento a 1.5 [Km] en la zona de estudio descrita anteriormente (ver Figura 4.3). Se presentan los días 1,21 y 29 del mes de Septiembre del año 2014. En cada día se presentan dos horas particulares, correspondientes a las 10:00 y 22:00 horas para el día 1/09/2014, 11:00 y

23:00 horas para los días 21/09/2014 y 29/09/2014, debido al cambio de horario que tuvo efecto el día 06/09/2014. Además, se muestra una ampliación de la imagen en una zona de alto contraste, evidenciando la propiedad del interpolador en cuanto a los contrastes.

En las Figuras 4.9, 4.10 y 4.11 se observa que las imágenes interpoladas a 1.5[Km] presentan contornos más irregulares que las originales a 1[Km] debido al upscaling, sin embargo, las fronteras entre zonas de contraste se mantienen, debido al uso del interpolador bicúbico.

4.3.1.2. Análisis temporal

En la parte temporal presentamos los elementos pertenecientes a tres vecindades $V \subset M^{(3)}$ ubicadas en diferentes zonas de la región: en el interior del mar, en un valle y en la montaña. En cada vecindad mostramos los elementos pertenecientes a $V' \subset M^{(1,5)}$ y los de $W \subset M^{(1)}$, cada una definida como en la Figura 3.11. Nuestro objetivo es analizar el comportamiento en el tiempo de las series temporales generadas mediante la interpolación bicúbica.

En la vecindad ubicada en el mar observamos que el comportamiento de las series temporales asociadas a la vecindad $W \subset M^{(1)}$ son prácticamente idénticas, por lo que es natural esperar que al realizar cualquier tipo de interpolación para obtener las series temporales asociadas a la vecindad $V' \subset M^{(1,5)}$, se obtienen resultados similares a las series de 1[Km]. En la Figura 4.12 apreciamos este hecho, en donde las series de tiempo a 1.5[Km] presentan el mismo comportamiento y son muy similares entre sí, diferenciándose levemente en los valores debido al suavizamiento del interpolador.

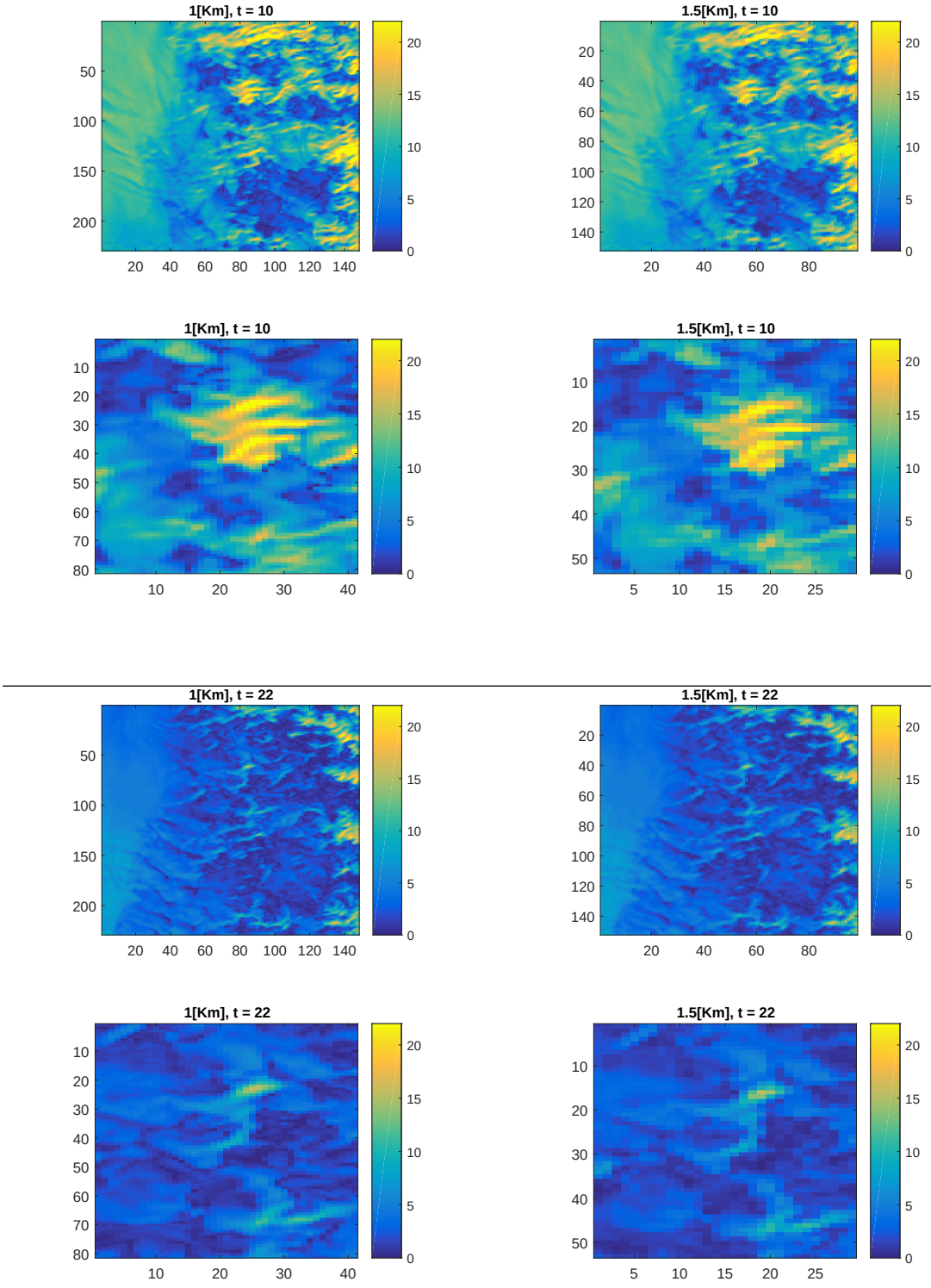


FIGURA 4.9: Campos de intensidad viento $X^{(1)}_{(\cdot,t_1)}$, $X^{(1,5)}_{(\cdot,t_1)}$ y $X^{(1)}_{(\cdot,t_2)}$, $X^{(1,5)}_{(\cdot,t_2)}$, donde $t_1 = 10$ y $t_2 = 22$ correspondientes al día 01 de Septiembre de 2014 a las 10:00 y 22:00 horas respectivamente. Además se muestran amplificaciones de los mismos campos entre las coordenadas $32°21'25{,}2''$ y $33°4'37{,}2''$ latitud sur; $70°52'24{,}24''$ y $71°18'10{,}8''$ longitud oeste.

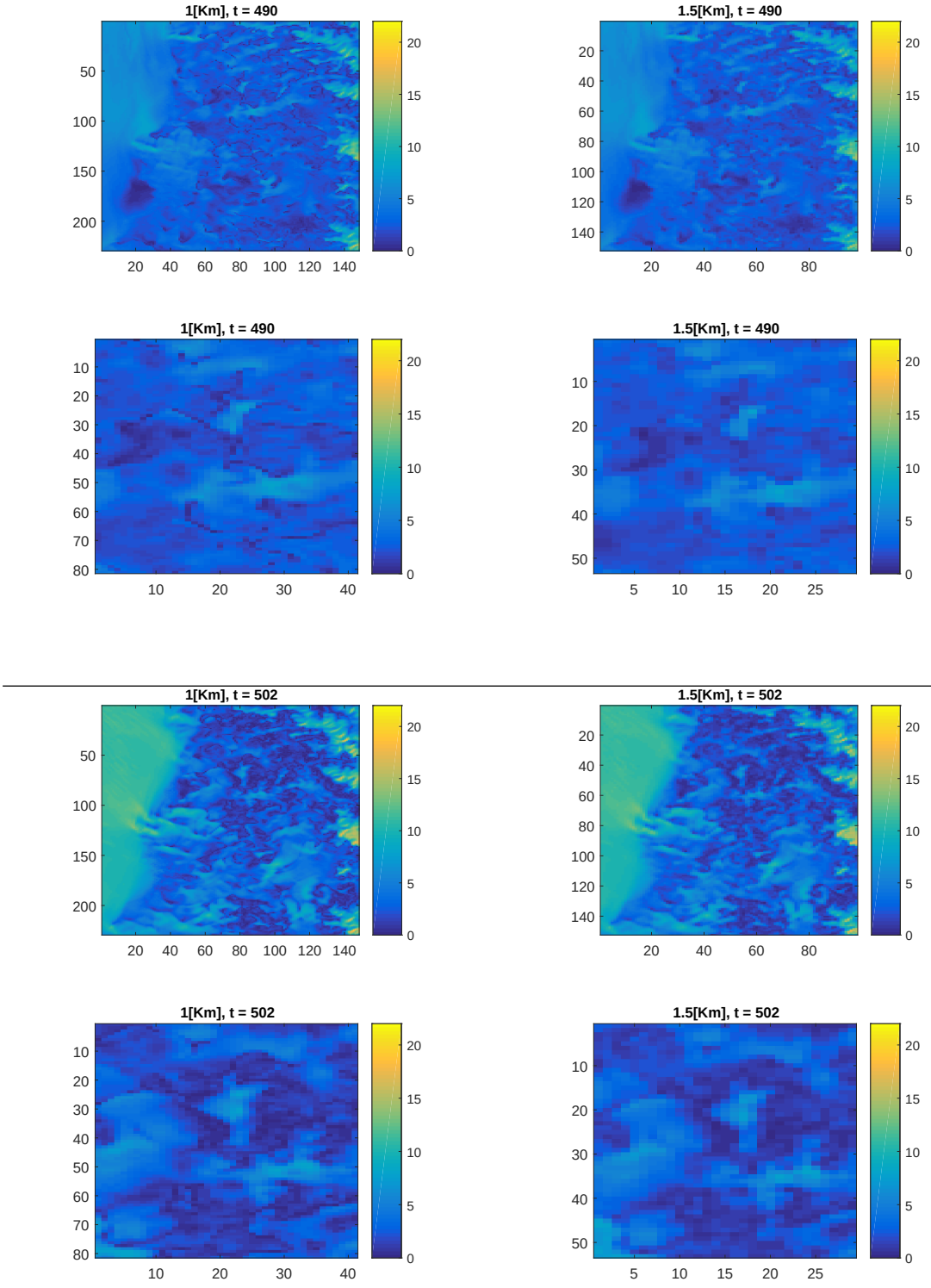


FIGURA 4.10: Campos de intensidad viento $X^{(1)}_{(\cdot,t_1)}$, $X^{(1,5)}_{(\cdot,t_1)}$ y $X^{(1)}_{(\cdot,t_2)}$, $X^{(1,5)}_{(\cdot,t_2)}$, donde $t_1 = 490$ y $t_2 = 502$ correspondientes al día 21 de Septiembre de 2014 a las 11:00 y 23:00 horas respectivamente. Además se muestran amplificaciones de los mismos campos entre las coordenadas 32°21′25,2″ y 33°4′37,2″ latitud sur; 70°52′24,24″ y 71°18′10,8″ longitud oeste.

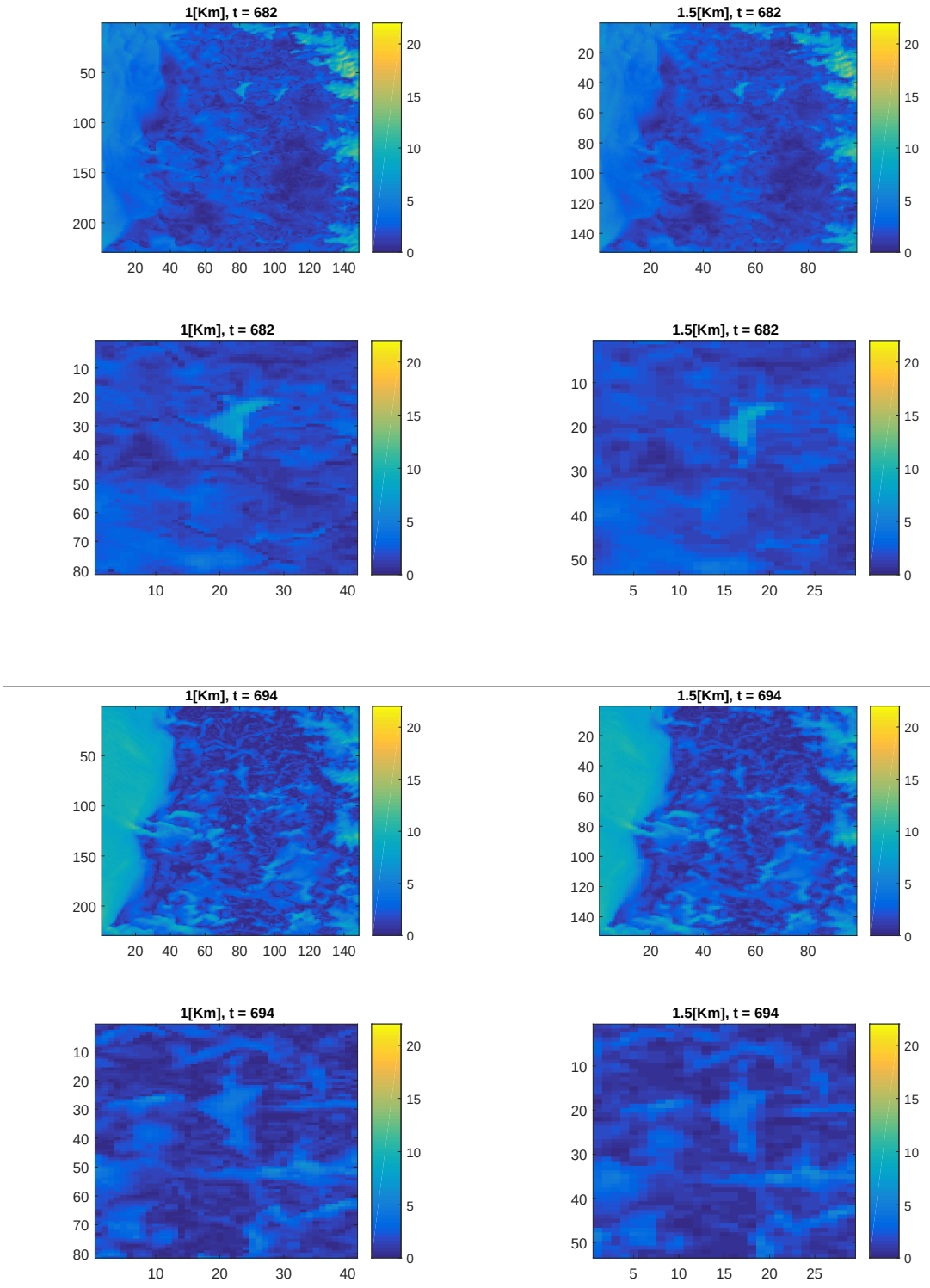


FIGURA 4.11: Campos de intensidad viento $X^{(1)}_{(\cdot,t_1)}$, $X^{(1,5)}_{(\cdot,t_1)}$ y $X^{(1)}_{(\cdot,t_2)}$, $X^{(1,5)}_{(\cdot,t_2)}$, donde $t_1 = 682$ y $t_2 = 694$ correspondientes al día 29 de Septiembre de 2014 a las 11:00 y 23:00 horas respectivamente. Además se muestran amplificaciones de los mismos campos entre las coordenadas 32°21′25,2″ y 33°4′37,2″ latitud sur; 70°52′24,24″ y 71°18′10,8″ longitud oeste.

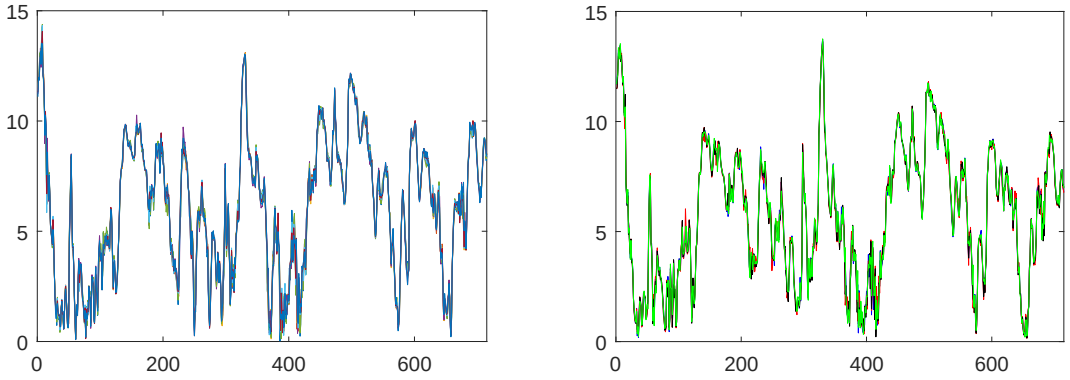


FIGURA 4.12: Series temporales para los elementos pertenecientes a $W \subset M^{(1)}$ y $V' \subset M^{(1,5)}$ al interior de la vecindad V ubicada en el mar y con coordenadas geográficas 31°37′1,56″ y 31°38′40,92″ latitud sur; 72°57′10,44″ y 72°59′7,08″ longitud oeste. A la izquierda se muestran las series correspondientes a W y a la derecha las correspondientes a V'.

La vecindad ubicada en el valle poseen un comportamiento localmente homogéneo, similar a las de zonas marítimas, siendo las series de tiempo asociadas a la vecindad $W \subset M^{(1)}$ parecidas entre sí, presentando leves diferencias en los valores extremos. Las series temporales asociadas a la vecindad $V' \subset M^{(1,5)}$ también son similares entre si y a las de W, salvo una ligera disminución en los valores extremos, tal como se aprecia en la Figura 4.13.

La vecindad presente en la zona montañosa presenta mayor heterogeneidad en cuanto a las series temporales asociadas a la vecindad $W \subset M^{(1)}$. En la Figura 4.14 se aprecian estas diferencias, incluso se puede observar que algunas tienen máximos donde otras presentan mínimos. Por el contrario, las series de tiempo asociadas a $V' \subset M^{(1,5)}$ son más homogéneas que las series temporales asociadas a W, pero preservan el comportamiento general. Debemos notar que, en algunos puntos de las montañas, la intensidad del viento aumenta bruscamente alcanzando sus valores más altos, contrastando con el comportamiento medio o usual. Este efecto produce que, en estas zonas, el suavizamiento provocado por el interpolador sea mayor que en el mar y en los valles.

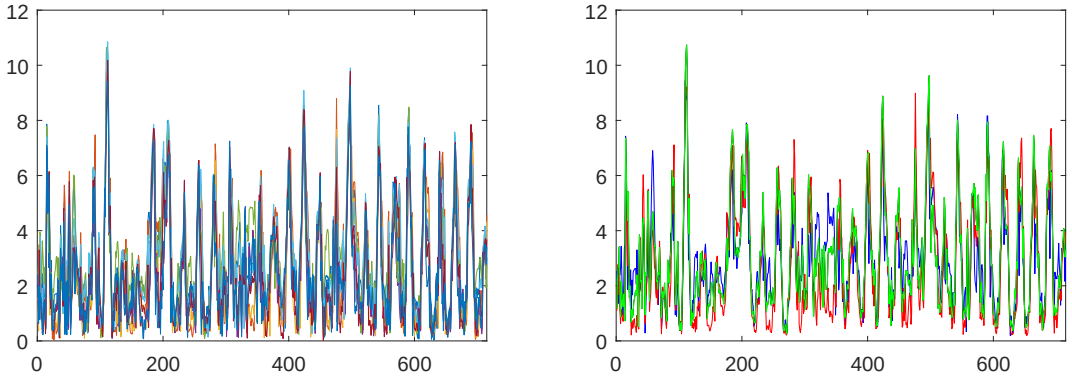


FIGURA 4.13: Series temporales para los elementos pertenecientes a $W \subset M^{(1)}$ y $V' \subset M^{(1,5)}$ al interior de la vecindad V ubicada en un valle al interior de la región y con coordenadas geográficas $32°44'9{,}6''$ y $32°45'47{,}88''$ latitud sur; $72°8'46{,}32''$ y $72°10'43{,}32''$ longitud oeste. A la izquierda se muestran las series correspondientes a W y a la derecha las correspondientes a V'.

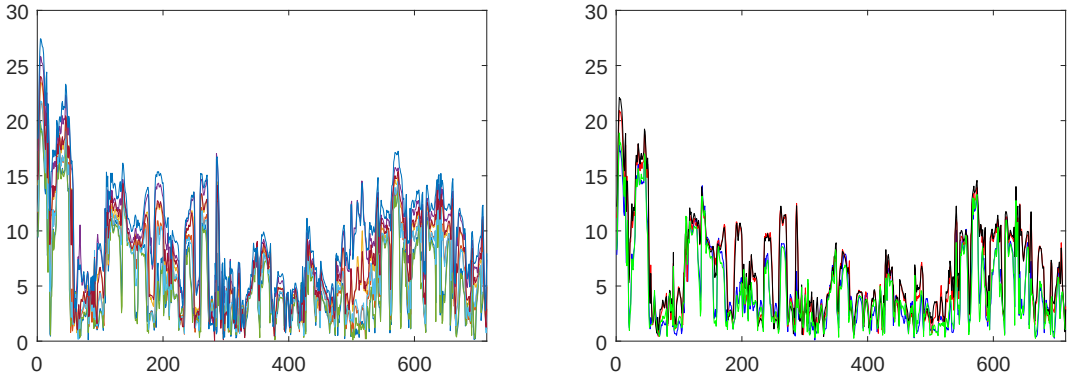


FIGURA 4.14: Series temporales para los elementos pertenecientes a $W \subset M^{(1)}$ y $V' \subset M^{(1,5)}$ al interior de la vecindad V ubicada en una zona montañosa de la región y con coordenadas geográficas $32°16'52{,}68''$ y $32°18'30{,}96''$ latitud sur; $71°41'31{,}2''$ y $71°43'27{,}12''$ longitud oeste. A la izquierda se muestran las series correspondientes a W y a la derecha las correspondientes a V'.

En conclusión, recalcamos que el interpolador bicúbico muestra sus buenas propiedades de manera espacial y, en general, preserva el comportamiento de las series temporales de manera local. Esto permite reflejar el traspaso de información desde la

malla a 3[Km] a mallas de mayor resolución, a través de un paso de upscaling que permite obtener una malla a 1.5[Km].

4.3.2. Regresión Lineal: 3km vs 1.5km

En esta sección analizaremos el modelo de regresión lineal ajustado entre las series temporales asociadas a los elementos de $V \subset M^{(3)}$ y $V' \subset M^{(1,5)}$. Justificaremos el uso de esta técnica mediante análisis exploratorio, compararemos espacial y temporalmente las estimaciones provenientes de la regresión y los resultados dados por la interpolación bicúbica.

4.3.2.1. Análisis exploratorio

Comenzamos realizando el análisis exploratorio en cada una de las tres vecindades mencionadas anteriormente. En la vecindad localizada en el mar, dada la gran homogeneidad presente en la zona, se espera que las series temporales asociadas a las vecindades $V \subset M^{(3)}$ y $V' \subset M^{(1,5)}$ presenten una marcada relación lineal. La Figura 4.15 ilustra la dependencia lineal existente entre dichas vecindades, apreciándose poca dispersión con respecto a la recta de regresión. La tabla 4.1 muestra los coeficientes de correlación lineal entre los elementos de la vecindad V, los cuales están en el intervalo $[0{,}9927, 1]$; también muestra los coeficientes de correlación lineal entre los elementos de V y V', los cuales están dentro del intervalo $[0{,}9846, 0{,}9904]$. De esta manera, es razonable asumir dependencia lineal en las vecindades ubicadas en el mar.

	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}
x_{s_1}	1,0000	0,9957	0,9950	0,9975
x_{s_2}	0,9957	1,0000	0,9973	0,9927
x_{s_3}	0,9950	0,9973	1,0000	0,9952
x_{s_4}	0,9975	0,9927	0,9952	1,0000

	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}
$x_{s'_1}$	0,9904	0,9881	0,9874	0,9885
$x_{s'_2}$	0,9869	0,9873	0,9866	0,9846
$x_{s'_3}$	0,9884	0,9875	0,9873	0,9868
$x_{s'_4}$	0,9900	0,9870	0,9872	0,9889

TABLA 4.1: Matrices de correlaciones para una vecindad $V \subset M^{(3)}$ ubicada en el mar, con coordenadas geográficas 31°37′1,56″ y 31°38′40,92″ latitud sur; 72°57′10,44″ y 72°59′7,08″ longitud oeste. A la izquierda se muestra la matriz de correlación para los elementos asociados a V y a la derecha la matriz de correlación entre los elementos asociados a V y $V' \subset M^{(1,5)}$.

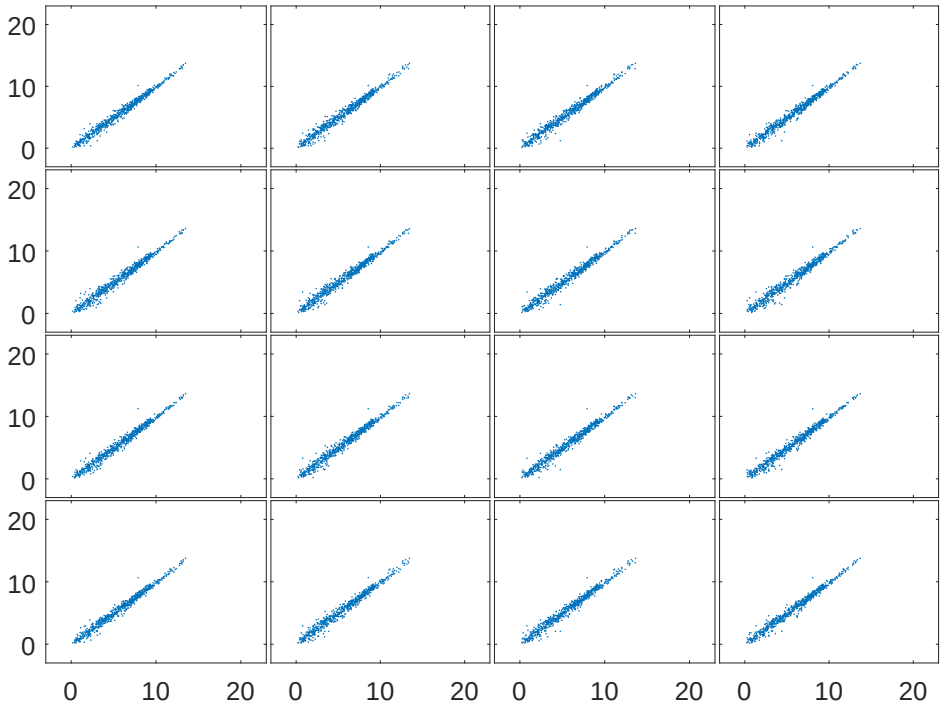


FIGURA 4.15: Gráficos de puntos de las series temporales de la intensidad del viento correspondientes a la vecindad $V \subset M^{(3)}$ contra los elementos de la vecindad $V' \subset M^{(1,5)}$ al interior de V. Estas vecindades están ubicadas al interior del mar, con coordenadas geográficas 31°37′1,56″ y 31°38′40,92″ latitud sur; 72°57′10,44″ y 72°59′7,08″ longitud oeste.

En la vecindad ubicada en el valle, los diagramas de dispersión presentes en la Figura 4.16 muestran que existe una relación lineal entre las series temporales asociadas a las vecindades $V \subset M^{(3)}$ y $V' \subset M^{(1,5)}$, la disperción con respecto a la recta de regresión es mayor que en el caso anterior, pero es homogénea en el rango de valores de intensidad del viento. La tabla 4.2 presenta los coeficientes de correlación lineal entre los elementos de la vecindad V, los cuales están en el intervalo $[0{,}8834, 1]$; también muestra los coeficientes de correlación lineal entre los elementos de V y V', los cuales están dentro del intervalo $[0{,}7566, 0{,}8891]$. La hipótesis de linealidad local se satisface, a pesar de que los datos en esta zona presenten mayor variabilidad.

	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}
x_{s_1}	1,0000	0,9250	0,8834	0,9252
x_{s_2}	0,9250	1,0000	0,9240	0,9134
x_{s_3}	0,8834	0,9240	1,0000	0,9497
x_{s_4}	0,9252	0,9134	0,9497	1,0000

	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}
$x_{s'_1}$	0,8591	0,7871	0,7566	0,8143
$x_{s'_2}$	0,8218	0,8649	0,7709	0,7873
$x_{s'_3}$	0,8891	0,9093	0,8702	0,8707
$x_{s'_4}$	0,8551	0,7664	0,7840	0,8359

TABLA 4.2: Matrices de correlaciones para una vecindad $V \subset M^{(3)}$ ubicada en un valle al interior de la región, con coordenadas geográficas 32°44′9,6″ y 32°45′47,88″ latitud sur; 72°8′46,32″ y 72°10′43,32″ longitud oeste. A la izquierda se muestra la matriz de correlación para los elementos asociados a V y a la derecha la matriz de correlación entre los elementos asociados a V y $V' \subset M^{(1,5)}$.

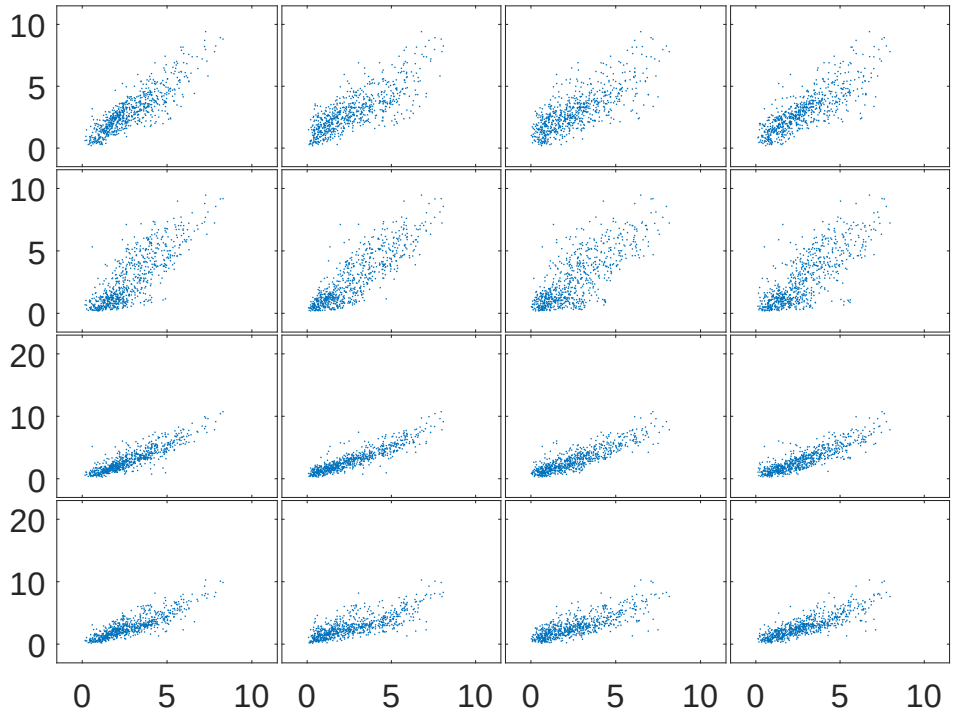


FIGURA 4.16: Gráficos de puntos de las series temporales de la intensidad del viento correspondientes a la vecindad $V \subset M^{(3)}$ contra los elementos de la vecindad $V' \subset M^{(1,5)}$ al interior de V. Estas vecindades están ubicadas en un valle al interior de la región, con coordenadas geográficas 32°44′9,6″ y 32°45′47,88″ latitud sur; 72°8′46,32″ y 72°10′43,32″ longitud oeste.

Para la vecindad presente en la montaña se observa en la Figura 4.17 que existe relación lineal entre las series temporales asociadas a las vecindades $V \subset M^{(3)}$ y $V' \subset M^{(1,5)}$. Se observa una gran dispersión de los datos con respecto a la recta de regresión, además de una variabilidad no homogénea, es decir, a intensidades del viento

mayores la dispersión aumenta. La tabla 4.3 presenta los coeficientes de correlación lineal entre los elementos de la vecindad V, los cuales están en el intervalo $[0{,}7891, 1]$; también muestra los coeficientes de correlación lineal entre los elementos de V y V', los cuales están dentro del intervalo $[0{,}6410, 0{,}9559]$. En líneas generales, en esta zona suponer dependencia lineal entre las mallas a 3[Km] y 1.5[Km] es razonable, pero la hipótesis de de homogeneidad local espacial es más débil debido a sus características geográficas.

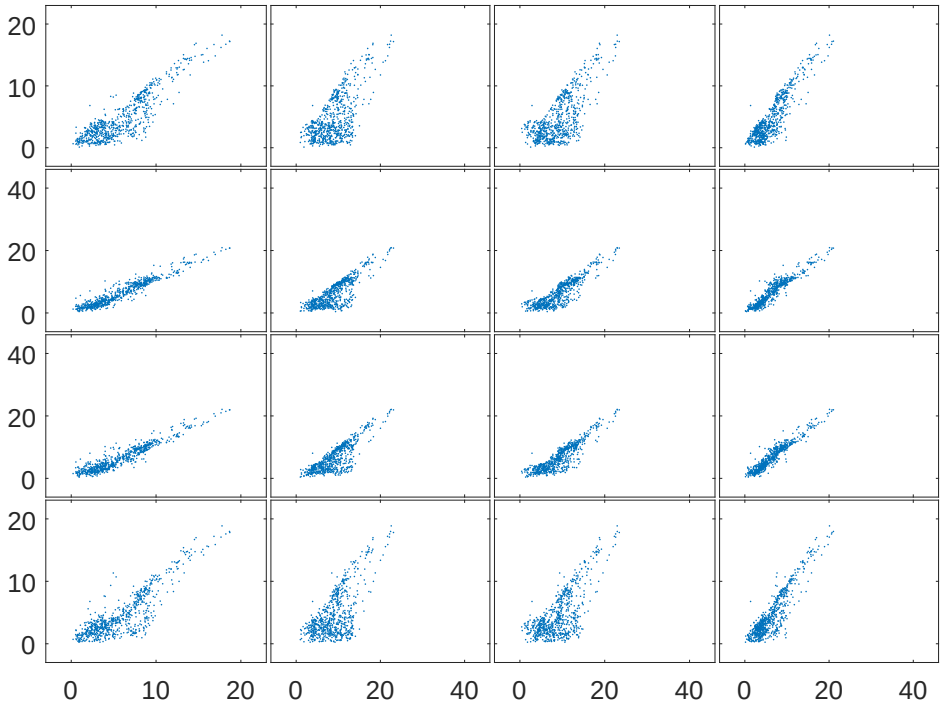


FIGURA 4.17: Gráficos de puntos de las series temporales de la intensidad del viento correspondientes a la vecindad $V \subset M^{(3)}$ contra los elementos de la vecindad $V' \subset M^{(1,5)}$ al interior de V. Estas vecindades están ubicadas en una zona montañosa de la región, con coordenadas geográficas $32°16'52{,}68''$ y $32°18'30{,}96''$ latitud sur; $71°41'31{,}2''$ y $71°43'27{,}12''$ longitud oeste.

En general observamos que la hipótesis de linealidad está bien justificada y que la topografía influye en la dispersión de los datos. En el mar hay muy poca dispersión, en el valle se presenta una dispersión cónica y en la montaña la dispersión parece tener relación con la intensidad del viento.

	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}
x_{s_1}	1,0000	0,8025	0,8760	0,9741
x_{s_2}	0,8025	1,0000	0,9641	0,7891
x_{s_3}	0,8760	0,9641	1,0000	0,8740
x_{s_4}	0,9741	0,7891	0,8740	1,0000

	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}
$x_{s'_1}$	0,8787	0,6495	0,7178	0,9005
$x_{s'_2}$	0,9559	0,8092	0,8825	0,9449
$x_{s'_3}$	0,9448	0,8161	0,8898	0,9524
$x_{s'_4}$	0,8617	0,6410	0,7107	0,9006

TABLA 4.3: Matrices de correlaciones para una vecindad $V \subset M^{(3)}$ ubicada en una zona montañosa, con coordenadas geográficas 32°16′52,68″ y 32°18′30,96″ latitud sur; 71°41′31,2″ y 71°43′27,12″ longitud oeste. A la izquierda se muestra la matriz de correlación para los elementos asociados a V y a la derecha la matriz de correlación entre los elementos asociados a V y $V' \subset M^{(1,5)}$.

4.3.2.2. Análisis espacial

Presentamos los resultados para el modelo de regresión lineal construido para explicar el campo de viento a 1.5[Km] mediante el campo de viento a 3[Km] en región de Valparaíso. Siguiendo el esquema, presentamos los días 1, 21 y 29 del mes de Septiembre del año 2014. En cada día se presentan dos horas particulares, correspondientes a las 10:00 y 22:00 horas para el día 1/09/2014, 11:00 y 23:00 horas para los días 21/09/2014 y 29/09/2014, debido al cambio de horario que tuvo efecto el día 06/09/2014 y, para cada hora, se muestran cuatro imágenes por cada fecha y hora: El campo de intensidad de viento a 3[Km], el campo de intensidad de viento a 1.5[Km] obtenido mediante la interpolación bicúbica, el campo de intensidad de viento a 1.5[Km] estimado mediante el ajuste de regresión lineal, y el error de estimación entre la interpolación y el modelo de regresión. Notamos, en las Figuras 4.18, 4.19 y 4.20 que la regresión lineal produce un efecto de suavizamiento y perdida de contraste en comparación con la obtenida mediante el interpolador bicúbico, produciendo perdidas en el detalle de la imagen del campo de intensidad del viento. Sin embargo, la estimación mediante regresión lineal proveniente desde el campo de viento a 3[Km] considera menos información que la interpolación bicúbica desde el campo a 1[Km], pues este último posee nueve veces más información que el campo a 3[Km], a pesar de lo anterior el ajuste de regresión lineal local es adecuado para nuestros propósitos, ya que capta el comportamiento general del viento y el rango de intensidad. Se observa que el residuo conserva una estructura de dependencia espacial que no es extraída por el ajuste de regresión lineal, por ejemplo, el comportamiento presente en el mar (lado izquierdo del residuo) y en las cordilleras (lado derecho del residuo) son diferentes entre sí. Además, el residuo en la mayoría de los casos es centrado.

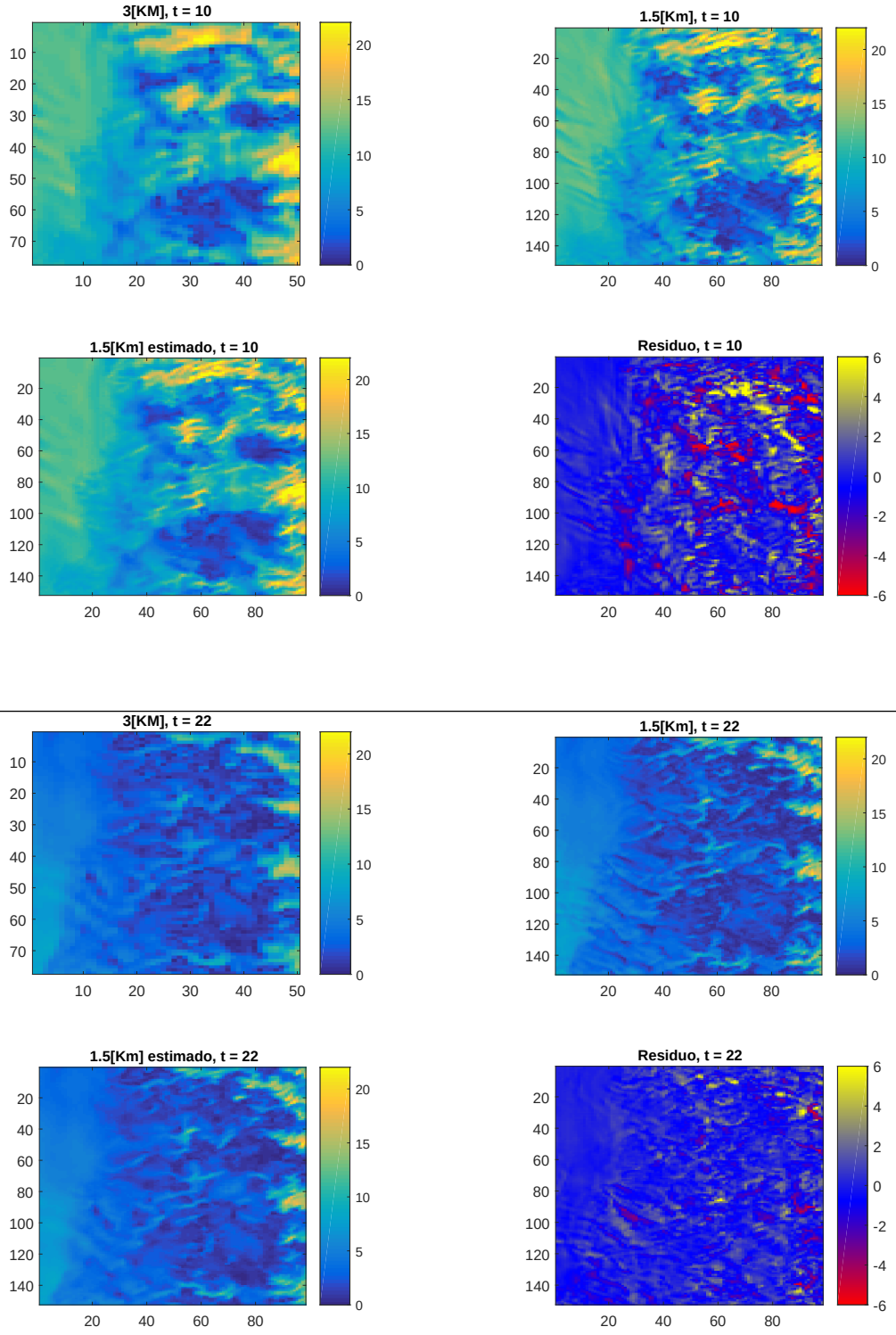


FIGURA 4.18: Campos de intensidad viento $X^{(1)}_{(\cdot,t_i)}$, $X^{(1,5)}_{(\cdot,t_i)}$, $\hat{X}^{(1)}_{(\cdot,t_i)}$ y $R_{(1,5:3)}$, con $i = 1, 2$ donde los tiempos $t_1 = 10$ y $t_2 = 22$ corresponden al día 01 de Septiembre de 2014 a las 10:00 y 22:00 horas respectivamente.

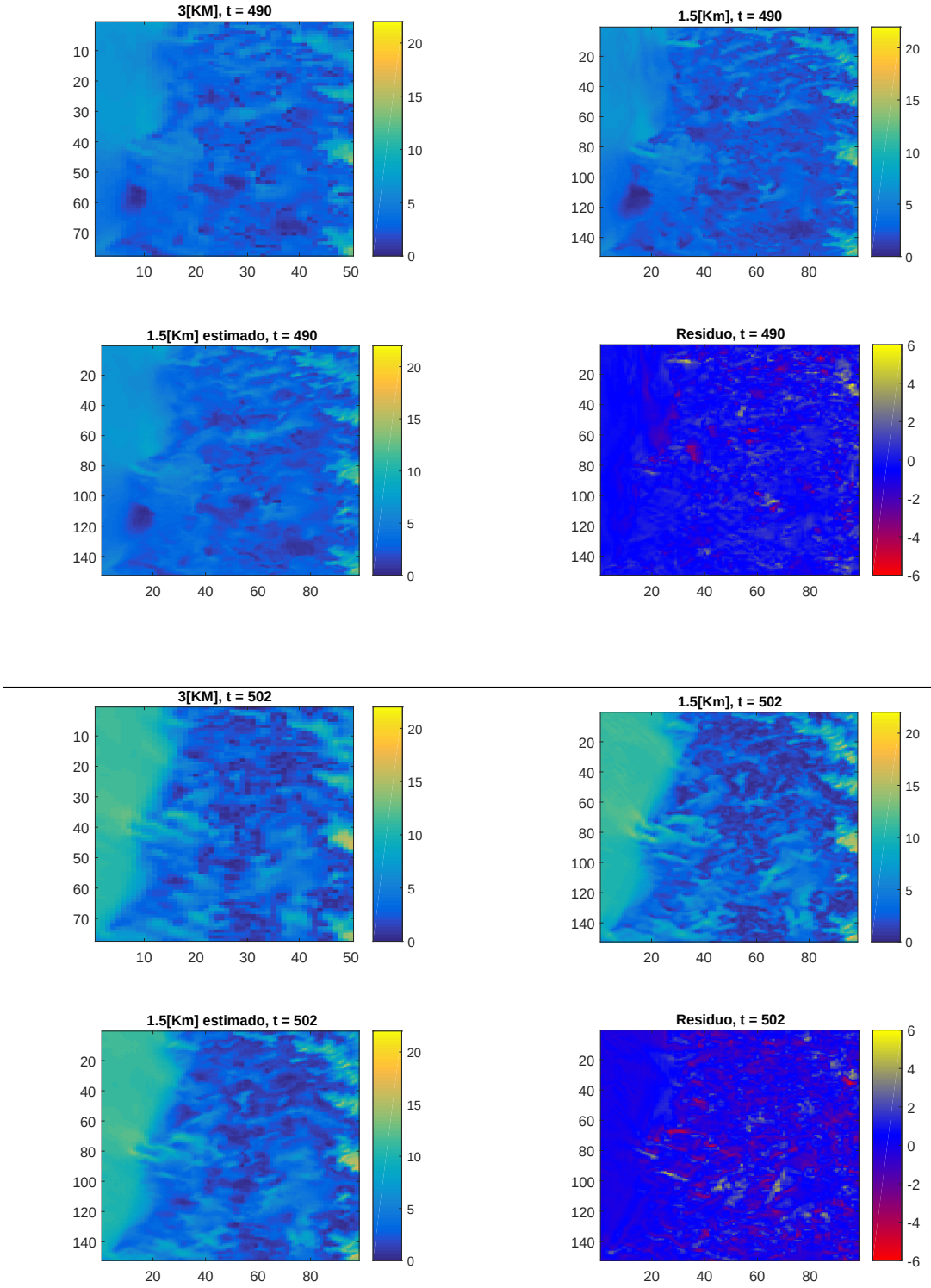


FIGURA 4.19: Campos de intensidad viento $X^{(1)}_{(\cdot,t_i)}$, $X^{(1,5)}_{(\cdot,t_i)}$, $\hat{X}^{(1)}_{(\cdot,t_i)}$ y $R_{(1,5:3)}$, con $i = 1, 2$ donde los tiempos $t_1 = 490$ y $t_2 = 502$ corresponden al día 21 de Septiembre de 2014 a las 11:00 y 23:00 horas respectivamente.

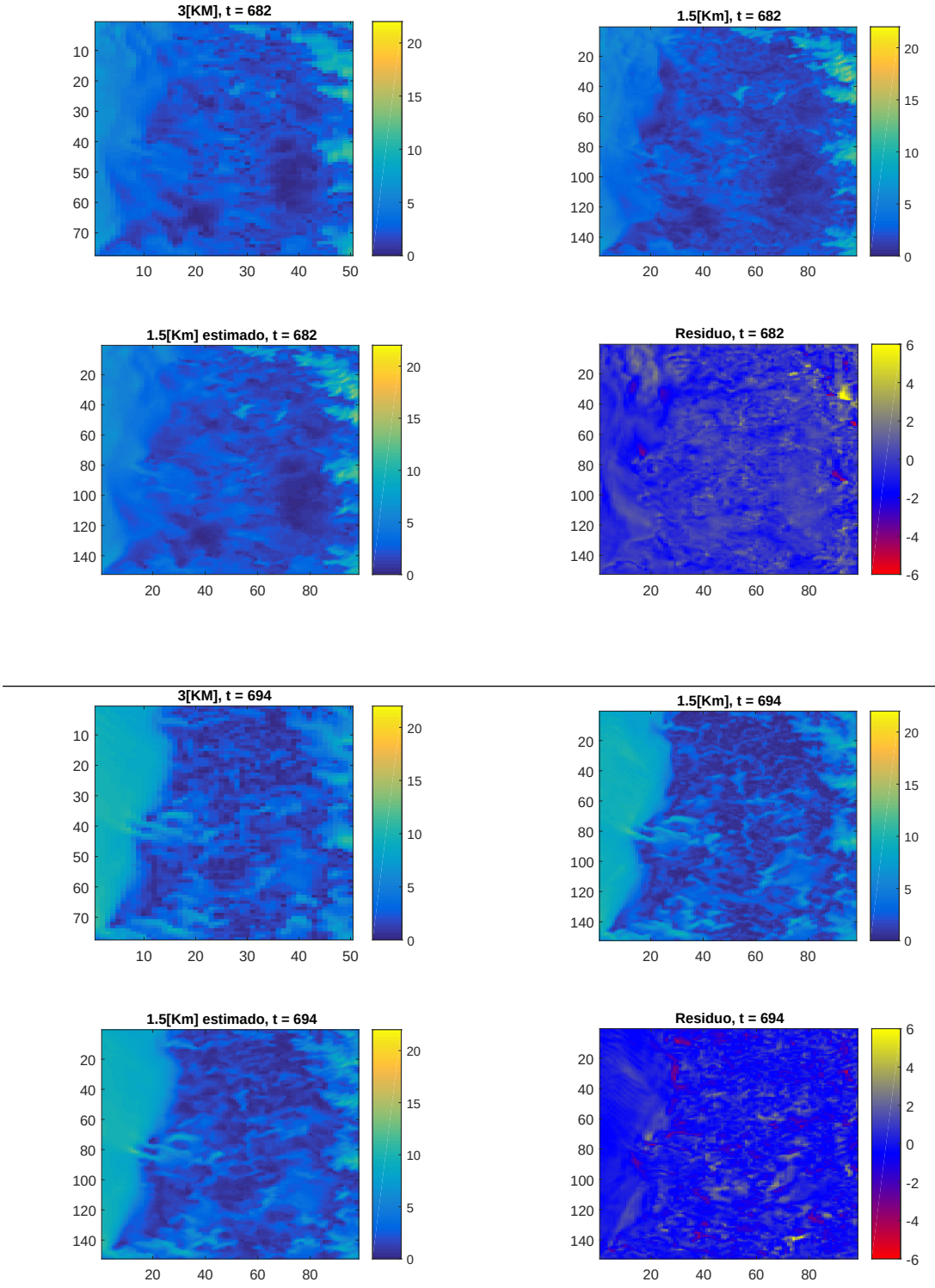


FIGURA 4.20: Campos de intensidad viento $X^{(1)}_{(\cdot,t_i)}$, $X^{(1,5)}_{(\cdot,t_i)}$, $\hat{X}^{(1)}_{(\cdot,t_i)}$ y $R_{(1,5:3)}$, con $i = 1, 2$ donde los tiempos $t_1 = 682$ y $t_2 = 694$ corresponden al día 29 de Septiembre de 2014 a las 11:00 y 23:00 horas respectivamente.

Realizamos un análisis exploratorio del error de estimación mediante histogramas y gráficos de caja o boxplot. En las Figuras 4.21, 4.22 y 4.23 apreciamos los histogramas y boxplot para el error de estimación en cada uno de los tiempos mostrados anteriormente. Los histogramas muestran que la distribución del error de estimación es simétrica, mientras que los gráficos de caja son muestran una gran cantidad de valores extremos o atípicos, evidenciando que la distribución espacial del error de estimación posee colas pesadas, por lo que la distribución espacia de los datos no siguen una distribución normal.

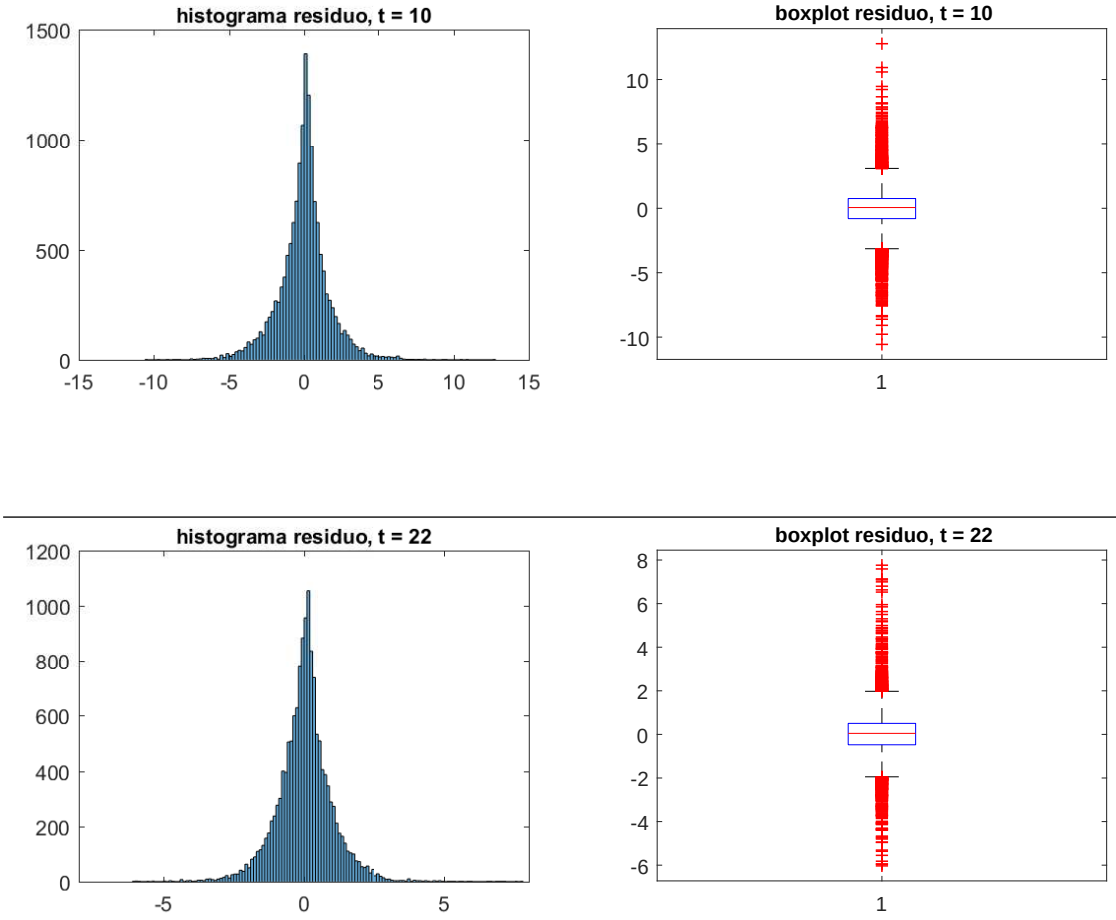


FIGURA 4.21: Histogramas y boxplot para el residuo $R_{(1,5:3)}$ de la intensidad viento producido por el modelo de regresión lineal entre los campos a 1.5[Km] y 3[Km], para los tiempos $t_1 = 10$ y $t_2 = 22$ correspondientes al día 01 de Septiembre de 2014 a las 10:00 y 22:00 horas respectivamente.

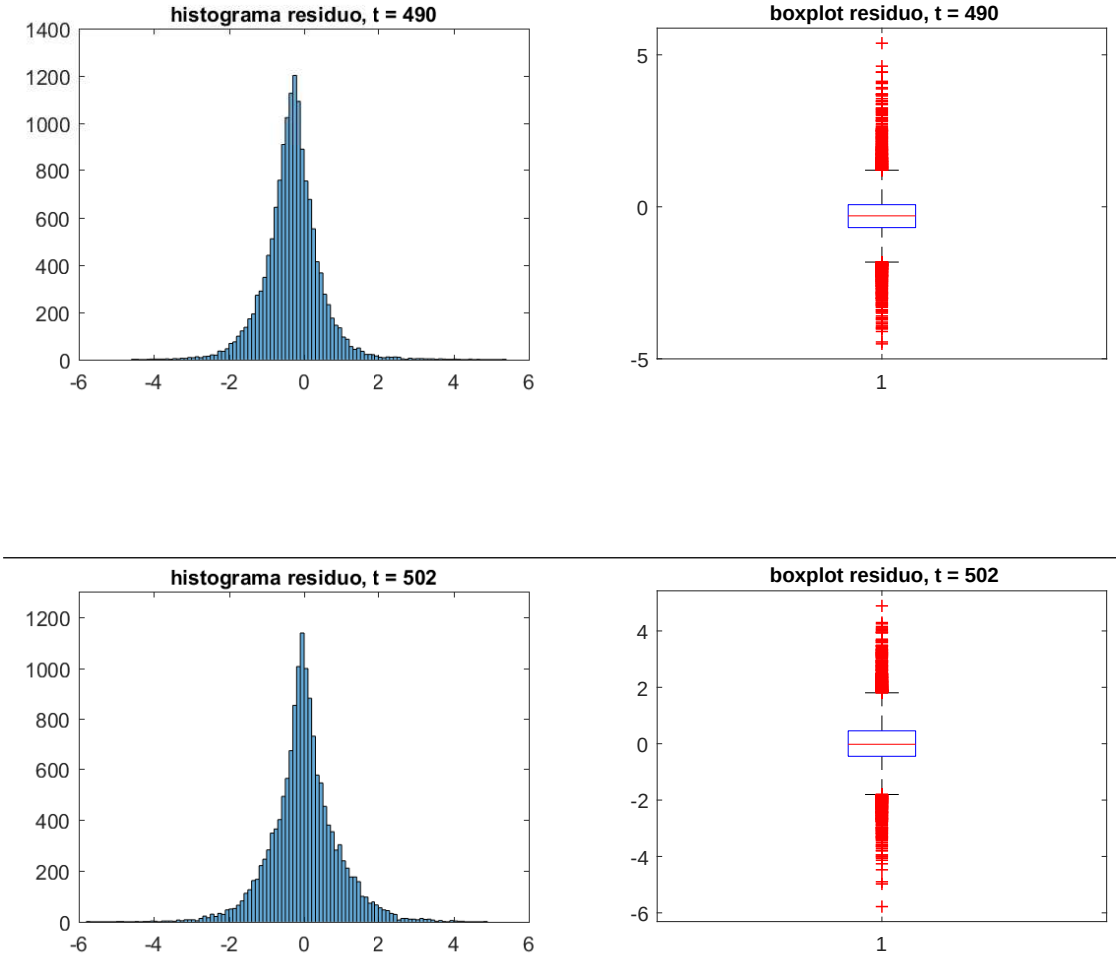


FIGURA 4.22: Histogramas y boxplot para el residuo $R_{(1,5:3)}$ de la intensidad viento producido por el modelo de regresión lineal entre los campos a 1.5[Km] y 3[Km], para los tiempos $t_1 = 490$ y $t_2 = 502$ correspondientes al día 21 de Septiembre de 2014 a las 11:00 y 23:00 horas respectivamente.

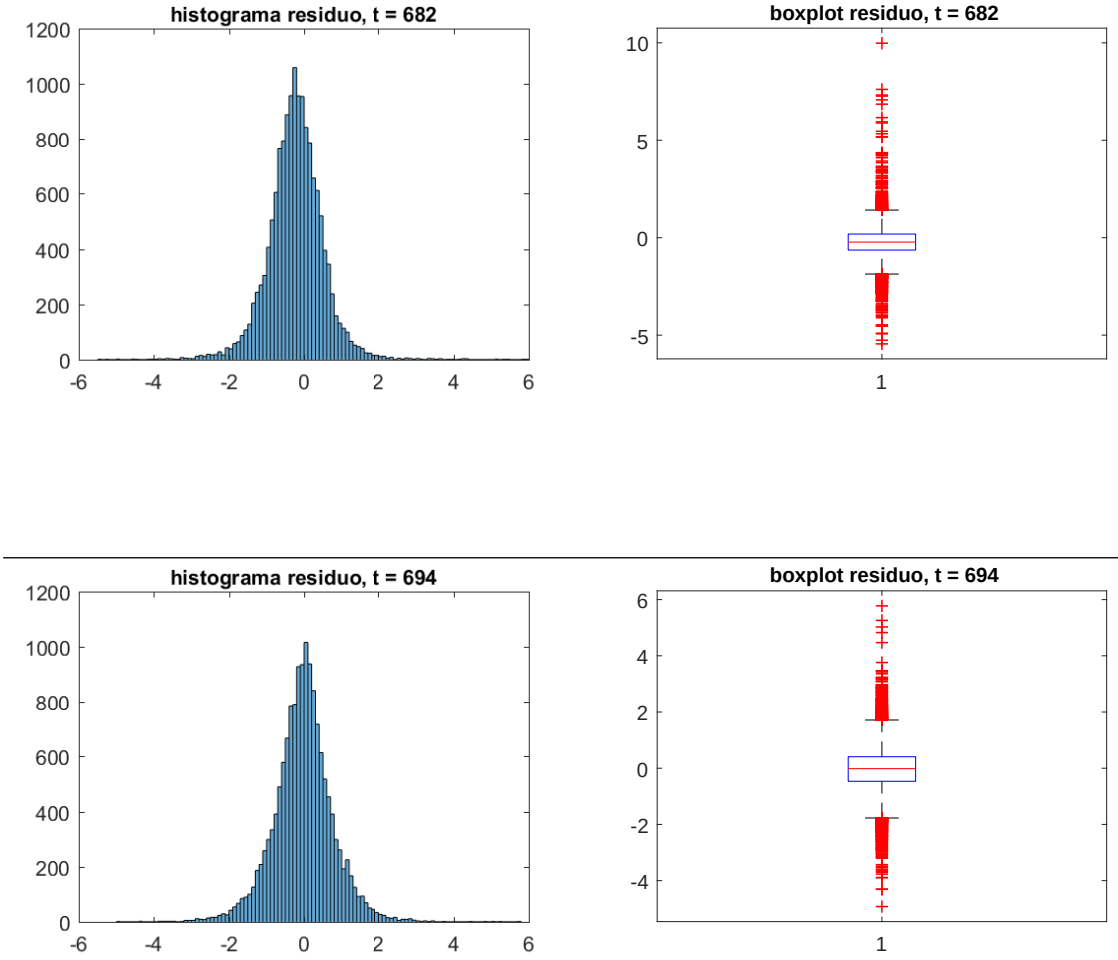


FIGURA 4.23: Histogramas y boxplot para el residuo $R_{(1,5:3)}$ de la intensidad viento producido por el modelo de regresión lineal entre los campos a 1.5[Km] y 3[Km], para los tiempos $t_1 = 682$ y $t_2 = 694$ correspondientes al día 29 de Septiembre de 2014 a las 11:00 y 23:00 horas respectivamente.

Dado lo anterior, se debe buscar alternativas a las técnicas clásicas para estudiar la dependencia especial del residuo. Nuestra propuesta consiste en usar técnicas de clasificación en el espacio de las distribuciones de las series de tiempo del residuo, y ver si esta guarda relación con las características geográficas de la zona de estudio.

4.3.2.3. Análisis temporal

A continuación, presentamos el estudio temporal del modelo de regresión lineal. Comenzamos presentando las series temporales en cada una de las tres vecindades, siguiendo el orden de los puntos presentado en la Figura 4.24, es decir, para cada vecindad generamos cuatro gráficos, cada uno muestra las series temporales de las vecindades $V \subset M^{(3)}$, $V' \subset M^{(1,5)}$ y las series estimadas en V' correspondientes a las posiciones $s_i \in V$ y $s'_i \in V'$ con $i = \{1, \ldots, 4\}$.

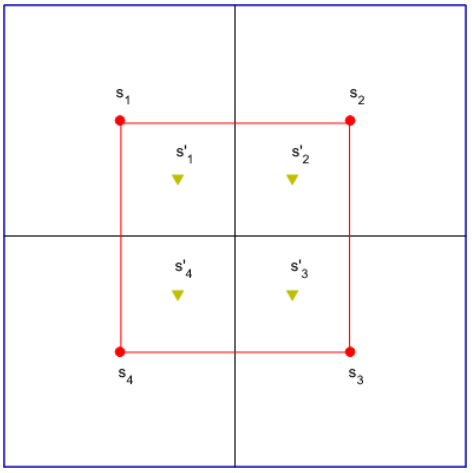


FIGURA 4.24: Ubicación de las posiciones pertenecientes a la vecindad $V \subset M^{(3)}$, representados por los puntos rojos, y las posiciones pertenecientes a la vecindad $V' \subset M^{(1,5)}$, representados por los triángulos verdes. La vecindad V' se ubica al interior de la vecindad V.

En la Figura 4.25 observamos los gráficos para las cuatro posiciones geográficas pertenecientes a la vecindad ubicada en el mar. Apreciamos que el comportamiento de las series temporales a 3[Km] $\left(x_{s_i} \in X^{(3)}, i = \{1, \ldots, 4\}\right)$ son similares a las de las series temporales a 1.5[Km] $\left(x_{s'_i} \in X^{(1,5)},\ i = \{1, \ldots, 4\}\right)$, debido a la homogeneidad del viento en esta zona, produciendo que las series temporales estimadas a 1.5[Km] $\left(\hat{x}_{s'_i} \in \hat{X}^{(1,5)}, i = \{1, \ldots, 4\}\right)$ también sean similares a la de las series de tiempo originales, salvo el suavizamiento de la regresión lineal, que hace que los máximos de las series temporales estimadas estén levemente por debajo. Recalcamos que los valores de las series temporales estimadas y las originales están dentro del mismo rango de valores.

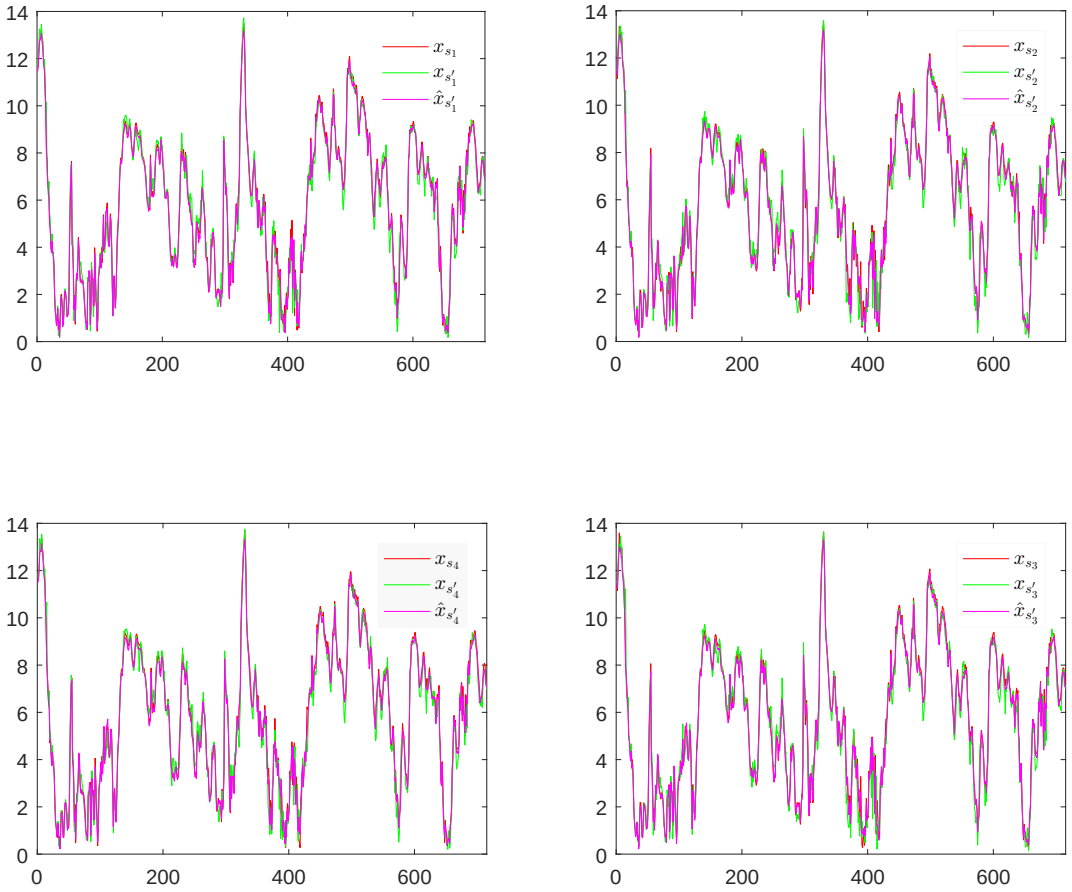


FIGURA 4.25: Series temporales dentro de la vecindad $V \subset M^{(3)}$ ubicada en el mar con coordenadas geográficas 31°37′1,56″ y 31°38′40,92″ latitud sur; 72°57′10,44″ y 72°59′7,08″ longitud oeste. Cada gráfico muestra: La serie temporal asociada a la posición correspondiente $s_i \in V$, además de la serie temporal asociada a la posición $s'_i \in V'$ junto con su estimación, para $i = \{1, \ldots, 4\}$.

En la Figura 4.26 apreciamos los gráficos para las cuatro posiciones geográficas pertenecientes a la vecindad ubicada en el valle. En estas zonas el comportamiento de la intensidad del viento también es homogéneo, similar al comportamiento en el mar, por lo que nuevamente las series temporales a 3[Km] $\left(x_{s_i} \in X^{(3)}, i = \{1, \ldots, 4\}\right)$ son parecidas a las de las series temporales a 1.5[Km] $\left(x_{s'_i} \in X^{(1,5)},\ i = \{1, \ldots, 4\}\right)$. Los resultados para las series temporales estimadas a 1.5[Km] $\left(\hat{x}_{s'_i} \in \hat{X}^{(1,5)}, i = \{1, \ldots, 4\}\right)$ son concordantes con el caso de la vecindad en el mar, es decir, captan el comportamiento de las series originales, sin embargo en los puntos extremos el efecto de suavizamiento es mayor, y se encuentran dentro del mismo rango de valores.

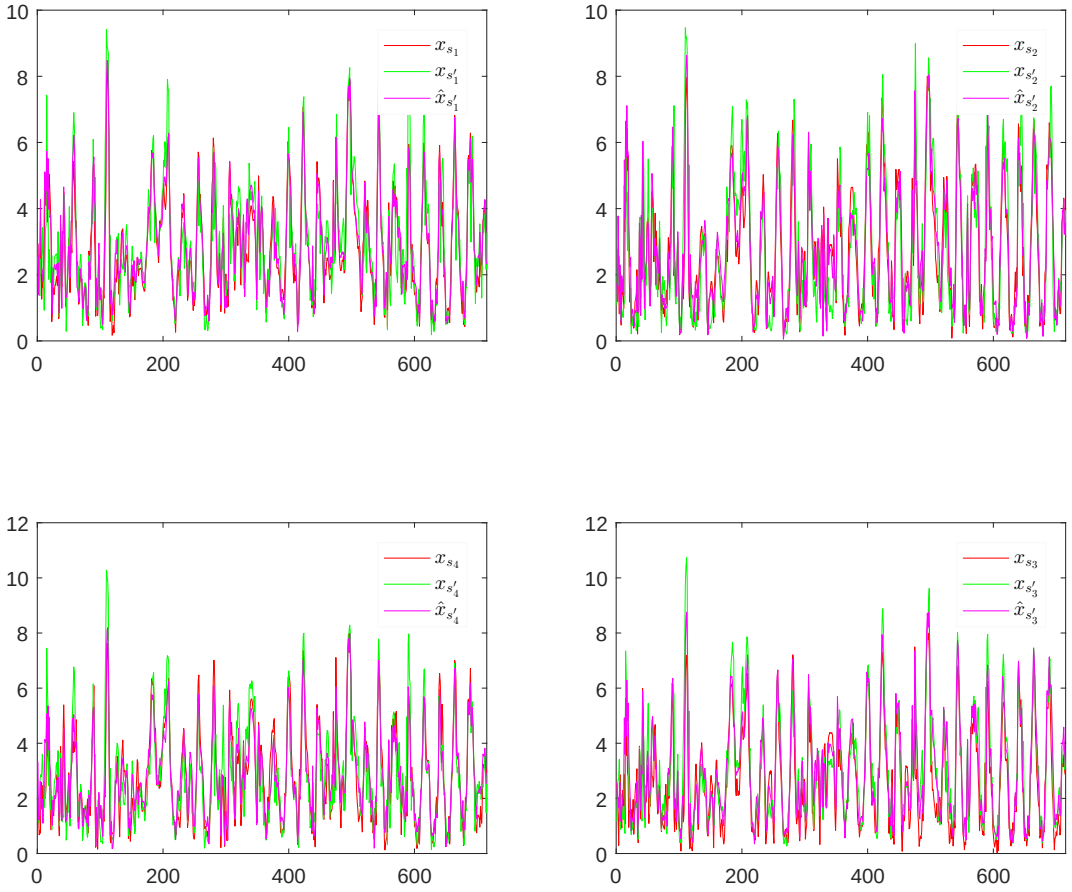


FIGURA 4.26: Series temporales dentro de la vecindad $V \subset M^{(3)}$ ubicada en el valle con coordenadas geográficas 32°44′9,6″ y 32°45′47,88″ latitud sur; 72°8′46,32″ y 72°10′43,32″ longitud oeste. Cada gráfico muestra: La serie temporal asociada a la posición correspondiente $s_i \in V$, además de la serie temporal asociada a la posición $s'_i \in V'$ junto con su estimación, para $i = \{1, \ldots, 4\}$.

En la Figura 4.27 se muestran los gráficos para las cuatro posiciones geográficas pertenecientes a la vecindad ubicada en la montaña. En esta zona el comportamiento para campo de intensidad del viento es más heterogéneo, lo que produce diferencias entre las series temporales a 3[Km] $\left(x_{s_i} \in X^{(3)}, i = \{1, \ldots, 4\}\right)$ y las series temporales a 1.5[Km] $\left(x_{s'_i} \in X^{(1,5)},\ i = \{1, \ldots, 4\}\right)$, debido a la diferencia de las posiciones geográficas (ver Figura 4.24). A pesar de lo anterior, las series de tiempo estimadas a 1.5[Km] $\left(\hat{x}_{s'_i} \in X^{(1,5)},\ i = \{1, \ldots, 4\}\right)$, captan el comportamiento general de las series originales, aunque dada la diferencia que se produce entre las series temporales a 3[Km] y 1.5[Km], la serie estimada puede presentar valores de intensidad negativo

debido al ajuste del modelo de regresión, lo que obliga a considerar un truncamiento a cero de estos valores, para asegurar que la intensidad de viento estimada sea positiva, estos valores negativos son de baja magnitud respecto al rango de los datos, por lo que al truncar estos valores no se produce un error considerable.

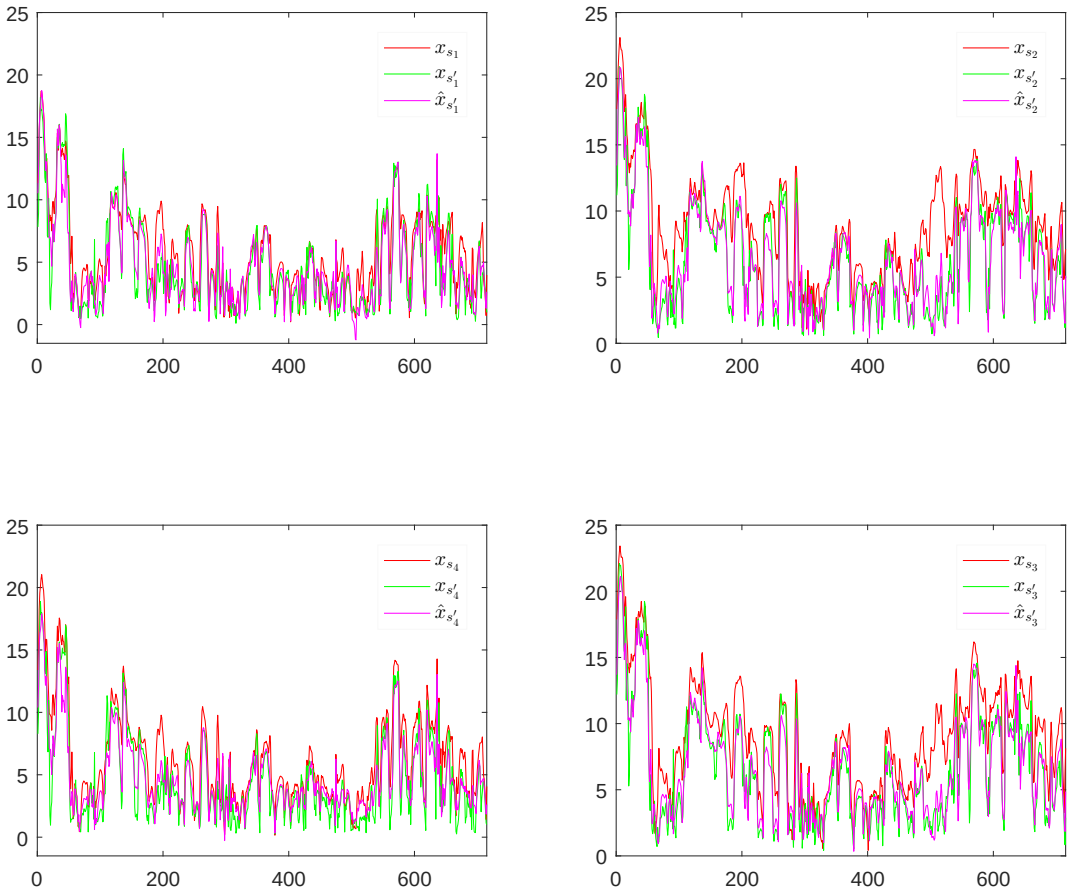


FIGURA 4.27: Series temporales dentro de la vecindad $V \subset M^{(3)}$ ubicada en la montaña con coordenadas geográficas $32°16'52{,}68''$ y $32°18'30{,}96''$ latitud sur; $71°41'31{,}2''$ y $71°43'27{,}12''$ longitud oeste. Cada gráfico muestra: La serie temporal asociada a la posición correspondiente $s_i \in V$, además de la serie temporal asociada a la posición $s'_i \in V'$ junto con su estimación, para $i = \{1, \ldots, 4\}$.

Las tablas 4.4, 4.5 y 4.6 muestran los errores cuadráticos medios y el estadístico R^2 (ver Capítulo 3.2.1) para el ajuste de regresión lineal en cada una de las posiciones pertenecientes a las vecindades $V' \subset M^{(1,5)}$. Observamos que, a medida que la zona pierde regularidad geográfica, el error cuadrático medio de la estimación aumenta, lo que concuerda con las conclusiones previas del análisis exploratorio espacial. Aun así,

el valor del estadístico R^2 parece no aumentar con respecto a la homogeneidad de las vecindades, debido a que este estadístico guarda relación con la variabilidad de los datos. En general el ajuste de regresión lineal da buenos resultados temporales, entregando valores del estadístico R^2 en el rango [0,78 0,98].

Vecindad 1	s'_1	s'_2	s'_3	s'_4
ECM	0.1573	0.1945	0.1923	0.1624
R^2	0.9821	0.9778	0.9783	0.9815

TABLA 4.4: Error cuadrático medio y R^2 de la regresión lineal para las posiciones $s'_i, i = \{1, \ldots, 4\}$ de la vecindad $V' \subset M^{(1,5)}$ ubicada en el mar.

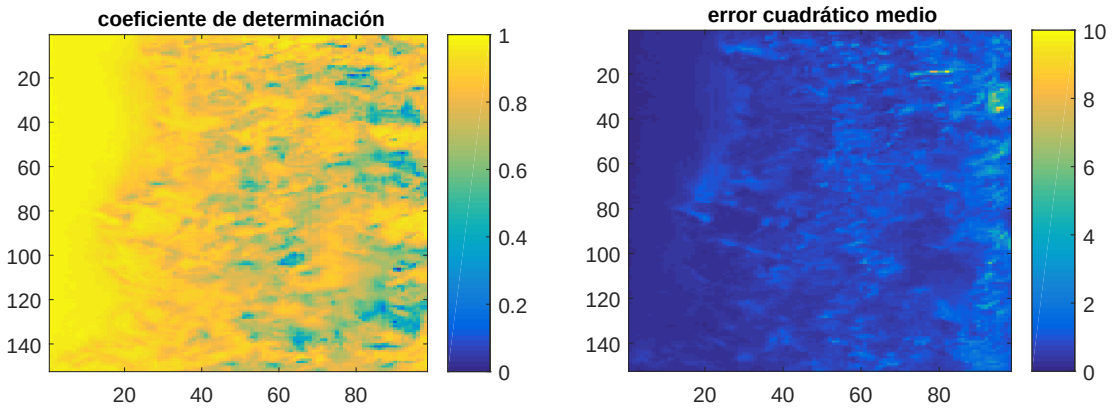
Vecindad 2	s'_1	s'_2	s'_3	s'_4
ECM	0.5778	0.8437	0.5725	0.6584
R^2	0.7825	0.8019	0.8446	0.7815

TABLA 4.5: Error cuadrático medio y R^2 de la regresión lineal para las posiciones $s'_i, i = \{1, \ldots, 4\}$ de la vecindad $V' \subset M^{(1,5)}$ ubicada en el valle.

Vecindad 3	s'_1	s'_2	s'_3	s'_4
ECM	1.6382	1.4103	1.4699	1.8589
R^2	0.8775	0.9174	0.9176	0.8589

TABLA 4.6: Error cuadrático medio y R^2 de la regresión lineal para las posiciones $s'_i, i = \{1, \ldots, 4\}$ de la vecindad $V' \subset M^{(1,5)}$ ubicada en la montaña.

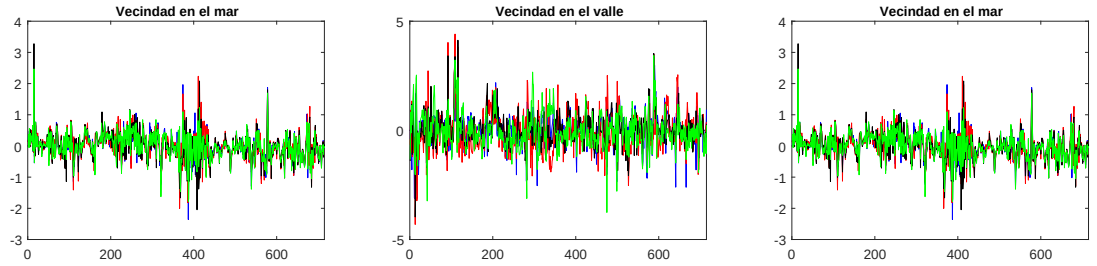
En la Figura 4.28 se muestran imágenes para todos los valores del coeficiente de determinación y error cuadrático medio del ajuste de regresión lineal sobre la malla $M^{(1,5)}$, apreciando que en general este ajuste presenta valores cercanos a uno para el coeficiente de determinación y cercanos a cero para el error cuadrático medio, por lo que en gran parte de la región este modelo es adecuado. Es conveniente resaltar que las zonas en donde el ajuste parece presentar problemas es en las zonas cordilleranas de la región (lado derecho de las imágenes), lo cual puede deberse a la heterogeneidad de estas zonas en cuanto a su topografía.



(a) (b)

FIGURA 4.28: Contraste global del ajuste de regresión lineal sobre el campo de intensidad de viento $M^{(1,5)}$ mediante: (a) El coeficiente de determinación, (b) El error cuadrático medio.

A continuación analizamos las series temporales del residuo $R(1{,}5:3)$, enfocándonos en las dos propiedades usuales que se desean para el error de estimación a nivel temporal: normalidad y estacionaridad, ya sea en el sentido débil o fuerte, recordando que si una serie temporal se distribuye normal y es estacionaria en el sentido débil, también es estacionaria en el sentido fuerte. Comenzamos mostrando en la Figura 4.29 las cuatro series de tiempo del residuo para cada una de las vecindades estudiadas. Se observa que todas las series temporales están centradas en cero, como es de esperarse, pues la regresión lineal recupera el comportamiento promedio de las series temporales originales.



(a) (b) (c)

FIGURA 4.29: Series temporales para el residuo en las cuatro vecindades estudiadas: (a) Vecindad en el mar, (b) Vecindad en el valle y (c) Vecindad en la montaña.

En la bibliografía [23],[2] y [3] se pueden ver estudios detallados sobre el error de estimación para campos de intensidad del viento. En estos trabajos se suelen ajustar modelos estacionarios, como los modelos autoregresivos (ver Capítulo 3.2.3) o más generalmente modelos autoregresivos con cambios de régimen markovianos, los cuales permiten considerar diferentes variabilidades para cada régimen, considerando incluso algún régimen volátil manteniendo la estacionaridad del modelo, como por ejemplo el comportamiento de las series temporales en la Figura 4.29 (ver [23]). En dichos trabajos muestran como el ajuste de estos modelos a los residuos pasan el test de la raíz unitaria, concluyendo que las series temporales del residuo son estacionarias.

Para visualizar si los residuos temporales siguen una distribución normal, presentamos en las Figuras 4.30, 4.31 y 4.32 los histogramas y gráficos cuartil-cuartil (qqplot en inglés) para las cuatro series temporales del error de estimación en cada una de las vecindades estudiadas. Los gráficos cuartil-cuartil comparan los cuartiles de los datos con los cuartiles provenientes de una distribución normal; si la distribución de los datos es normal, el gráfico se visualiza como una linea recta, en caso contrario, presentan alguna curvatura o desviación con respecto a la recta central. Los histogramas reflejan que las distribuciones de los residuos presentan un comportamiento simétrico con respecto al valor medio, sin embargo, los gráficos cuartil-cuartil muestran curvatura en los extremos de la recta, evidenciando la existencia de colas pesadas, por lo que estas distribuciones no se comportan como la distribución normal. En la tabla 4.7 se muestran los resultados para el test de normalidad Jarque-Bera, el cual contrasta la curtosis y el coeficiente de simetría de los datos con una muestra proveniente de la distribución normal, dando el valor de 1 si los datos no provienen de una distribución normal o 0 si es posible que los datos provengan de una distribución normal. Apreciamos que todas las series temporales del residuo no pasan este test, restringiendo el uso de técnicas como kriging, campos gaussianos, análisis de componentes principales, entre otros.

	Vecindad 1				**Vecindad 2**				**Vecindad 3**			
	s'_1	s'_2	s'_3	s'_4	s'_1	s'_2	s'_3	s'_4	s'_1	s'_2	s'_3	s'_4
h	1	1	1	1	1	1	1	1	1	1	1	1
p	0.0036	0.0033	0.0036	0.0025	e-3	e-3	e-3	e-3	e-3	e-3	e-3	e-3

Tabla 4.7: Resultados para el test de normalidad de Jarque-Bera para cada una de las tres vecindades.

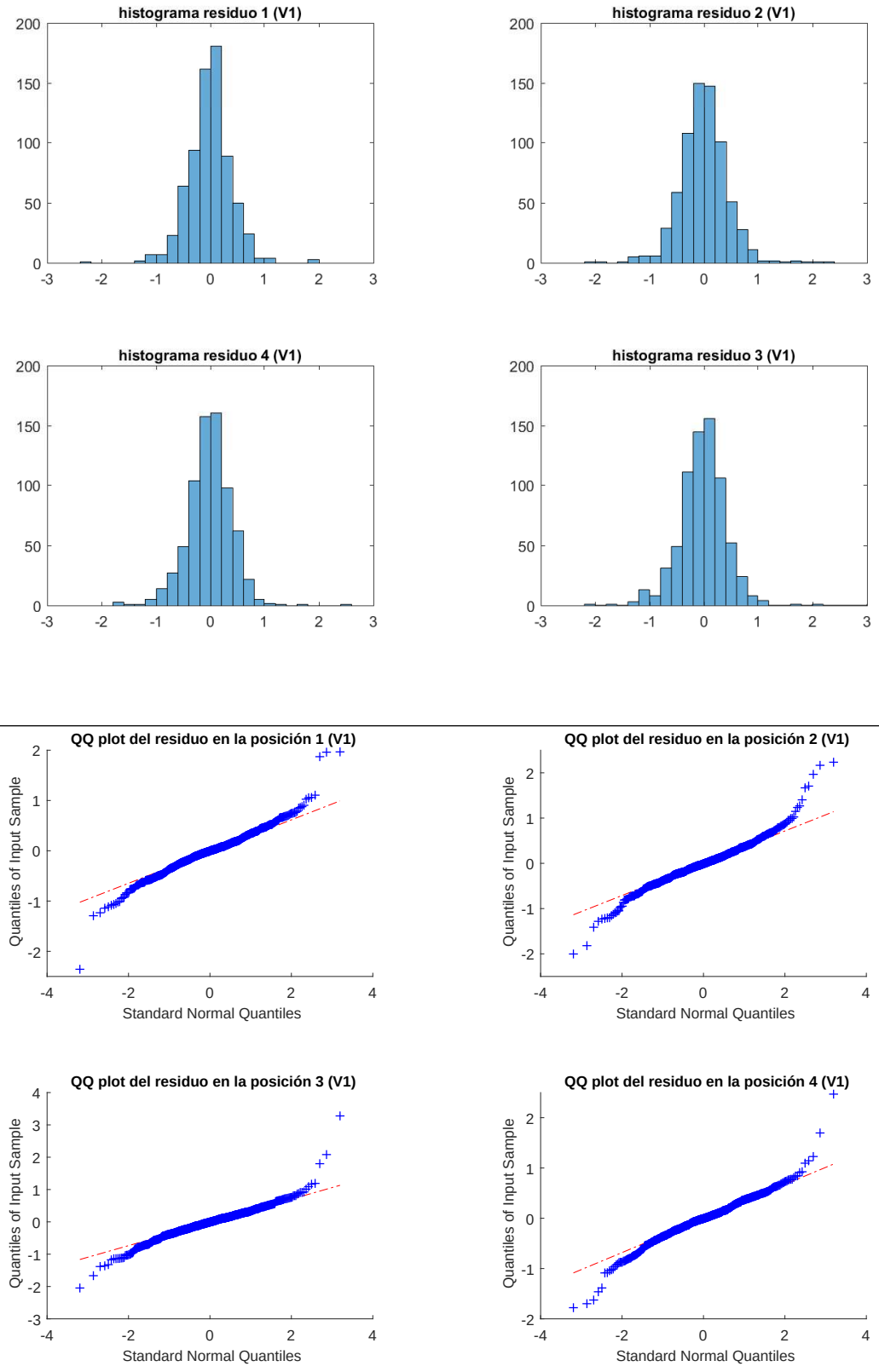


FIGURA 4.30: Histogramas y gráfico cuartil-cuartil para las series temporales del residuo en la vecindad ubicada en el mar, con posición geográfica 31°37′1,56″ y 31°38′40,92″ latitud sur; 72°57′10,44″ y 72°59′7,08″ longitud oeste.

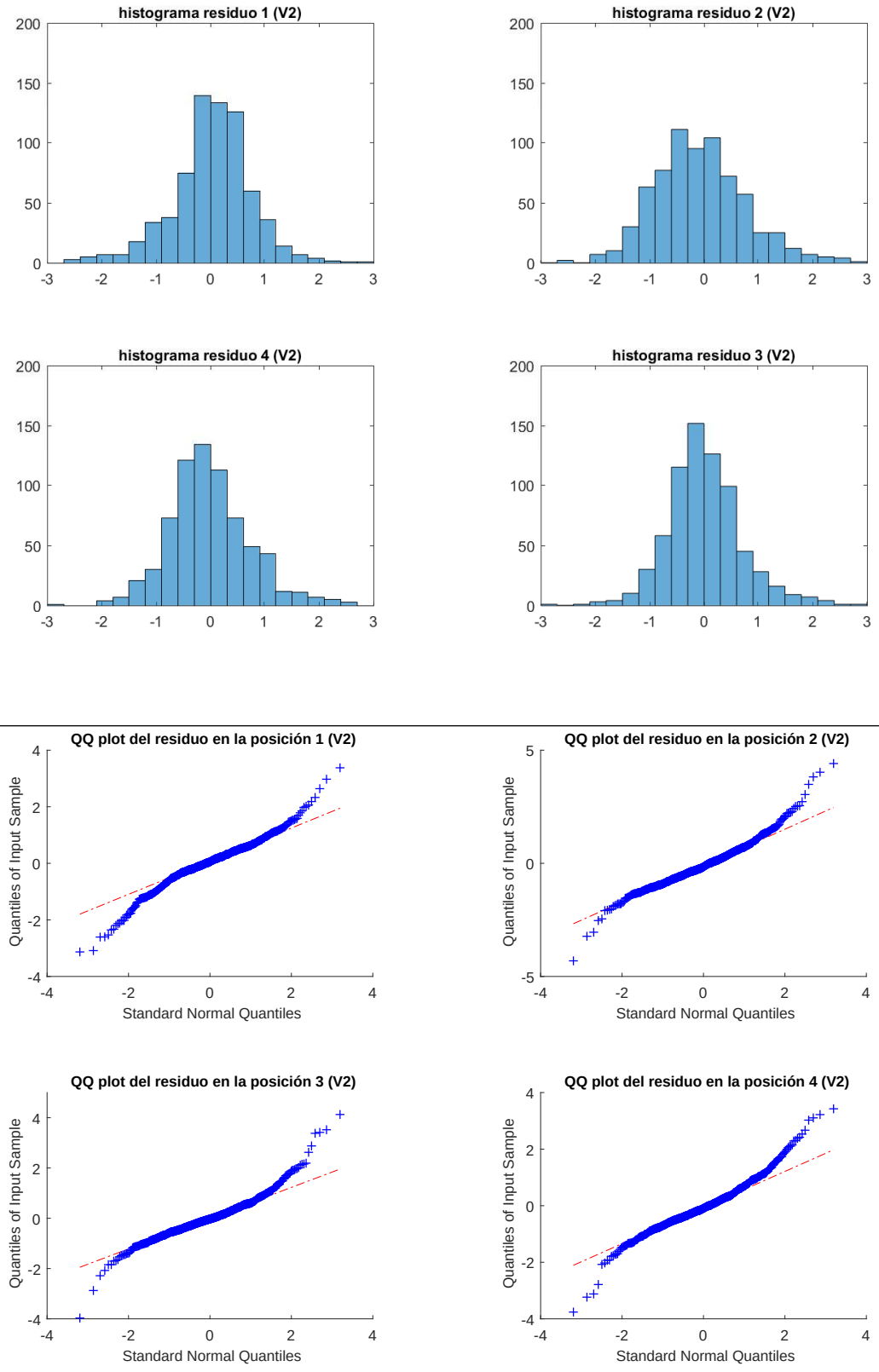


FIGURA 4.31: Histogramas y gráfico cuartil-cuartil para las series temporales del residuo en la vecindad ubicada en el valle, con posición geográfica 32°44′9,6″ y 32°45′47,88″ latitud sur; 72°8′46,32″ y 72°10′43,32″ longitud oeste.

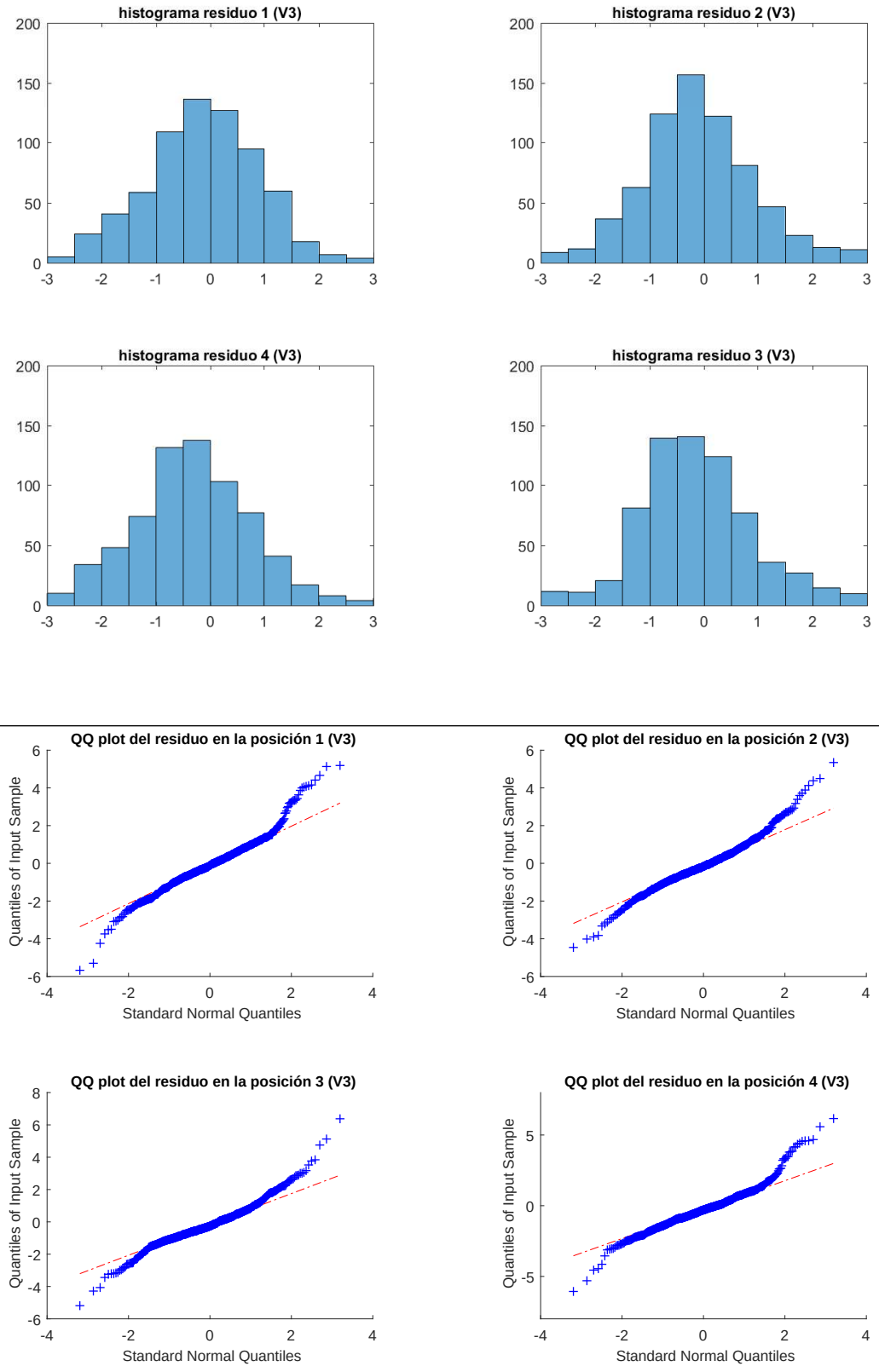


FIGURA 4.32: Histogramas y gráfico cuartil-cuartil para las series temporales del residuo en la vecindad ubicada en la montaña, con posición geográfica 32°16′52,68″ y 32°18′30,96″ latitud sur; 71°41′32,2″ y 71°43′27,12″ longitud oeste.

En líneas generales, podemos concluir que el error de estimación temporal no sigue una distribución normal, debido a que posee colas pesadas. A pesar de lo anterior, estas series temporales son estacionarias, siendo esta propiedad la que más nos interesa, debido a que es la única necesaria en el paso de clasificación usando el algoritmo del punto más lejano, dadas las hipótesis para usar la distancia telescópica como medida de similaridad entre las series de tiempo (ver Capítulo 2.3 y 3.2.3).

4.3.3. Downscaling: 0.5 [Km]

Procedemos a mostrar los resultados para la estimación del campo de intensidad de viento con resolución de 0.5[Km] en la zona de estudio, obtenido a través del campo de viento con resolución de 1[Km], y la estimación de los coeficientes del ajuste de regresión lineal entre los campos a 1.5[Km] y 3[Km].

4.3.3.1. Análisis espacial

Comenzamos analizando el efecto espacial y, al igual que en los análisis espaciales previos, presentamos los resultados para los días 1, 21 y 29 del mes de Septiembre del año 2014 en las Figuras 4.33, 4.34 y 4.35. Para cada día se presentan dos horas particulares, correspondientes a las 10:00 y 22:00 horas para el día 1/09/2014, 11:00 y 23:00 horas para los días 21/09/2014 y 29/09/2014, debido al cambio de horario que tuvo efecto el día 06/09/2014. Para cada día y hora se muestran el campo de viento a 1[Km], la estimación a 0.5[Km] y una ampliación de las imágenes en una zona de alto contraste, a modo de visualizar con más detalle lo que ocurre en estas zonas al realizar la estimación.

El campo de viento estimado a 0.5[Km] agrega detalles en las fronteras de las zonas de contraste, suavizando las transiciones entre una zona a otra, un ejemplo de este comportamiento se muestra en la Figura 4.34 en la ampliación para el tiempo $t = 502$. Además, en general, la estimación del campo de viento a 0.5[Km] presenta problemas al pasar desde una zona de alto contraste (colores amarillos) a zonas de más bajo contraste (verdes y azules), ocasionando que las fronteras entre ellas se vuelven difusas, pues tiende a mezclar pixeles, tal como se aprecia claramente en las ampliaciones presentes en las Figuras 4.33 y 4.35.

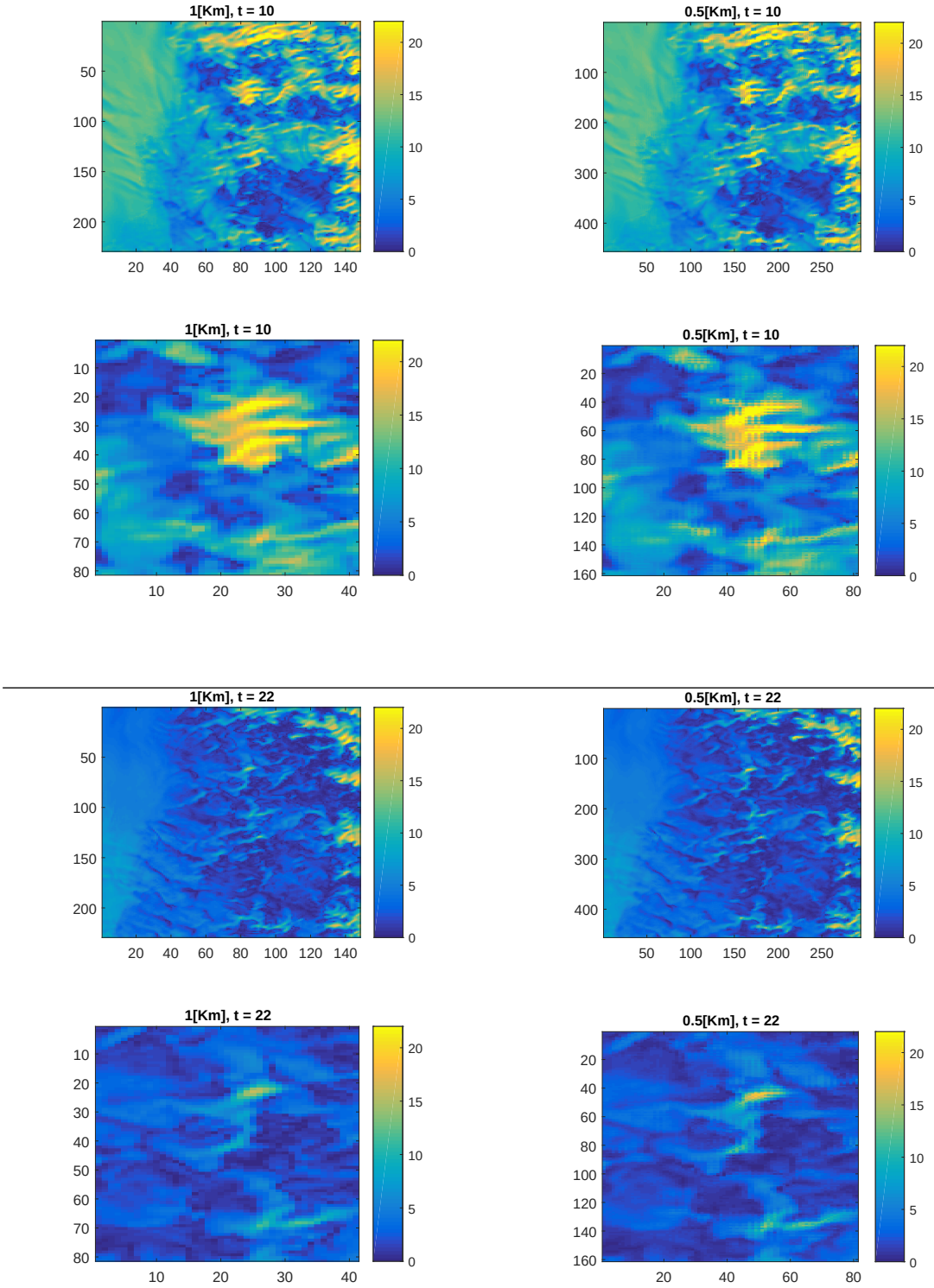


FIGURA 4.33: Campos de intensidad viento $X^{(1)}_{(\cdot,t_1)}$, $\hat{X}^{(0,5)}_{(\cdot,t_1)}$ y $X^{(1)}_{(\cdot,t_2)}$, $\hat{X}^{(0,5)}_{(\cdot,t_2)}$, donde $t_1 = 10$ y $t_2 = 22$ correspondientes al día 01 de Septiembre de 2014 a las 10:00 y 22:00 horas respectivamente. Además se muestran amplificaciones de los mismos campos entre las coordenadas 32°21′25,2″ y 33°4′37,2″ latitud sur; 70°52′24,24″ y 71°18′10,8″ longitud oeste.

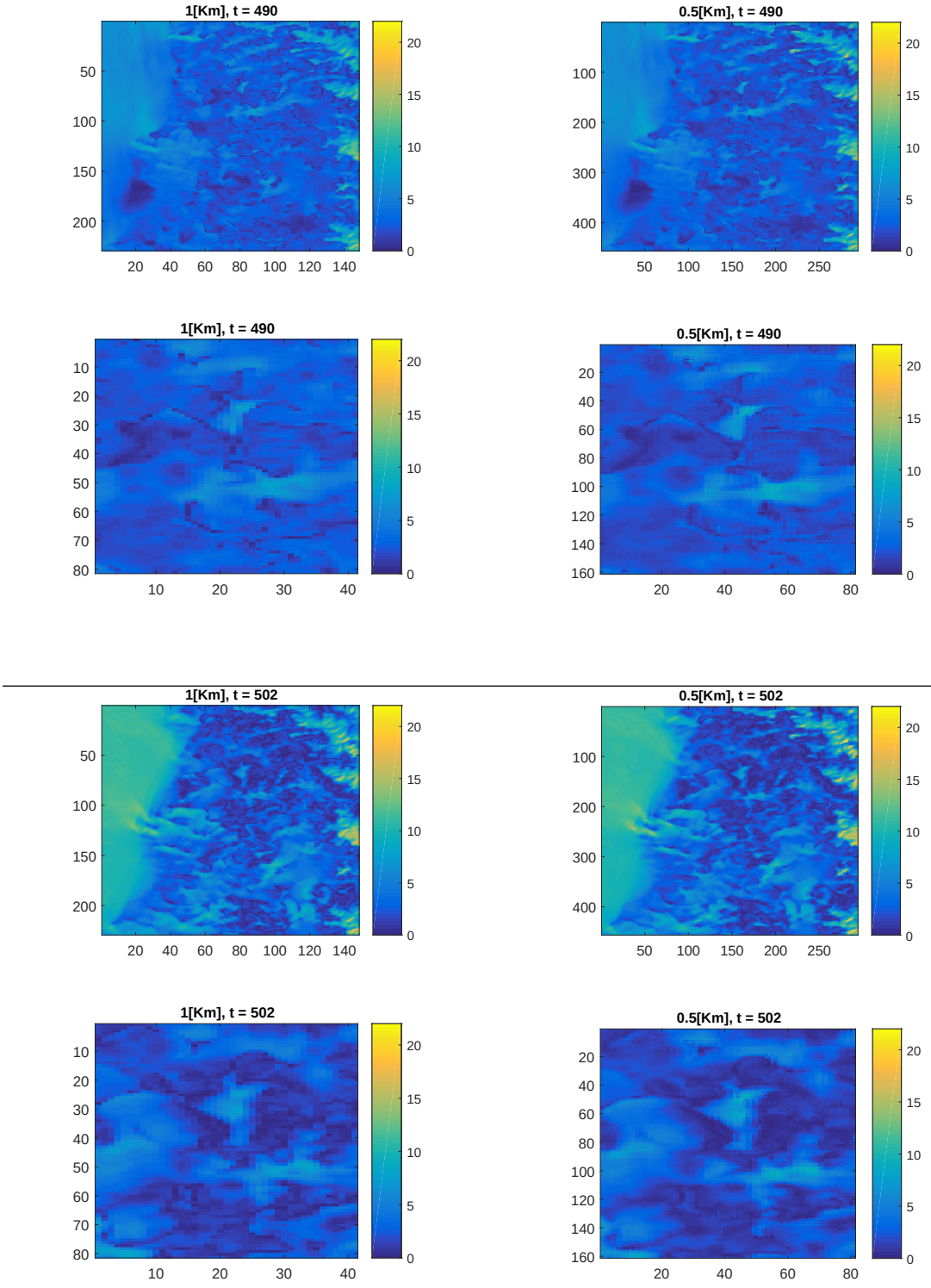


FIGURA 4.34: Campos de intensidad viento $X^{(1)}_{(\cdot,t_1)}$, $\hat{X}^{(0,5)}_{(\cdot,t_1)}$ y $X^{(1)}_{(\cdot,t_2)}$, $\hat{X}^{(0,5)}_{(\cdot,t_2)}$, donde $t_1 = 490$ y $t_2 = 502$ correspondientes al día 01 de Septiembre de 2014 a las 11:00 y 23:00 horas respectivamente. Además se muestran amplificaciones de los mismos campos entre las coordenadas 32°21′25,2″ y 33°4′37,2″ latitud sur; 70°52′24,24″ y 71°18′10,8″ longitud oeste.

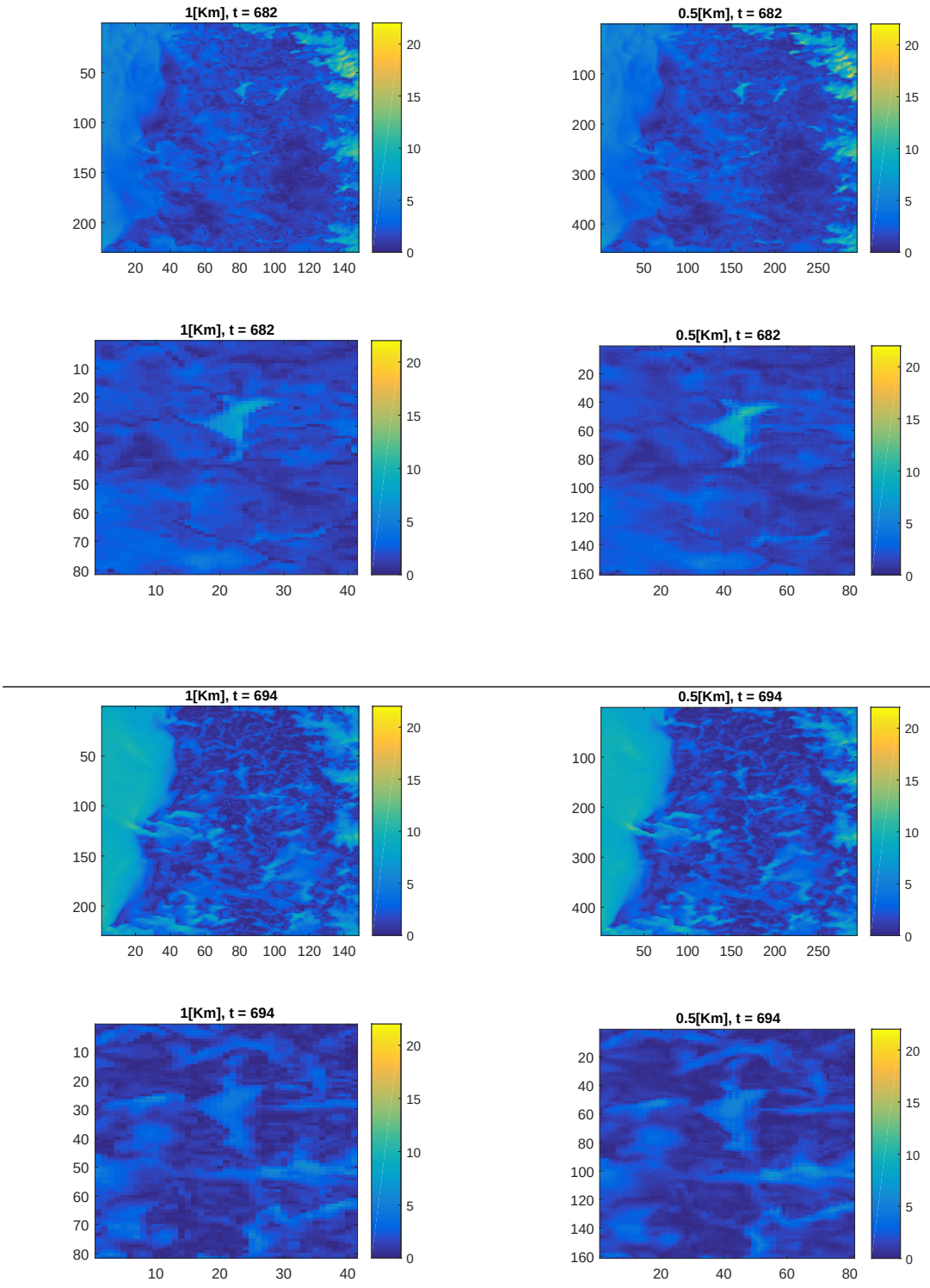


FIGURA 4.35: Campos de intensidad viento $X^{(1)}_{(\cdot,t_1)}$, $\hat{X}^{(0,5)}_{(\cdot,t_1)}$ y $X^{(1)}_{(\cdot,t_2)}$, $\hat{X}^{(0,5)}_{(\cdot,t_2)}$, donde $t_1 = 682$ y $t_2 = 694$ correspondientes al día 01 de Septiembre de 2014 a las 11:00 y 23:00 horas respectivamente. Además se muestran amplificaciones de los mismos campos entre las coordenadas 32°21′25,2″ y 33°4′37,2″ latitud sur; 70°52′24,24″ y 71°18′10,8″ longitud oeste.

Las zonas en donde se producen los problemas de estimación y mezcla de pixeles presentan heterogeneidad espacial de las series de tiempo de intensidad de viento. Una posibilidad de trabajar este problema es reajustar el modelo de regresión lineal en cada zona de homogeneidad espacial, las que definiremos a través de la clasificación en el espacio de las distribuciones de los residuos.

4.3.3.2. Análisis temporal

Continuamos mostrando los resultados temporales para la estimación del campo de viento a 0.5[Km], de una manera similar a la presentada en la sección 4.3.2.3. Debido a que en el interior de cada vecindad $V \subset M^{(3)}$ hay 16 puntos de $M^{(1)}$ y 36 puntos de $M^{(0,5)}$, con el fin de no ser redundante en la presentación de los resultados, elegimos solamente las cuatro series temporales de $\hat{X}^{(0,5)}$ tal que sus posiciones geográficas asociadas tienen un representante en $M^{(0,5)}$ y $M^{(1,5)}$ (ver Capitulo 3.3.1 la hipótesis de auto similaridad-espacial).

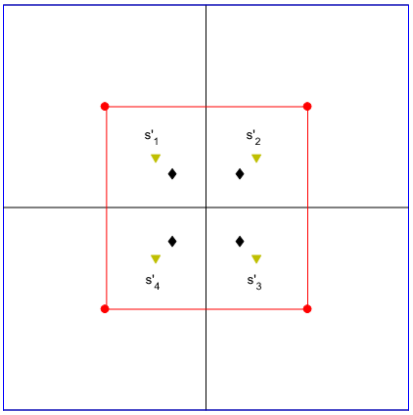


FIGURA 4.36: Ubicación de las posiciones pertenecientes a la vecindad $V' \subset M^{(3)}$, representados por los triángulos verdes y las posiciones pertenecientes a $W \subset M^{(1)}$, representados por los diamantes negros.

De esta forma, para cada una de las tres vecindades usadas para el análisis, ubicadas en el mar, valle y montaña, seguimos el orden presentado en la Figura 4.36, es decir en cada una de ellas generamos cuatro gráficos, donde cada uno contiene las series temporales de $\hat{X}^{(0,5)}$ y $X^{(1,5)}$ (obtenida mediante la interpolación bicúbica) asociadas a la posición $s'_i, i = \{1, \ldots, 4\}$, representadas por los triángulos verdes, las cuales pertenecen a $M^{(0,5)}$ y $M^{(1,5)}$; y la serie temporal a 1[Km] cuya posición es la más cercana a s_i, representadas por los diamantes negros.

En la Figura 4.37 mostramos los gráficos para la vecindad ubicada en el mar. Debido a que los campos de viento en estas zonas son homogéneos, las tres series temporales se comportan de manera similar, de hecho solo es posible distinguir una de otra en los extremos de las series, debido al suavizamiento que ocurre a causa del ajuste de regresión lineal.

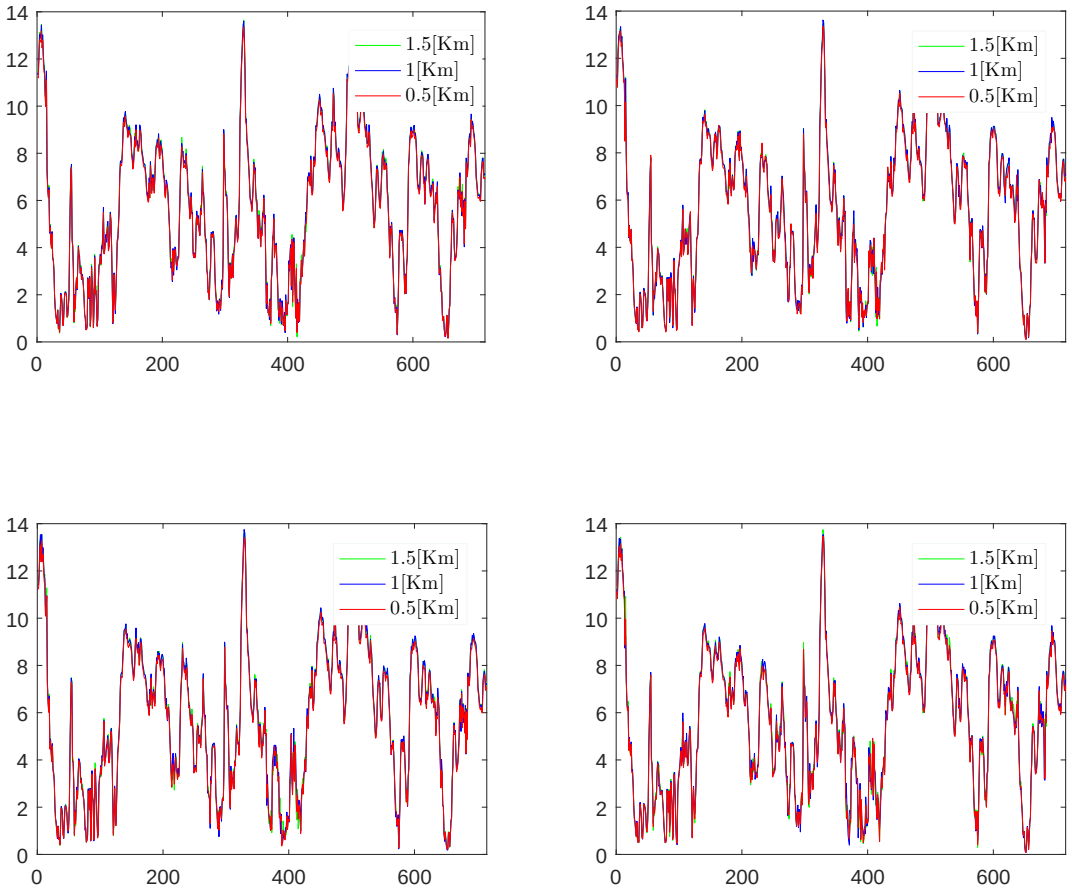


FIGURA 4.37: Series temporales dentro de la vecindad $V \subset M^{(3)}$ ubicada en el mar con coordenadas geográficas 31°37′1,56″ y 31°38′40,92″ latitud sur; 72°57′10,44″ y 72°59′7,08″ longitud oeste. Cada gráfico muestra: Las series temporales a 1.5[Km] y 0.5[Km] asociadas a la posición $s'_i \in V'$ correspondiente, además de la serie temporal a 1[Km] cuya posición geográfica está más cerca a s'_i, para $i = \{1, \ldots, 4\}$.

En la Figura 4.38 observamos los gráficos para la vecindad ubicada en el valle. Notamos que en esta vecindad las series temporales son más distinguibles entre sí que en la vecindad del mar, aun así, sus comportamientos son parecidos, debido a que esta zona

también es bastante homogénea en cuanto a la intensidad del viento. Nuevamente las mayores diferencias entre las series de tiempo ocurren en sus puntos extremos.

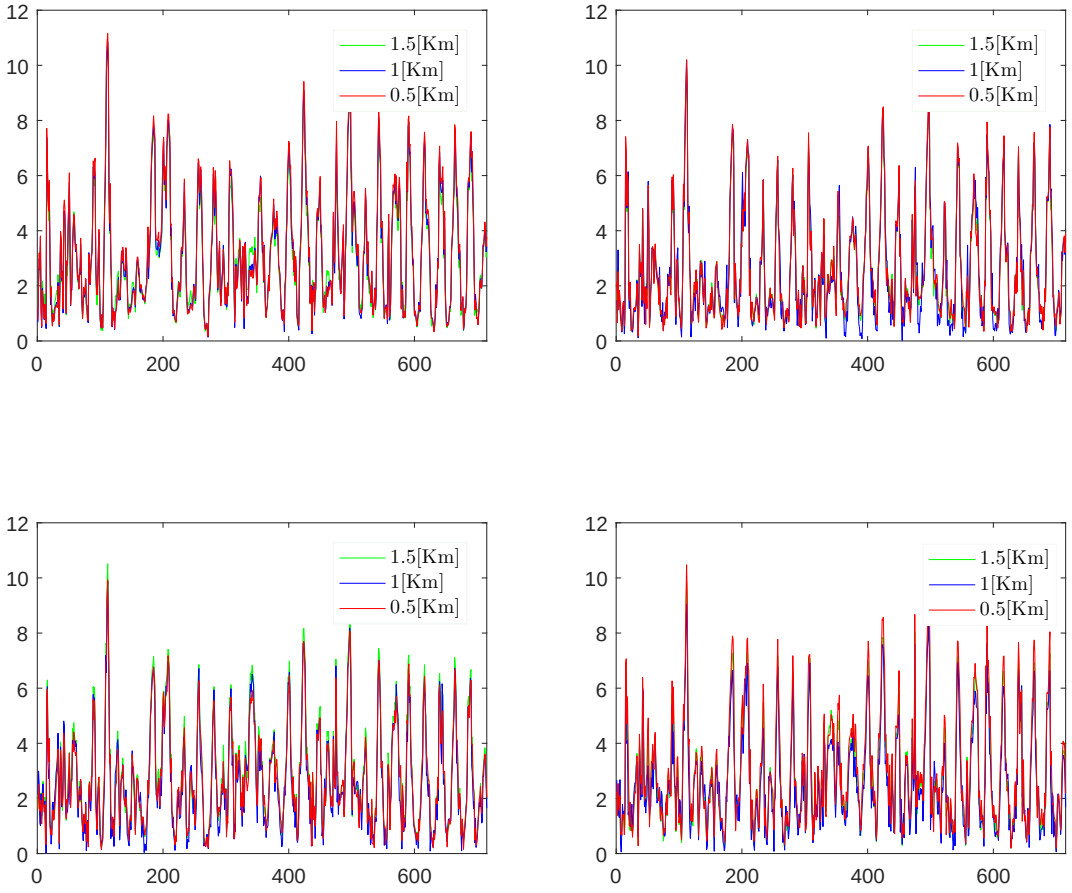


FIGURA 4.38: Series temporales dentro de la vecindad $V \subset M^{(3)}$ ubicada en el valle con coordenadas geográficas 32°44′9,6″ y 32°45′47,88″ latitud sur; 72°8′46,32″ y 72°10′43,32″ longitud oeste. Cada gráfico muestra: Las series temporales a 1.5[Km] y 0.5[Km] asociadas a la posición $s'_i \in V'$ correspondiente, además de la serie temporal a 1[Km] cuya posición geográfica está más cerca a s'_i, para $i = \{1, \ldots, 4\}$.

En la Figura 4.39 muestran los gráficos para la vecindad ubicada en la montaña. Se aprecia que la serie temporal estimada $\hat{X}^{(0,5)}$ capta las fluctuaciones, cambios abruptos en el comportamiento de las series de tiempo a escala mayor.

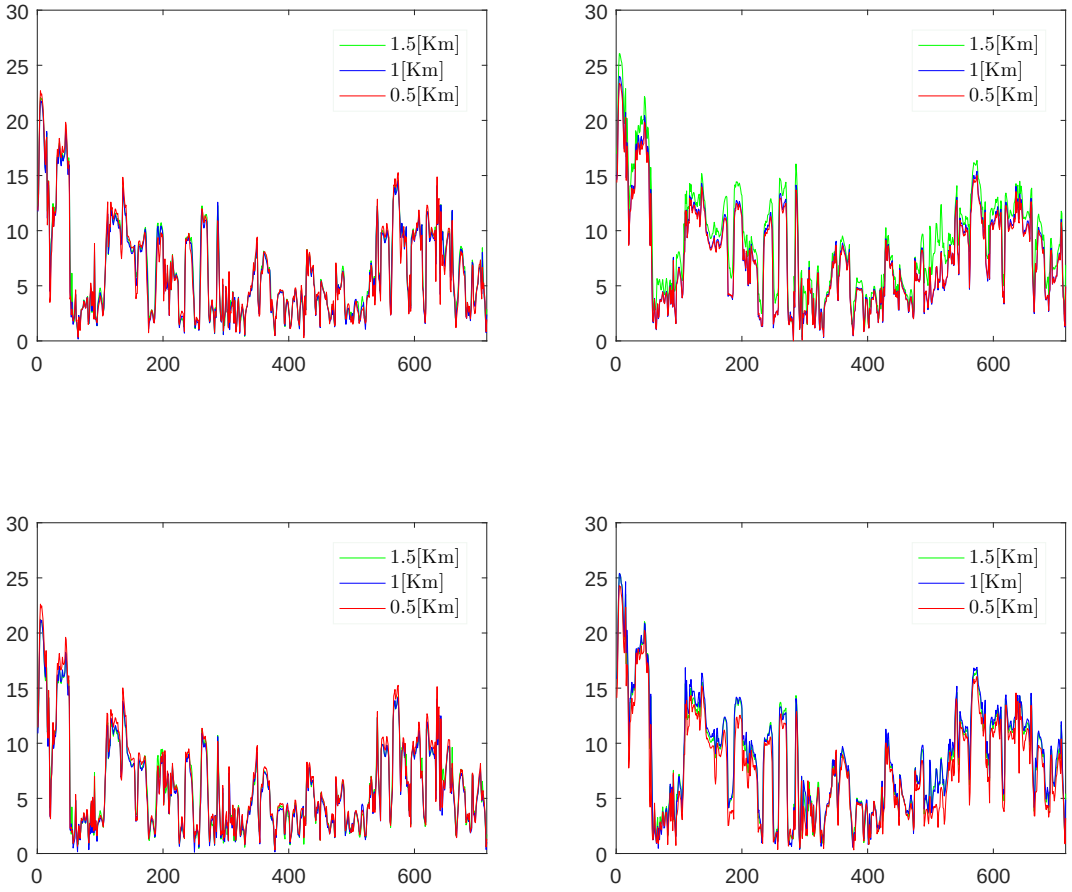


FIGURA 4.39: Series temporales dentro de la vecindad $V \subset M^{(3)}$ ubicada en la montaña con coordenadas geográficas 32°16′52,68″ y 32°18′30,96″ latitud sur; 71°41′31,2″ y 71°43′27,12″ longitud oeste. Cada gráfico muestra: Las series temporales a 1.5[Km] y 0.5[Km] asociadas a la posición $s'_i \in V'$ correspondiente, además de la serie temporal a 1[Km] cuya posición geográfica está más cerca a s'_i, para $i = \{1, \ldots, 4\}$.

En general podemos concluir que el ajuste de regresión lineal resulta un buen método de estimación en zonas de homogeneidad espacial. El estimador a 0.5[Km] capta de buena manera el comportamiento de las fluctuaciones de las series de tiempo. Sin embargo hay zonas en las cuales el ajuste de regresión lineal realizado no es capaz de captar de manera adecuada el comportamiento temporal, por ejemplo si consideramos la vecindad $V \subset M^{(3)}$ con coordenadas geográficas 32°33′5,4″ y 32°34′43,68″ latitud sur; 71°39′46,08″ y 71°41′42,36″ notamos, a través de la información del coeficiente de determinación y error cuadrático medio contenido en la tabla 4.8, que el ajuste de

regresión lineal en esta vecindad no es capaz de captar la variabilidad presente en los datos, por lo que no es un buen modelo para ellos.

Vecindad XX	s'_1	s'_2	s'_3	s'_4
ECM	1.5088	1.0733	0.9744	1.0264
R^2	0.1869	0.0388	0.1889	0.4715

TABLA 4.8: Error cuadrático medio y R^2 de la regresión lineal para las posiciones $s'_i, i = \{1, \ldots, 4\}$ de la vecindad $V' \subset M^{(1,5)}$ con coordenadas geográficas 32°33′5,4″ y 32°34′43,68″ latitud sur; 71°39′46,08″ y 71°41′42,36″.

Más aún, en la Figura 4.40, la cual presenta los diagramas de puntos entre los elementos de $V \subset M^{(3)}$ y $V' \subset M^{(1,5)}$ apreciamos que solo en algunos existe una leve componente de tendencia, incluso algunos parecen tener un patrón de puntos aleatorios.

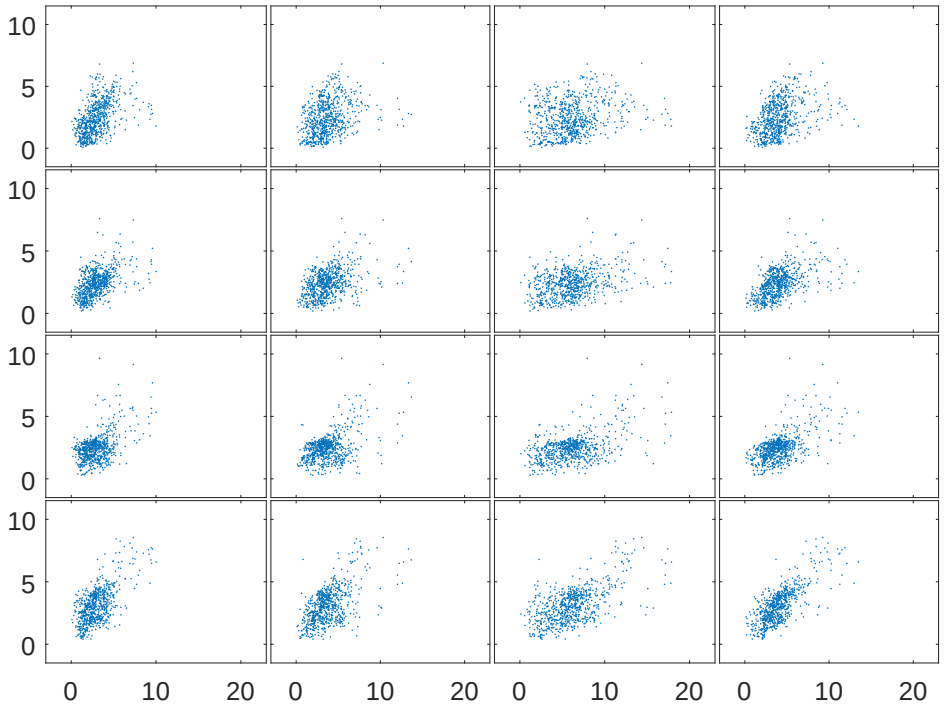


FIGURA 4.40: Gráficos de puntos de las series temporales de la intensidad del viento correspondientes a la vecindad $V \subset M^{(3)}$ contra los elementos de la vecindad $V' \subset M^{(1,5)}$ al interior de V. Estas vecindades poseen coordenadas geográficas 32°33′5,4″ y 32°34′43,68″ latitud sur; 71°39′46,08″ y 71°41′42,36″.

A pesar de lo anterior, al analizar los resultados para el downscaling en esta vecindad V, presentes en la Figura 4.41, observamos resultados bastantes positivos, donde las

series temporales a 0.5[Km] son capaces de captar las fluctuaciones provenientes de escalas de menor resolución, pero presentan una disminución en su intensidad.

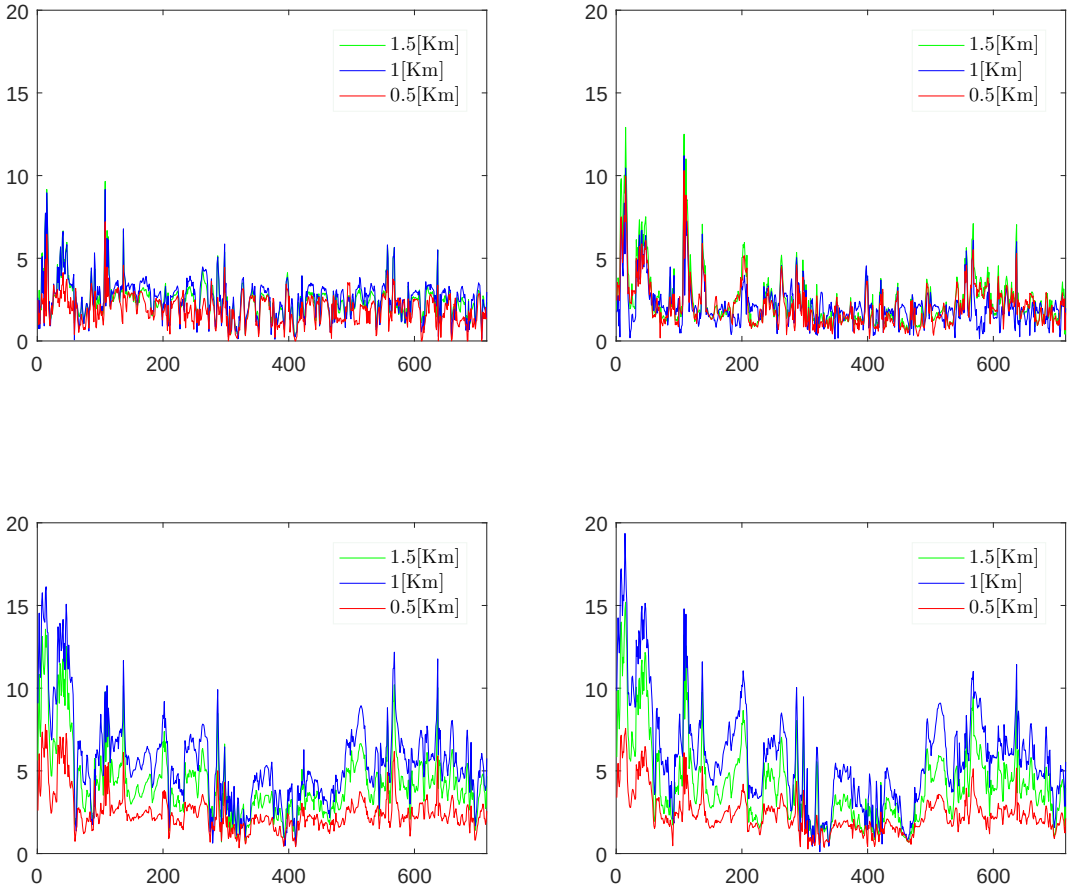


FIGURA 4.41: Series temporales dentro de la vecindad $V \subset M^{(3)}$ coordenadas geográficas 32°33′5,4″ y 32°34′43,68″ latitud sur; 71°39′46,08″ y 71°41′42,36″. Cada gráfico muestra: Las series temporales a 1.5[Km] y 0.5[Km] asociadas a la posición $s'_i \in V'$ correspondiente, además de la serie temporal a 1[Km] cuya posición geográfica está más cerca a s'_i, para $i = \{1, \ldots, 4\}$.

Por lo tanto, el esquema de downscaling planteado entrega buenos resultados, siendo capaz de captar en la mayoría de los casos el comportamiento medio y fluctuaciones a nivel temporal, además de añadir detalles y suavizando las transiciones entre contrastes a nivel espacial. Se concluye que, en las zonas en donde el paso de ajuste mediante regresión lineal no es significativo, el modelo sigue siendo capaz de captar las fluctuaciones temporales y agregar información espacial.

4.4. Clasificación

En esta sección discutiremos los resultados obtenidos mediante la clasificación de las series temporales para el residuo obtenido desde el ajuste del modelo de regresión lineal entre los campos de viento $X^{(1,5)}$ y $X^{(3)}$ realizado en la sección 4.3.

Comenzamos recordando las ideas principales del algoritmo de clasifación mencionadas en las secciones 2.3 y 3.2.2. Nuestro objetivo es, dadas las N series temporales de residuos agruparlas en m clases disjuntas, en donde dos series temporales pertenecen a una misma clase si y solo si fueron generadas por funciones de distribución "similares". Para esto se necesitan dos ingredientes: una función de distancia que indique la similaridad entre las funciones de distribución de las series temporales, y un algoritmo que tome esta información para definir a que clase pertenece cada serie temporal. En nuestro caso usamos el algoritmo del punto más lejano en conjunto con la distancia telescópica [29].

Recordemos que el algoritmo del punto más lejano consta de dos pasos. El primero busca a los representantes de cada clase, escogiendo las m series temporales provenientes de las distribuciones más diferentes, y luego el resto de las series de tiempo son asignados a la clase cuyo representante este más cercano. La distancia telescópica es una medida de similaridad que nos permite comparar las distribuciones de procesos estocásticos estacionarios mediante las distribuciones finito dimensionales empíricas de sus muestras.

En la practica el número m de clases es desconocido, para un primer análisis supondremos que $m \leq 12$, por lo que realizamos la clasificación fijando $m = \{2, \ldots, 12\}$. En la Figura 4.42 mostramos el gráfico de la menor distancia entre los representantes de las clases, para cada valor de m, observando que, dada la estructura del algoritmo del punto más lejano, la relación entre el número de clases y las distancias es decreciente.

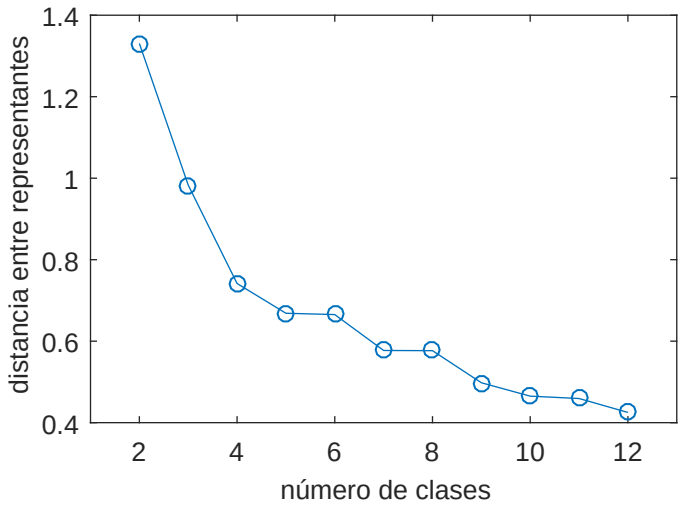


FIGURA 4.42: Gráfico entre el la menor distancia entre las clases versus el número de clases.

Para estudiar los valores entregados por la distancia telescópica, consideramos un ejemplo de series de tiempo independientes con distribuciones gaussianas de media cero y varianzas diferentes. Generamos 100 series temporales provenientes de distribuciones normales N(0,1) y N(0,1.25), la distancia telescópica promedio obtenida es de 0,4 con una desviación estándar de 0,02, mientras que al calcular la divergencia de Kullback entre estas dos distribuciones obtenemos valores de 0,05 y 0,04, por lo que la distancia telescópica discrimina más la diferencia entre estas distribuciones. La selección del número de clases utilizando la distancia telescópica es un problema abierto; en nuestro caso, no es claro definir un buen criterio de selección mediante el análisis de la Figura 4.42, sin embargo, observando el decaimiento de la distancia entre los centros de las clases, el primer quiebre importante se produce para $m = 4$, por lo que debemos considerar al menos este número de clases.

En la Figura 4.43 mostramos como se distribuyen las clases de manera espacial para diferentes números de clases que van desde $m = 2$ hasta $m = 7$, donde se aprecia que color se puede asociar a diferentes atributos geográficos. Para $m = 2$ la zona amarilla corresponde a sectores cordilleranos, principalmente cordillera de los andes y una pequeña zona de la cordillera de la costa en donde se encuentran las cúspides más altas, mientras que la zona azul corresponde al resto de la zona, sin discriminar mayor detalle. Para $m = 3$ la zona azul se asocia al mar, valles transversales y zonas urbanas, mientras que el amarillo corresponde a zonas mayoritariamente montañosas y la celeste corresponde a las zonas en donde la cordillera de los andes alcanza sus cúspides más altas. Para $m = 4$ la clase azul y celeste mantienen su interpretación,

mientras el amarillo se transforma en una zona de transición entre la clase azul y la verde, que incluye la costa norte de la región de Valparaíso, mientras que la clase verde está asociada a cordones cordilleranos incluyendo la cordillera de la costa y la cordillera de los andes.

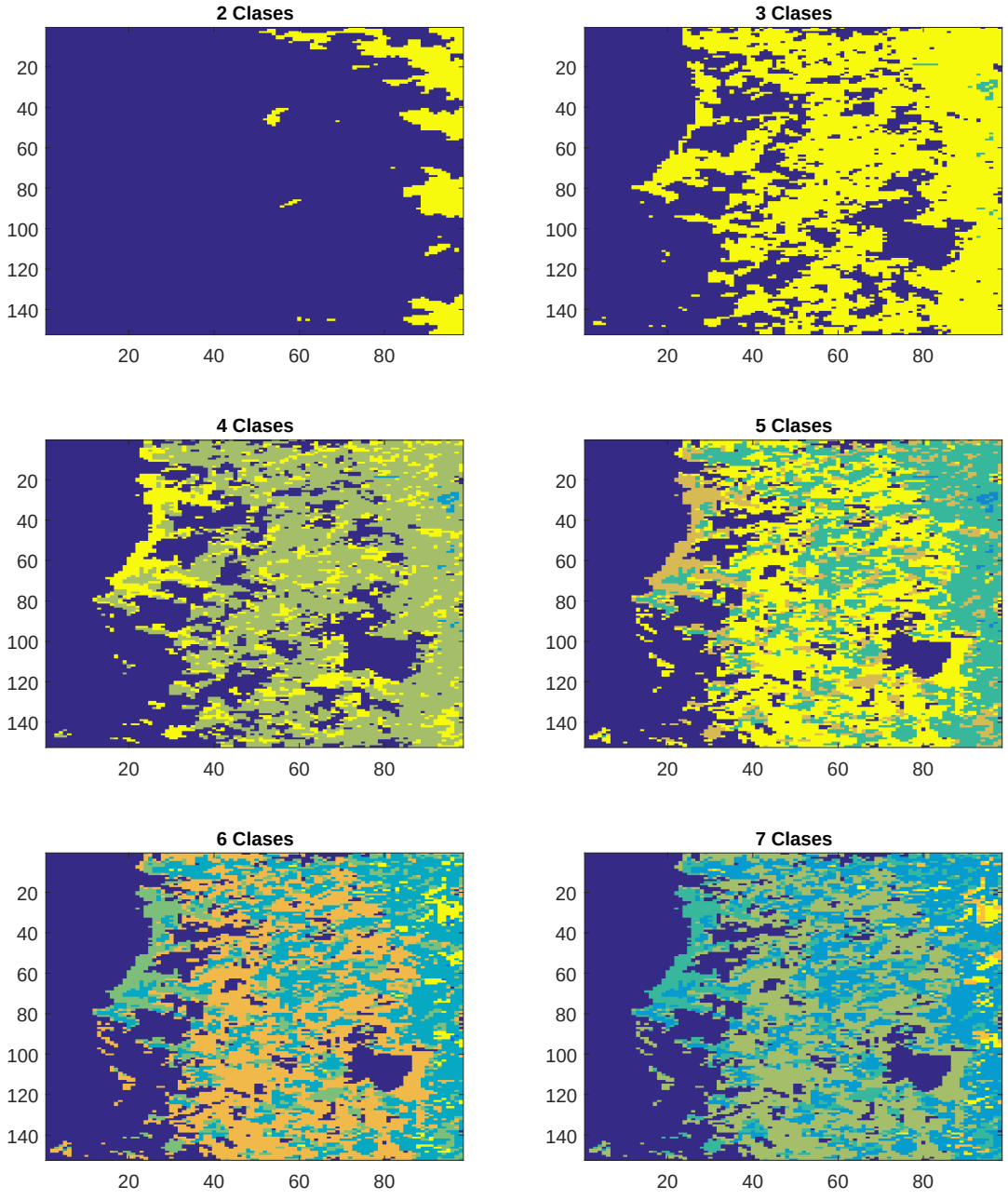


FIGURA 4.43: Distribución espacial de las clases obtenidas para $m = 2, 3, \ldots, 7$.

Para $m = 5$ aparece la zona café, que explicar de la transición en la costa norte de la región, aquí se empiezan a segmentar la clase azul, perdiendo parte de la interpretabilidad geográfica. Para $m \geq 6$ las clases descritas anterior empiezan a segregarse hasta el punto de dividir clases con poca cantidad de elementos (como la clase celeste, la cual es casi inexistente para 6 y 7 clases), por lo que algunas clases quedan subrepresentadas.

En la Figura 4.44 mostramos las distribuciones estimadas mediante un método no paramétrico usando funciones de kernel (ver Capítulo 2.4) para las siete clases mostradas en la Figura 4.43. Observamos que, salvo la segunda clase (la cual en la Figura 4.43 está representada por la zona celeste), la mayor diferencia entre estas densidades se presenta en la variabilidad, pues estas distribuciones están centradas en cero, además se puede observar que la sexta y séptima clase son similares. Tomando en cuentas estas observaciones una elección conveniente para el número m de clases esta entre cuatro y cinco.

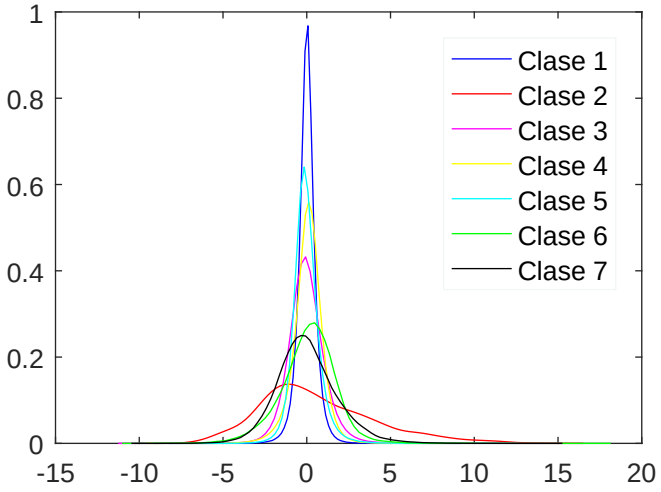
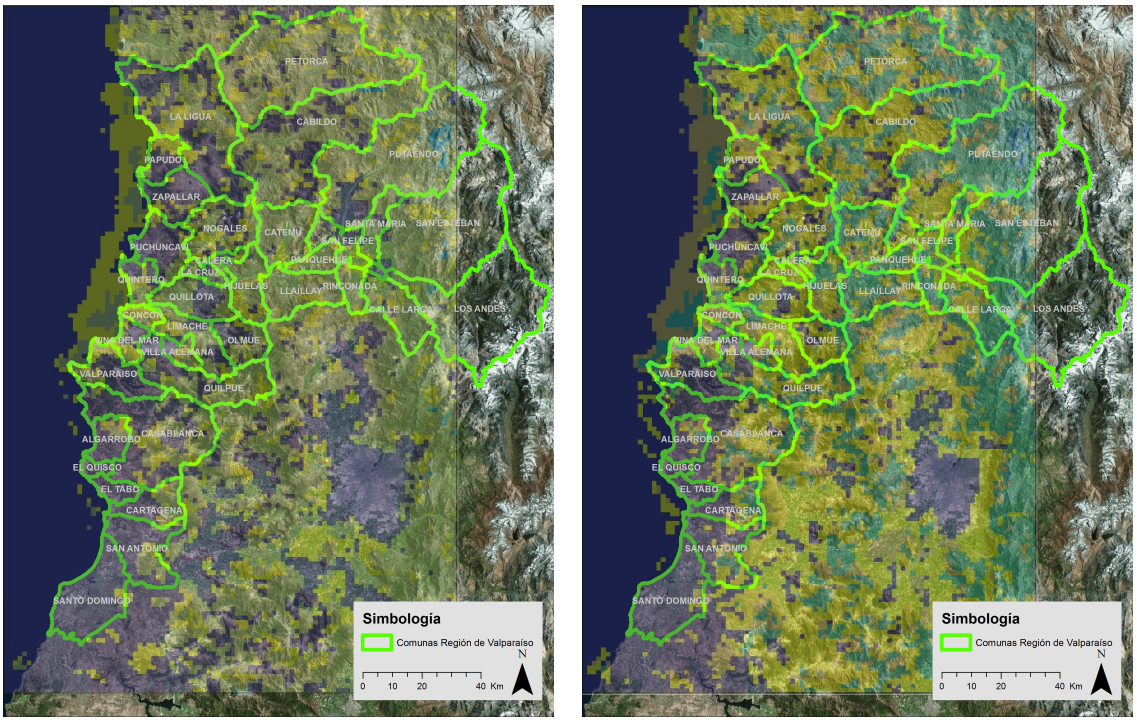


FIGURA 4.44: Gráfico de las distribuciones estimadas de manera no paramétrica por funciones de kernel para 7 clases.

Para discernir entre $m = 4$ y $m = 5$, visualizamos Figura 4.45 en la como se distribuyen las clases sobre la topografía de la región de estudio, observando al pasar de cuatro a cinco clases se pierden al interior de la región zonas azules, las cuales están asociadas a valles y zonas urbanas.

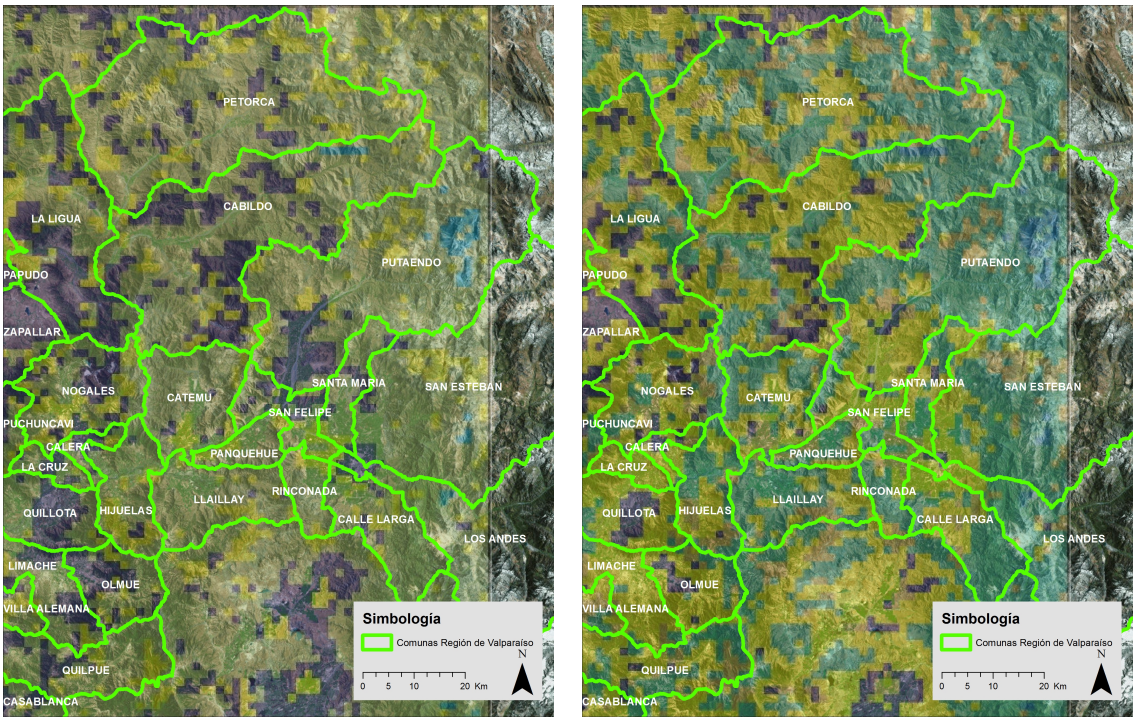


(a)4 Clases

(a)5 Clases

FIGURA 4.45: Distribución geografía de las clases para (a)$m = 4$ y (b)$m = 5$.

En la Figura 4.46 se muestra una ampliación de la Figura 4.45. Apreciamos que al considerar cinco clases hay zonas urbanas, por ejemplo, las ciudades de Zapallar, San Felipe y Quillota, dejan de estar en la clase asociada a zonas urbanas y pasan a ser parte de una nueva clase de transición, dificultando la interpretación geográfica. Por estas razones elijemos trabajar con cuatro clases ($m = 4$), pues facilita la interpretación geográfica de cada una de las ellas.



(a)4 Clases (a)5 Clases

FIGURA 4.46: Distribución geografía de las clases al interior de la región de Valparaíso para (a)$m = 4$ y (b)$m = 5$.

4.4.1. Regresión Lineal por clases.

La regresión lineal por clases consiste en estimar localmente solo series temporales pertenecientes a una misma clase. Por ejemplo en la Figura 4.47 para la clase $C = 1$, s'_1, s'_2 y s_1, s_2 pertenecen a esta clase, así que para el ajuste de regresión solo se consideran las series temporales asociadas a dichas posiciones. De la misma manera, para la clase $C = 2$ solo se consideran las series temporales asociadas a las posiciones s'_3, s'_4, s_3 y s_4. Este reajuste permite considerar en la estimación local solo las series temporales que poseen distribuciones similares.

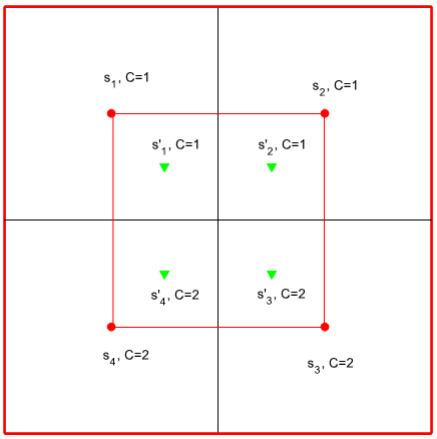


FIGURA 4.47: Ilustración de una posible asignación de clases para los elementos pertenecientes a una vecindad $V = \{s_1, s_2, s_3, s_4\}$ y $V' = \{s'_1, s'_2, s'_3, s'_4\}$. En general, para la clase j se considera para la regresión las series temporales $V \cap \{C = j\}$.

Para construir el campo de intensidad de viento a 0.5[Km] se utiliza el mismo esquema. Los resultados obtenidos mediante este reajuste no difieren lo suficiente con respecto a los mostrados en la sección 4.3. De hecho, la máxima diferencia en valor absoluto, punto a punto, entre ambos campos estimados mediante el paso de downscaling es de orden 10^{-11}. Con esto podemos concluir que el esquema de regresión lineal local es capaz de discriminar los puntos con diferentes características a la hora de explicar el campo de intensidad de viento 1.5[Km] desde el campo de viento a 3[Km] y el reajuste considerando la clasificación no es necesario.

4.4.2. Bandas de confianza.

A continuación, procedemos a construir bandas de confianza para las series temporales del residuo mediante la información obtenida de la clasificación. Con ello generamos bandas de confianza para las series temporales de intensidad de viento estimadas a 0.5[Km].

Para ello construimos intervalos de confianza en todos los instantes de tiempo para cada clase, es decir, para todas las series temporales del residuo perteneciente a una misma clase, consideramos sus valores para cada uno de los tiempos, por ejemplo, en la Figura 4.48 mostramos un histograma y la densidad estimada para todos los valores de las series temporales del residuo pertenecientes a la clase cuatro, en el tiempo $t = 200$.

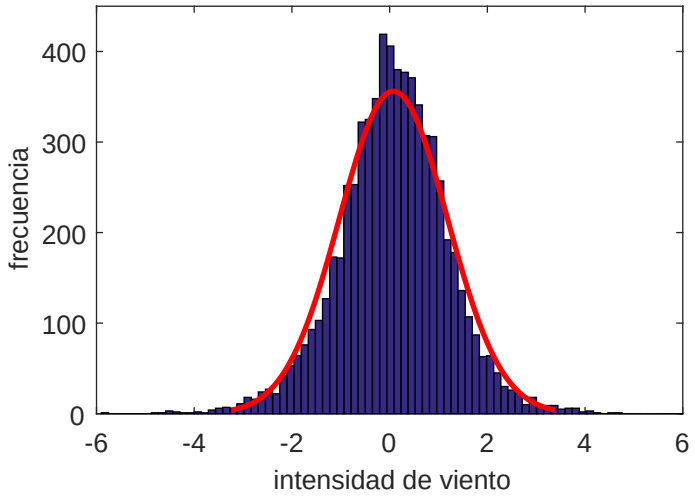


FIGURA 4.48: Histograma de valores de las series temporales del residuo para $t = 200$ pertenecientes a la clase 4.

Luego ordenamos estos valores de menor a mayor, con el fin de calcular los percentiles. De esta manera, obtenemos un intervalo de confianza al $(1-\alpha)\,\%$ mediante los datos correspondientes a los percentiles $\alpha/2$ y $1-\alpha/2$, denotados por $\theta_{\alpha/2}$ y $\theta_{1-\alpha/2}$ respectivamente (ver Figura 4.49), el intervalo de confianza está dado por $\left[\theta_{\alpha/2}, \theta_{1-\alpha/2}\right]$. Esta forma de construir intervalos de confianza es similar al denominado *boostrap pivotal confidence*, el cual puede ser revisada en mayor detalle en [38].

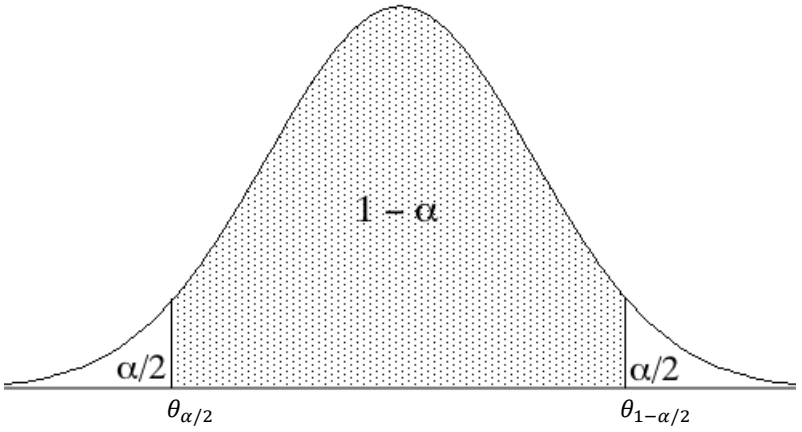


FIGURA 4.49: Ilustración de un intervalo de confianza al $(1-\alpha)\,\%$, representado por la zona punteada.

Obtenemos los siguientes resultados para intervalos de confianza al 95 %:

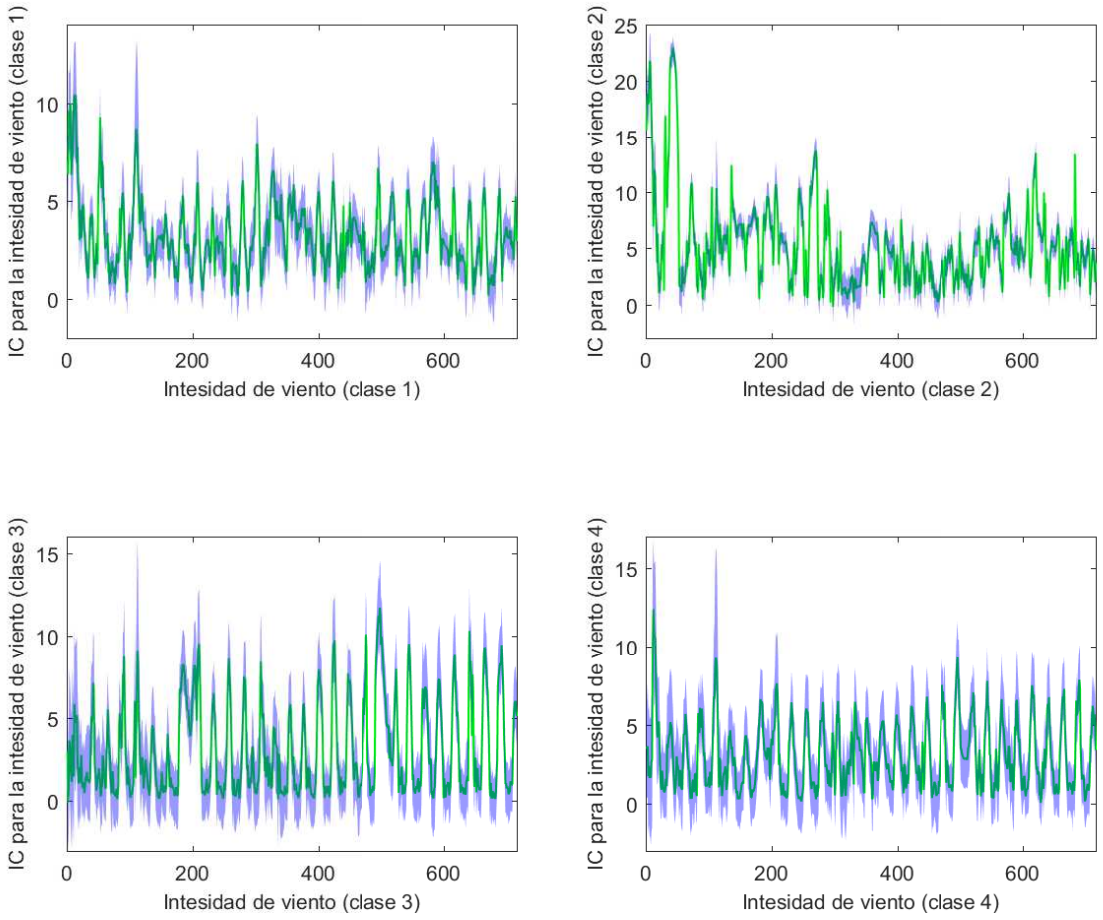


FIGURA 4.50: Bandas de confianza al 95 % para las series temporales de intensidad de viento pertenecientes a cada una de las cuatro clases obtenidas mediante la clasificación.

Notamos que las bandas de confianza al 95 % son pequeñas con respecto al rango de variabilidad de los datos, salvo en los puntos donde se presentan cambios bruscos en el comportamiento de la intensidad del viento, aumentando la variabilidad para la intensidad de viento y provocando mayor incertidumbre en la estimación de sus valores.

La construcción de las bandas de confianza nos da información respecto a los diferentes escenarios que pueden ocurrir al estimar la intensidad del viento en un rango de tiempo, además de entregar casos pesimistas y optimistas, por ejemplo, en el caso

de un incendio forestal, un caso pesimista es que la intensidad del viento sea alta. Además, las bandas de confianza son de utilidad en modelos que estiman el potencial eólico de una zona de interés, pues estos se enfocan en prever la menor y mayor producción posible en un cierto horizonte de tiempo.

En conclusión, el uso de técnicas de clasificación para las distribuciones temporales del residuo mediante la distancia telescópica y el algoritmo del punto más lejano nos permiten en este estudio identificar diferentes comportamientos estadísticos con características geográficas de la región, a pesar de que teóricamente no exista un criterio definido para determinar el número de clases. Además, nos permite construir bandas de confianza para las series temporales de intensidad de viento estimadas a 0.5[Km], permitiendo visualizar la variabilidad que no es captada por el ajuste de regresión lineal.

Capítulo 5

Conclusión

En este trabajo se propuso un esquema de downscaling en tres pasos, sustentándose en las hipótesis de auto-similaridad espacial, regularidad local espacial y homogeneidad espacial, las cuales influyen de diferentes maneras en la metodología planteada. La auto-similaridad induce una estructura de dependencia entre campos de intensidad de viento con resoluciones diferentes, por lo que la elección de la técnica a usar en el paso de upscaling debe ser capaz de captar la relación existente entre los campos de viento a 1[Km] y 1.5[Km]. En este sentido el interpolador bicúbico es una opción viable para el paso de upscaling, pues mantiene las fronteras y no sobresuaviza las imágenes de los campos de intensidad de viento, preservando el rango de intensidad, además logra captar el comportamiento de las series temporales de manera local, manteniendo la información proveniente del campo de viento a 1[Km].

Con lo anterior, se debe platear un ajuste entre los campos de viento a 3[Km] y 1.5[Km] que preserve la estructura de dependencia, pues la hipótesis de auto-similaridad nos indica que la relación existente entre estos campos es la misma que entre los campos de viento a 1[Km] y 0.5[Km]. La regularidad espacial local permite utilizar el método de regresión lineal para estimar la relación entre los campos de viento a 3[Km] y 1.5[Km] observando que, a nivel espacial, es captas de captar la componente de tendencia y conservar el rango de intensidad de viento para el campo a 1.5[Km], produciendo errores de estimación centrados en cero. El residuo presenta una estructura de dependencia espacial relacionada con la regularidad de la topografía, pues en zonas marítimas y de valles se presenta poca dispersión, mientras que en zonas montañosas

los valores de la intensidad del viento presentan mayor variabilidad. Además, las distribuciones espaciales del residuo presentan colas pesadas, concordando con la gran cantidad de valores extremos o atípicos observados, sobre todo en las zonas de mayor irregularidad geográfica identificas a través del análisis de clasificación.

En cuanto a la influencia de la auto-similaridad en la variabilidad temporal el ajuste de regresión capta la variabilidad presente en los datos, pues un 85 % de las estimaciones realizadas poseen coeficiente de determinación mayores a 0,7. Sin embargo, en zonas de gran heterogeneidad topográfica, este modelo no es adecuado, pues el coeficiente de determinación presenta valores entre 0,09 y 0,04, por lo que el ajuste de regresión lineal no capta de buena manera la variabilidad de la intensidad del viento en estas zonas. A pesar que el ajuste presenta buenos resultados en general, el error de estimación temporal posee colas pesadas, por lo que falla todos los test de normalidad, sin embargo, en las referencias es conocido que estas series temporales resultan ser estacionarias.

El downscaling desde el campo de viento de resolución 1[Km] al campo de viento a 0.5[Km] presenta resultados satisfactorios, siendo capaz de captar en la mayoría de los casos el comportamiento medio y fluctuaciones a nivel temporal provenientes de las series temporales a 1[Km], además de añadir detalles y suavizar las transiciones entre contrastes de intensidad a nivel espacial. Otra característica importante es que este esquema logra captar el aumento del rango de intensidad de viento al pasar a una escala de mayor resolución, mientras que técnicas de interpolación determinista suelen disminuir el rango de valores, perdiendo propiedades físicas de la intensidad del viento. En las zonas en donde el ajuste mediante regresión lineal no es significativo, el esquema de downscaling sigue siendo capaz de captar las fluctuaciones temporales provenientes de las series a 1[Km], aunque a nivel espacial puede presentar problemas de mezcla de pixeles al pasar de una zona con valores de intensidad de viento altas a una zona de baja intensidad.

Con respecto a la clasificación de las distribuciones de series temporales para el residuo, podemos concluir que, considerando cuatro clases, estas tienen una interpretabilidad geográfica. Además, facilita la construcción de bandas de confianza para las series temporales de intensidad de viento a 0.5[Km], tomando en cuenta la homogeneidad espacial dentro de cada clase. Esto permite visualizar la variabilidad en los datos que el modelo de downscaling no es capaz de capturar, entregando estimaciones en casos

extremos de la intensidad del viento. Estas estimaciones son de gran interés la hora de enfrentar catástrofes relacionadas a la intensidad de viento, como incendios e interrupción del suministro eléctrico, tanto como en la planificación energética mediante la evaluación del potencial. Una observación importante es que las bandas de confianza construidas, en general, son pequeñas con respecto al rango de variabilidad de los datos, mientras que en temporadas en el que la intensidad del viento presenta mayor variabilidad, las bandas muestran mayor valor, indicando una incertidumbre en la estimación.

En general, el esquema de downscaling propuesto utiliza fuertemente las hipótesis impuestas, entregando resultados satisfactorios, pues capta las fluctuaciones de la intensidad del viento en toda la zona de estudio y, en general, estima de buena manera la componente de tendencia proveniente desde las escalas de menor resolución. Un punto importante a destacar es que este método es altamente dependiente de la calidad de la estimación del campo a 1.5[Km], pues si se utiliza un interpolador que no sea capaz de captar la relación existente entre los campos con diferentes resoluciones, los resultados para la estimación del viento a 0.5[Km] puede ser pobres, con intensidades sub y sobre determinadas.

Este estudio representa un análisis preliminar de la estructura de dependencia escala-espacio-temporal de los campos de viento obtenidos desde el modelo WRF. Los resultados que se obtienen permitirán desarrollar un nuevo modelo jerárquico espacio-temporal que incorpore la estructura de dependencia del cambio de escala de observación e incorporar nuestra metodología de downscaling para aumentar la resolución de campos aleatorios. Por ejemplo, modificaciones de modelos que usan bases de kernel para predecir lluvias [40] o modelos jerárquicos bayesianos [22] que consideren la información proveniente de mallas de menor resolución, como también modelos que buscan predecir patrones puntuales del comportamiento del viento que consideren las dinámicas provenientes de otras escalas y sus influencias [15].

Bibliografía

[1] T. M. Adams and A. B. Nobel. Uniform approximation of vapnik–chervonenkis classes. *Bernoulli*, 18(4):1310–1319, 11 2012. doi: 10.3150/11-BEJ379. URL http://dx.doi.org/10.3150/11-BEJ379.

[2] P. Ailliot. *Modèles autorégressifs à changements de régimes markoviens. Applications aux séries tempo-relles de vent*. PhD thesis, Université de Rennes 1, 2004.

[3] P. Ailliot, E. Frenod, and V. Monbet. Long term object drift forecast in the ocean with tide and wind. *Multiscale Modeling & Simulation*, 5(2):514–531, 2006.

[4] M.-F. Balcan, A. Blum, and S. Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, pages 671–680, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-047-0. doi: 10.1145/1374376.1374474. URL http://doi.acm.org/10.1145/1374376.1374474.

[5] D. Barker, X.-Y. Huang, Z. Liu, T. Auligné, X. Zhang, S. Rugg, R. Ajjaji, A. Bourgeois, J. Bray, Y. Chen, et al. The weather research and forecasting model's community variational/ensemble data assimilation system: Wrfda. *Bulletin of the American Meteorological Society*, 93(6):831–843, 2012.

[6] R. E. Benestad, I. Hanssen-Bauer, and D. Chen. *Empirical-statistical downscaling*, volume 41. World Scientific, 2008.

[7] F. Bernardin, M. Bossy, C. Chauvin, P. Drobinski, A. Rousseau, and T. Salameh. Stochastic downscaling method: application to wind refinement. *Stochastic Environmental Research and Risk Assessment*, 23(6):851–859, 2009.

[8] P. Billingsley. *Convergence of probability Measures*. Wiley, New York, 1968.

[9] M. Borga. Canonical correlation: a tutorial. Tutorial, Institutionen för medicinsk teknik (IMT), 2001.

[10] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting*. Springer-Verlag, New York, 2002.

[11] G. Casella and R. L. Berger. *Statistical inference*. Duxbury advanced series. Brooks Cole, 2002.

[12] R. Chadwick, E. Coppola, and F. Giorgi. An artificial neural network technique for downscaling gcm outputs to rcm spatial scale. *Nonlinear Processes in Geophysics*, 18(6), 2011.

[13] U. Cubasch, J. Waszkewitz, Hamburg, H. v. Storch, and E. Zorita. Estimates of climate change in southern europe using different downscaling techniques. *Climate Res*, 7:129–149, 1996.

[14] R. Durrett. *Probability : theory and examples*. The Wadsworth & Brooks/Cole statistics/probability series. Wadsworth Inc. Duxbury Press, 1991.

[15] M. Fuentes, L. Chen, J. M. Davis, and G. M. Lackmann. Modeling and predicting complex space–time structures and patterns of coastal wind fields. *Environmetrics*, 16(5):449–464, 2005.

[16] E. Gil. Evaluating the impact of wind power uncertainty on power system adequacy. In *Proceedings of PMAPS*, pages 664–8, 2012.

[17] R. Gray. Probability, random processes, and ergodic properties. 1988.

[18] S. L. Grotch and M. C. MacCracken. The use of general circulation models to predict regional climatic change. *Journal of Climate*, 4(3):286–303, 1991.

[19] L. E. Hay, R. L. Wilby, and G. H. Leavesley. A comparison of delta change and downscaled gcm scenarios for three mounfainous basins in the united states1. *JAWRA Journal of the American Water Resources Association*, 36(2):387–397, Apr. 2000. doi: 10.1111/j.1752-1688.2000.tb04276.x.

[20] N. Helbig, R. Mott, A. Herwijnen, A. Winstral, and T. Jonas. Parameterizing surface wind speed over complex topography. *Journal of Geophysical Research: Atmospheres*, 122(2):651–667, 2017.

[21] C. Lindberg and A. J. Broccoli. Representation of topography in spectral climate models and its effect on simulated precipitation. *Journal of Climate*, 9:2641–2659, nov 1996. doi: 10.1175/1520-0442(1996)009⟨2641:ROTISC⟩2.0.CO;2.

[22] N. McMillan, S. M. Bortnick, M. E. Irwin, and L. M. Berliner. A hierarchical bayesian model to estimate and forecast ozone through space and time. *Atmospheric Environment*, 39(8):1373–1382, 2005.

[23] A. Molinet. Modelizaciòn estocàstica para series de tiempo de vientos. Master's thesis, Universidad Simòn Bolìvar, 2014.

[24] I. Orlanski. A rational subdivision of scales for atmospheric processes. *Bulletin of the American Meteorological Society*, 56:527–530, 1975.

[25] H. Paeth, N. Hall, M. Gartner, M. Alonso, S. Moumouni, J. Polcher, P. Ruti, A. Fink, M. Gosset, T. Lebel, A. Gaye, D. Rowell, W. Moufouma-Okia, D. Jakob, B. Rockel, F. Giorgi, and M. Rummukainen. Progress in regional downscaling of west african precipitation. *Atmospheric Science Letters*, 12:75–82, 2011. doi: 10.1002/asl.306.

[26] J. Risbey and P. Stone. A case study of the adequacy of gcm simulations for input to regional climate change assessments. *Journal of Climate*, 9(7), Jul 1996. doi: 10.1175/1520-0442(1996)009⟨1441:ACSOTA⟩2.0.CO;2.

[27] A. W. Robertson, S. Kirshner, and P. Smyth. Hidden markov models for modeling daily rainfall occurrence over brazil. Technical report, University of California, 2003.

[28] D. Ryabko. Clustering processes. In *Proc. the 27th International Conference on Machine Learning (ICML 2010)*, pages 919–926, Haifa, Israel, 2010.

[29] D. Ryabko and J. Mary. A binary-classification-based metric between time-series distributions and its use in statistical and learning problems. *Journal of Machine Learning Research*, 14:2837–2856, 2013.

[30] L. Seaby, J. Refsgaard, T. Sonnenborg, S. Stisen, J. Christensen, and K. Jensen. Assessment of robustness and significance of climate change signals for an ensemble of distribution-based scaled climate projections. *Journal of Hydrology*, 486: 479–493, 2013. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2013.02.015.

[31] A. Shashua. Introduction to machine learning: Class notes 67577. *CoRR*, abs/0904.3664, 2009.

[32] R. H. Shumway and D. S. Stoffer. *Time series analysis and its applications : with R examples*. Springer texts in statistics. Springer, 2006.

[33] S. Trzaska and E. Schnarr. A review of downscaling methods for climate change projection. Technical report, U..., 2014.

[34] T. Vandal, E. Kodra, and A. R. Ganguly. Intercomparison of machine learning methods for statistical downscaling: The case of daily and extreme precipitation. *arXiv preprint arXiv:1702.04018*, 2017.

[35] M. T. Vu, T. Aribarg, S. Supratid, S. V. Raghavan, and S.-Y. Liong. Statistical downscaling rainfall using artificial neural network: significantly wetter bangkok? *Theoretical and applied climatology*, 126(3-4):453–467, 2016.

[36] Q. Wang, W. Shi, P. M. Atkinson, and Y. Zhao. Downscaling modis images with area-to-point regression kriging. *Remote Sensing of Environment*, 166:191–204, 2015.

[37] Q. Wang, W. Shi, P. M. Atkinson, and E. Pardo-Igúzquiza. A new geostatistical solution to remote sensing image downscaling. *IEEE Transactions on Geoscience and Remote Sensing*, 54(1):386–396, 2016.

[38] L. Wasserman. *All of Nonparametric Statistics, ser.* Springer Texts in Statistics. New York: Springer-Verlag, 2006.

[39] R. L. Wilby, J. Troni, Y. Biot, L. Tedd, B. C. Hewitson, D. M. Smith, and R. T. Sutton. A review of climate risk information for adaptation and development planning. *International Journal of Climatology*, 29(9):1193–1215, 2009. ISSN 1097-0088. doi: 10.1002/joc.1839. URL http://dx.doi.org/10.1002/joc.1839.

[40] K. Xu, C. K. Wikle, and N. I. Fox. A kernel-based spatio-temporal dynamical model for nowcasting weather radar reflectivities. *Journal of the American statistical Association*, 100(472):1133–1144, 2005.

[41] Z. Zeng, W. W. Hsieh, W. R. Burrows, A. Giles, and A. Shabbar. Surface wind speed prediction in the canadian arctic using non-linear machine learning methods. *Atmosphere-Ocean*, 49(1):22–31, 2011.

[42] V. M. Zolotarev and B. Seckler. Probability metrics. *Theory of Probability and its Applications*, 28:278–302, 1983.

[43] E. Zorita and H. von Storch. The Analog Method as a Simple Statistical Downscaling Technique: Comparison with More Complicated Methods. *J. Climate*, 12(8):2474–2489, Aug. 1999. doi: 10.1175/1520-0442(1999)012\%3C2474:tamaas\%3E2.0.co;2. URL http://dx.doi.org/10.1175/1520-0442(1999)012%3C2474:tamaas%3E2.0.co;2.