UNIVERSIDAD
NACIONAL
DE COLOMBIA

Pronóstico de velocidad de viento para generación eólica Offshore basado en programación genética

Oscar Alberto Garrido Atencia

Universidad Nacional de Colombia
Facultad de Ingeniería
Departamento Ingeniería Eléctrica y Electrónica
Bogotá D.C., Colombia
2020

Pronóstico de velocidad de viento para generación eólica Offshore basado en programación genética

Oscar Alberto Garrido Atencia

Tesis presentada como requisito parcial para optar al título de:

Magister en Ingeniería Eléctrica

Director (a):

Sergio Raúl Rivera Rodríguez, Ph.D

Línea de Investigación:

Sistemas de Potencia

Grupo de Trabajo:

Inteligencia Computacional aplicada al Sector Eléctrico, EMC - UN

Grupo de Investigación:

Electromagnetic Compatibility EMC - UN

Universidad Nacional de Colombia

Facultad de Ingeniería

Departamento Ingeniería Eléctrica y Electrónica

Bogotá D.C., Colombia

2020

A mi madre y hermanos, quienes siempre me han acompañado en los diferentes retos que he emprendido.

A esa persona especial que ha estado conmigo en esta etapa y siembre ha tenido palabras de aliento que me exhortaban a no darme por vencido hasta conseguir la meta.

Declaración de obra original

Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

_____

Nombre

Oscar Garrido Atencia

Fecha 15/02/2021

Agradecimientos

A mi director de tesis, el ing. Sergio Raul Rivera, Ph.D, por alentarme a sumir nuevos retos y conseguir grandes metas.

A la Universidad Nacional, por dame la oportunidad de conocer personas brillantes y por darme las herramientas para ser un mejor profesional.

A Alejandra, quien fue más que una compañera durante este reto; ya que mutuamente fuimos un apoyo para seguir adelante y no desfallecer hasta alcanzar la meta.

Resumen

Título: Pronóstico de velocidad de viento para generación eólica Offshore basado en programación genética

Las energías renovables han surgido como la alternativa más viable para solucionar los problemas que presentan las fuentes de generación convencionales. En este sentido, la generación eólica offshore cuenta con gran potencial de crecimiento para los próximos años es por esto que el presente trabajo plantea una metodología que implementa la programación genética para realizar pronósticos de vientos promedio a mediano y largo plazo, con el fin de minimizar la incertidumbre asociada a este tipo de generación.

Para esto, inicialmente se realiza el planteamiento del algoritmo regresión simbólica híbrida por medio del cual se realizarán los pronósticos de vientos propuestos; realizando una descripción del funcionamiento de este.

Posteriormente se realiza la implementación del algoritmo planteado en cuatro casos de estudio ubicados en zonas costeras y en islas, de tal manera que se disponga de históricos de datos meteorológicos con los cuales poder realizar las pruebas del algoritmo. Posterior a esto, se evaluarán los errores obtenidos para seleccionar una cantidad de datos para entrenamiento y prueba del algoritmo.

Palabras clave: Pronóstico, energía eólica, generación eólica offshore, programación genética, regresión simbólica.

Abstract

Título en inglés: Wind speed forecast for offshore wind generation based on genetic programming

Renewable energies have emerged as the most viable alternative to solve the problems presented by conventional generation sources. offshore wind generation has great growth potential for the next years, which is why this work proposes a methodology that implements genetic programming to make forecasts of average winds in the medium and long term, to minimize the uncertainty associated with this kind of generation.

To that, initially the approach of the hybrid symbolic regression algorithm is carried out by means of which the proposed wind forecasts will be made; making a description of how it works.

Subsequently, the implementation of the algorithm proposed in four case studies located in coastal areas and on islands is carried out, so that historical meteorological data are available with which to carry out the algorithm tests. After this, the errors obtained will be evaluated to select an amount of data for training and testing the algorithm.

Keywords: forecast, wind power, offshore wind generation, genetic programming, symbolic regression.

Contenido

Pág.

Resumen ..... **IX**

Lista de figuras ..... **XII**

Lista de tablas ..... **XV**

Lista de Símbolos y abreviaturas ..... **XVI**

Introducción **1**

1. Estado del arte **3**
1.1 Técnicas para la elaboración de pronósticos de vientos 4
1.2 Fuentes para la obtención de datos meteorológicos 5

2. Modelo de algoritmo genético para pronósticos de vientos **6**
2.1 Aplicación de la programación genética para la realización de pronósticos 7
2.2 Planteamiento de algoritmo de regresión simbólica para pronósticos de vientos 9
2.2.1 Algoritmo de regresión simbólica multigen 9
2.2.2 Métodos de kernel en procesos de regresión 11

3. Implementación del algoritmo de regresión simbólica para pronósticos de viento **17**
3.1 Planteamiento de los casos de estudio 17
3.2 Resultados obtenidos de los casos de estudio 19
3.2.1 Resultados caso de estudio 1 20
3.2.2 Resultados caso de estudio 2 27
3.2.3 Resultados caso de estudio 3 35
3.2.4 Resultados caso de estudio 4 41
3.3 Análisis de resultados obtenidos 48
3.4 Cumplimiento de los objetivos 50

4. Conclusiones y recomendaciones **53**
4.1 Conclusiones 53
4.2 Trabajos futuros 54

A. Anexo: Algoritmo de regresión simbólica híbrido en Matlab **55**

Bibliografía **65**

Lista de figuras

Pág.

Figura 2-1: Funcionamiento general del GP. 7
Figura 2-2: Ejemplo de individuo multi-gen. 10
Figura 3-1: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de un año de datos con un porcentaje de datos de entrenamiento del 90%. 21
Figura 3-2: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de un año de datos con un porcentaje de datos de entrenamiento del 50%. 21
Figura 3-3: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de un año de datos con un porcentaje de datos de entrenamiento del 30%. 22
Figura 3-4: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de dos años de datos con un porcentaje de datos de entrenamiento del 90%..... 23
Figura 3-5: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de dos años de datos con un porcentaje de datos de entrenamiento del 50%..... 24
Figura 3-6: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de dos años de datos con un porcentaje de datos de entrenamiento del 30%..... 24
Figura 3-7: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de tres años de datos con un porcentaje de datos de entrenamiento del 90%. 26
Figura 3-8: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de tres años de datos con un porcentaje de datos de entrenamiento del 50%. 26
Figura 3-9: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de tres años de datos con un porcentaje de datos de entrenamiento del 30%. 27
Figura 3-10: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de un año de datos con un porcentaje de datos de entrenamiento del 90%. 28
Figura 3-11: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de un año de datos con un porcentaje de datos de entrenamiento del 50%. 29
Figura 3-12: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de un año de datos con un porcentaje de datos de entrenamiento del 30%. 29
Figura 3-13: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de dos años de datos con un porcentaje de datos de entrenamiento del 90%..... 31
Figura 3-14: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de dos años de datos con un porcentaje de datos de entrenamiento del 50%..... 31
Figura 3-15: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de dos años de datos con un porcentaje de datos de entrenamiento del 30%..... 32
Figura 3-16: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de tres años de datos con un porcentaje de datos de entrenamiento del 90%. 33

Figura 3-17: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de tres años de datos con un porcentaje de datos de entrenamiento del 50%..... 34
Figura 3-18: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de tres años de datos con un porcentaje de datos de entrenamiento del 30%..... 34
Figura 3-19: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de un año de datos con un porcentaje de datos de entrenamiento del 90%… 36
Figura 3-20: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de un año de datos con un porcentaje de datos de entrenamiento del 50%… 36
Figura 3-21: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de un año de datos con un porcentaje de datos de entrenamiento del 30%… 37
Figura 3-22: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de dos años de datos con un porcentaje de datos de entrenamiento del 90%. 38
Figura 3-23: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de dos años de datos con un porcentaje de datos de entrenamiento del 50%. 38
Figura 3-24: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de dos años de datos con un porcentaje de datos de entrenamiento del 30%. 39
Figura 3-25: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de tres años de datos con un porcentaje de datos de entrenamiento del 90%. 40
Figura 3-26: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de tres años de datos con un porcentaje de datos de entrenamiento del 50%. 40
Figura 3-27: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de tres años de datos con un porcentaje de datos de entrenamiento del 30%. 41
Figura 3-28: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de un año de datos con un porcentaje de datos de entrenamiento del 90%… 42
Figura 3-29: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de un año de datos con un porcentaje de datos de entrenamiento del 50%… 43
Figura 3-30: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de un año de datos con un porcentaje de datos de entrenamiento del 30%… 43
Figura 3-31: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de dos años de datos con un porcentaje de datos de entrenamiento del 90%. 44
Figura 3-32: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de dos años de datos con un porcentaje de datos de entrenamiento del 50%. 45
Figura 3-33: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de dos años de datos con un porcentaje de datos de entrenamiento del 30%. 45

Figura 3-34: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de tres años de datos con un porcentaje de datos de entrenamiento del 90%. 46
Figura 3-35: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de tres años de datos con un porcentaje de datos de entrenamiento del 50%. 47
Figura 3-36: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de tres años de datos con un porcentaje de datos de entrenamiento del 30%. 47
Figura 3-37: Logro de objetivos por secciones del documento. 51

Lista de tablas

Pág.

Tabla 3.1: Cantidad de datos por caso de estudio 19
Tabla 3.2: Resultados de pronósticos del caso de estudio 1 para un año. 20
Tabla 3.3: Resultados de pronósticos del caso de estudio 1 para dos años..... 23
Tabla 3.4: Resultados de pronósticos del caso de estudio 1 para tres años. 25
Tabla 3.5: Resultados de pronósticos del caso de estudio 2 para un año. 27
Tabla 3.6: Resultados de pronósticos del caso de estudio 2 para dos años..... 30
Tabla 3.7: Resultados de pronósticos del caso de estudio 2 para tres años. 32
Tabla 3.8: Resultados de pronósticos del caso de estudio 3 para un año. 35
Tabla 3.9: Resultados de pronósticos del caso de estudio 3 para dos años..... 37
Tabla 3.10: Resultados de pronósticos del caso de estudio 3 para tres años..... 39
Tabla 3.11: Resultados de pronósticos del caso de estudio 4 para un año. 42
Tabla 3.12: Resultados de pronósticos del caso de estudio 4 para dos años. 44
Tabla 3.13: Resultados de pronósticos del caso de estudio 4 para tres años..... 46

Lista de Símbolos y abreviaturas

Símbolos con letras latinas

Símbolo	Término	Unidad SI	Definición
f	Función simbólica		
F	Espacio característico de kernel		
G	Matriz Gram		$X'X$
I	Matriz identidad		
k	Vector de producto escalar		
l	Cantidad de observaciones		
X	Matriz de variables experimentales		
Y	Matriz de resultados de inferencias		

Símbolos con letras griegas

Símbolo	Término	Unidad SI	Definición
α	Vector de factores de peso		
β			$X(X'X)^{-2}X'Y$
λ	Termino de compensación matricial		
$\varphi(\cdot)$	Función de mapeo en espacio característico de kernel		

Subíndices

Subíndice	Término
p	Número de variables predictivas

Superíndices

Superíndice	Término
$'$	Transponer matriz
$\hat{Y}$	Promedio de Y
n	Exponente, potencia

Abreviaturas

Abreviatura	Término
DEEPSO	Enjambre de partículas evolutivas diferenciales
ELM	Máquina de aprendizaje extremo
GP	Programación genética
MCO	Mínimos cuadrados ordinarios
ML	Aprendizaje automático
PSO	optimización de enjambre de partículas
SPM	Máquina de soporte vectorial
SR	Regresión simbólica

Introducción

Los sistemas eléctricos de potencia han ido evolucionado para poder adaptarse a las condiciones cambiantes y poder tener así un buen desempeño en aspectos técnicos y de seguridad [1].

Los sistemas eléctricos de potencia basados en energías renovables se encuentran en constante desarrollo debido a su potencial de crecimiento. Para el caso de la generación de energía eólica offshore, en los últimos años se han realizado diversas investigaciones sobre el potencial de generación eólico offshore a lo largo del mundo [2], [3].

Uno de los principales inconvenientes en la aplicación de fuentes renovables de energía para la generación de energía eléctrica es la incertidumbre asociada a dichas fuentes, como lo es la radiación solar, la velocidad del viento, entre otras.

Debido a lo anterior, se han ido desarrollando diferentes técnicas y metodologías que permitan minimizar las incertidumbres mencionadas. El presente trabajo propone la implementación de una metodología derivada de la programación genética para elaborar pronósticos de vientos promedio aplicado a ubicaciones costeras o islas donde se pueda evaluar la posibilidad de instalar un parque de generación eólica offshore.

La metodología a implementar es una regresión simbólica hibrida, que toma las propiedades y ventajas de la programación genética y las combina con regresiones de cresta por medio de matrices de kernel para dar como resultado un algoritmo híbrido con el cual se puedan tomar un conjunto de variables meteorológicas correlacionadas con la velocidad del viento para elaborar pronósticos a mediano y largo plazo.

En el presente trabajo se encuentra distribuido de la siguiente manera: en el capítulo 1 se realizó una revisión bibliográfica del estado del arte respecto a las diferentes metodologías y técnicas que se han implementado en la realización de pronósticos de vientos para

plantas eólicas. En el capítulo 2, se realizó el planteamiento del método de programación genética a implementar para la elaboración de pronósticos de vientos. En el tercer capítulo se realizaron las simulaciones con el modelo planteado y con base al cuál se realizaron los pronósticos de vientos para plantas eólicas offshore en diferentes ubicaciones geográficas para diferentes períodos de tiempos y así evaluar el desempeño del modelo. Finalmente, en el cuarto capítulo se presentan las conclusiones recomendaciones fruto de los resultados obtenidos.

A continuación, se presentan los objetivos propuesto a desarrollar en el presente trabajo:

Objetivos

Objetivo general:

Realizar el pronóstico de la velocidad promedio del viento por medio de la implementación de programación genética basada en regresión simbólica para locaciones con fines de generación eólica Offshore.

Objetivos específicos:

1. Obtener diferentes conjuntos de datos de históricos meteorológicos con base a los cuales se realizarán los pronósticos de vientos.

2. Plantear un método de regresión simbólica con base al cual se van a realizar los pronósticos de las velocidades de vientos promedio correlacionadas a las variables meteorológicas disponibles.

3. Implementar el modelo de pronóstico propuesto en diferentes conjuntos de datos decidiendo un número apropiado de datos de entrenamiento y prueba, para validar su funcionamiento.

1. Estado del arte

Las incertidumbres en la generación de energía con fuentes renovables es uno de los grandes retos que afrontan los operadores de los sistemas de potencia ante la penetración de este tipo de generación debido al alto componente de aleatoriedad que introducen al sistema; debido a esto se deben tomar diferentes medidas encaminadas a mantener la operación del sistema dentro de los parámetros de seguridad [4].

En particular, las plantas de generación eólicas se han ido introduciendo al sector de la generación en los sistemas de potencia debido a que entre sus características principales está el beneficio de que su fuente primaria -el viento- es limpia, inagotable y no contaminante; gracias a esto se considera una buena alternativa para mitigar los problemas ambientales ocasionados por la generación de energía tradicional [5]. De las diferentes plantas de generación eólica, la generación eólica offshore se proyecta como una alternativa que cuenta con un potencial de crecimiento esperado para los próximos 10 años de 30 GW de capacidad instalada [6].

Como se mencionaba anteriormente, uno de los grandes inconvenientes de la inclusión de las energías renovables en los sistemas de potencia es su alto contenido de incertidumbre en los niveles de generación. En particular, la energía eólica tiene la complejidad de que el viento posee una naturaleza estocástica, no lineal y una alta correlación con las condiciones meteorológicas del entorno [7]. Como medida para contrarrestar esto, a lo largo de los años se han ido desarrollando diferentes técnicas y metodologías que permiten realizar predicciones de las fuentes primarias usadas para la generación y de esta forma disminuir las incertidumbres asociadas al viento. A continuación, se plantea una revisión del estado del arte de las técnicas implementadas para la predicción de las velocidades de vientos en la generación eólica:

1.1 Técnicas para la elaboración de pronósticos de vientos

Algunos de los principales modelos clásicos que se han utilizado para la proyección de velocidades de vientos son los estocásticos, determinísticos y por medio de regresiones lineales y no lineales [8]–[10].

Otros métodos más recientes utilizados para pronósticos de vientos son las redes neuronales [11], [12], máquenas de soporte vectorial (SVM-por sus siglas en inglés) [13], machine learning (ML) y extreme learning machine (ELM) [14]–[17] y fuzzy logic [18]. Esto debido a que las capacidades en el reconocimiento de patrones de los modelos mencionados dan mejores resultados que los modelos tradicionales [19].

Hoy en día, se han desarrollado metodologías híbridas que combinan las diferentes metodologías para resaltar los puntos fuertes de cada y minimizar las desventajas individuales que presentan obteniendo como resultado una mejora significativa en la precisión de los pronósticos que involucran series de viento [20]. Algunos de los modelos híbridos que se han desarrollado se presentan a continuación:

En [21] se desarrolló un nuevo marco generalizado para realizar pronósticos probabilísticos de energía eólica con alta flexibilidad y confiabilidad mediante la implementación de un algoritmo híbrido inteligente con extreme machine learning y la optimización de enjambre de partículas (PSO-por sus siglas en inglés) para obtener intervalos óptimos de predicción de generación de energía eólica.

Un híbrido entre algoritmo neuronal y difuso fue propuesto en [7] donde se utilizó una red de sensores inalámbricos para contemplar en el modelo los factores primordiales que tienen relevancia en la potencia de salida de un parque eólico – temperatura del aire, velocidad del viento, densidad y presión del aire- para luego usar un conjunto de reglas difusas para implementar el método de predicción.

Por su parte, en [22] los autores usaron valores medios diarios de velocidades de viento para entrenar diversos algoritmos basados en machine learning; dichos algoritmos entrenados para realizar los pronósticos fueron xGBoost, SVR y RF. Luego, estos fueron usados para realizar pronósticos de vientos para diferentes ubicaciones. El horizonte de tiempo para pronósticos de vientos con esta metodología fue de un año.

Otra metodología, planteada en [23], buscaba realizar pronósticos de viento para un horizonte de tiempo de 72 horas por medio de la implementación de los algoritmos GRPE y DEPE por medio del entrenamiento de modelos de pronósticos recurrentes. Otros autores han utilizado diferentes métodos para realizar pronósticos de vientos para intervalos desde el orden de algunos segundos hasta algunos días [10], [20], [24], [25].

1.2 Fuentes para la obtención de datos meteorológicos

En el proceso de revisión del estado del arte se encontró que muchos los artículos consultados tomaban los valores de históricos de vientos de conjuntos de estaciones meteorológicas o de entidades gubernamentales según los lugares donde se realizaron los estudios pertinentes. Sin embargo, no se hacía especial aclaración de la fuente para la obtención de los datos; debido a esto, se realizó una búsqueda de diferentes fuentes meteorológicas que permitiesen acceder a conjuntos de datos históricos de vientos para varios años consecutivos. Algunas de las fuentes consultadas fueron el data set de IEEE PES [26], el data set de la NASA [27] y las presentadas en [28], [29]. De esta última se tomaron los datos históricos de diferentes parámetros meteorológicos para las locaciones seleccionadas dado que permite el uso de sus datos para fines no comerciales.
Con lo anterior se le da cumplimiento al primer objetivo específico:

1. Obtener diferentes conjuntos de datos de históricos meteorológicos con base a los cuales se realizarán los pronósticos de vientos.

2. Modelo de algoritmo genético para pronósticos de vientos

La programación genética (GP -por sus siglas en inglés) es una técnica de computación evolutiva que busca resolver problemas de manera automática sin la necesidad de ingresarle a la computadora textualmente cómo lo debe realizar. En síntesis, GP es un método sistemático y de dominio independiente para conseguir computadoras que puedan resolver problemas automáticamente partiendo desde una declaración de alto nivel de lo que se debe realizar [30].

La GP está basada en la evolución biológica y lo que busca es desarrollar algoritmos que se adapten a una tarea particular. La GP se inicia generando de manera aleatoria una población de programas individuales, los cuales serán evaluados según una función de comportamiento específica según la tarea. Posterior a esto, se aplican operaciones genéticas como la mutación y los cruces que generan nuevas poblaciones descendientes a partir de los individuos con mejores aptitudes, esperando que las aptitudes de los programas individuales mejoren a lo largo que se van desarrollando nuevas poblaciones [31]. En la **Figura 2-1** se presenta el funcionamiento de la programación genética por medio de un diagrama de flujo.

Figura 2-1: Funcionamiento general del GP.

Las aplicaciones de la GP son bastantes numerosas y en campos variados, pero se puede consideras a groso modo que las principales áreas de aplicación son en ajuste de curvas, procesamiento de señales e imágenes, modelamiento de datos, predicción en series de tiempo, modelamiento económico y regresión simbólica [30]. Es en esta última en la que se centra el presente trabajo para realizar pronósticos de tiempos con vistas a la generación de energía eólica offshore.

2.1 Aplicación de la programación genética para la realización de pronósticos

Gracias a su flexibilidad en la generación de modelos matemáticos de manera libre a partir de datos observados sin ningún tipo de conocimiento experto, la regresión simbólica (SR - por sus siglas en inglés) es una de las aplicaciones más populares de programación genética así como una alternativa atractiva a los enfoques de regresión estándares [32].

La regresión simbólica es considerada una actividad de aprendizaje inductivo que tiene como finalidad hallar una expresión matemática -o modelo- que permita ajustar los datos disponibles por medio de un proceso de entrenamiento [33].

El análisis por medio de regresión es un método conceptual simple que sirve para investigar relaciones funcionales entre múltiples variables. Esta relación se expresa como una ecuación o un modelo que conecta la variable dependiente, representada como Y, con una o más variables predictivas o independientes, denotadas como $X_1, X_2, \ldots, X_p$, donde p es el número de variables predictivas [34]. Un modelo típico de regresión es presentado en Ecuación (2.1):

$$Y = f(X_1, X_2, \ldots, X_p) \tag{2.1}$$

Donde f es representa una función simbólica no lineal. El modelo resultante tendrá funciones lineales como no lineales, donde la estructura y adaptación de los términos son seleccionados de manera automática por el algoritmo de SR [35].

Para la predicción de datos, se utiliza la representación de las expresiones matemáticas que correspondientes a diferentes modelos como una estructura de árbol; donde los nodos interiores (unión entre terminales y otros operadores) vendrían a ser los operadores matemáticos y las hojas del árbol serían las variables y parámetros. El proceso algorítmico está basado en la búsqueda dentro de un campo de ecuaciones factibles a través de la aplicación de operadores genéticos tales como la mutación, el cruce y la reproducción [36]. Para la creación del árbol este relaciona los genes con expresiones aritméticos y funciones trigonométricas. Luego, la población es organizada por jerarquías y crea una nueva generación basado en el principio de recombinación genética [37].

Normalmente la evaluación de los individuos se realiza con el error de aproximación promedio, como el error cuadrático medio sobre todo el conjunto de observaciones. De esta forma, se calcula la probabilidad de participación de cada individuo en las operaciones genéticas. Así se crea una nueva población que luego será evaluada nuevamente produciendo un proceso iterativo hasta que se cumpla la regla de terminación, la cual está dada por el número máximo de generaciones que establezca el usuario; de esta forma se espera que cada población nueva tenga un mejor desempeño que la anterior. Luego, el modelo simbólico que se selecciona es aquel donde se haya tenido un menor grado de error aproximado.

2.2 Planteamiento de algoritmo de regresión simbólica para pronósticos de vientos

El algoritmo de regresión simbólica que se plantea utilizar fue el implementado en [38] y en [39]; en el primer caso fue implementado como una etapa inicial de un algoritmo híbrido que combina la regresión simbólica con la optimización por medio del algoritmo DEEPSO para realizar el pronóstico de demanda anual.

El presente trabajo busca utilizar el algoritmo de regresión simbólica no lineal antes mencionado para realizar el pronóstico de vientos de cuatro locaciones costeras para que pueda ser implementado en pronósticos de generación eólica offshore. A continuación, se explicará el funcionamiento de dicho algoritmo:

2.2.1 Algoritmo de regresión simbólica multigen

El algoritmo de regresión simbólica normalmente contempla cada modelo de solución individual como un solo gen (o estructura de árbol) el cual da como salida una sola ecuación matemática. Sin embargo, el algoritmo seleccionado implementa una variación propuesta en [30] y [40] donde se modifica el enfoque tradicional de tal manera que se considera que cada individuo está conformado por un conjunto de genes que combinados linealmente, como se muestra en la **Figura 2-2**, dan como resultado la inferencia.

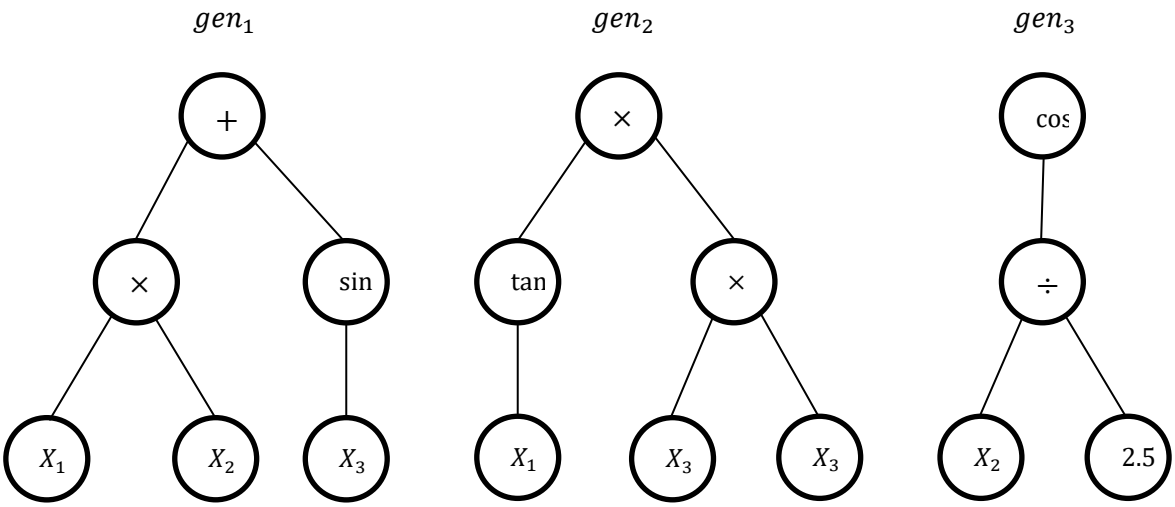
El modelo resultante dado por el individuo multi-gen se calcula como se observa en las Ecuaciones (más adelante) y (más adelante).

$$Y = \alpha_0 + \alpha_1 gen_1 + \alpha_2 gen_2 + \alpha_3 gen_3 \tag{2.2}$$

$$Y = \alpha_0 + \alpha_1((X_1 X_2) + \sin X_3) + \alpha_2({X_3}^2 * \tan X_1) + \alpha_3 \cos \frac{X_2}{2.5} \tag{2.3}$$

Donde X_i, con $i = \{1,2,3\}$, son los valores de las variables predictivas (o independientes), α_i son los factores de regresión o factores de peso y Y es el resultado de la inferencia.

Figura 2-2: Ejemplo de individuo multi-gen.



En la siguiente etapa del algoritmo se deben hallar los valores de los factores de regresión que dan como resultado una expresión que ajusta las observaciones con la mayor exactitud con respecto al valor real. Para esto, en [30] y [40] se planteó el método de los mínimos cuadrados ordinarios (MCO); de lo cual se tiene como solución la expresión mostrada en la Ecuación $\alpha=(X'X)^{-1}X'Y$ (2.4). De dicha ecuación se tiene que se genera una matriz como resultado de calificar el conjunto de observaciones en los genes de los diferentes individuos; dicha matriz; el tamaño de la matriz está determinado por la cantidad de observaciones y de genes, donde la cantidad de filas está dada por el número de observaciones y la cantidad de columnas está dada por el número de genes más uno.

$$\alpha = (X'X)^{-1}X'Y \tag{2.4}$$

Los resultados de la inferencia Y se pueden calcular mediante la Ecuación 5 una vez que se cuenta con los valores del vector α. Luego, dado que el rendimiento está definido por el error, este puede ser calculado como indica la Ecuación 6 por medio de la sumatoria de cuadrados, donde n es el número de observaciones.

$$Y = X\alpha \tag{2.5}$$

$$e = \sum_{i=1}^{n} \left\| Y_i - \hat{Y}_i \right\|^2 \tag{2.6}$$

2.2.2 Métodos de kernel en procesos de regresión

El procedimiento de regresión simbólica multi-gen que se ha considerado hasta el momento lleva a realizar la búsqueda de un vector compuesto por los factores de ponderación que combinen linealmente el conjunto de genes que componen un individuo. Aun cuando los autores de este enfoque encontraron valores apropiados de inferencias, se genera una inquietud sobre el comportamiento del algoritmo si el supuesto de que la relación lineal entre los genes de un individuo se mantiene verdadera cuando se considera una relación no-lineal.

Partiendo de la teoría detrás del algoritmo multigen y las operaciones matemáticas relacionadas con la regresión lineal por MCO, se puede considerar la adición de una técnica de regresión (principalmente lineal) cuya representación puede funcionar en la exploración de relaciones no lineales.

Teniendo en cuenta el procedimiento que se ha planteado hasta el momento, el encontrar de un vector de pesos en el método de MCO está basado en la existencia de una matriz inversa $(X'X)^{-1}$. Una alternativa viable para evitar la singularidad de dicha matriz es el proceso conocido como regresión de cresta. En este método, el problema de los mínimos cuadrados es expresado por medio de su representación dual y la regularización se considera incluyendo un término de equilibrio entre la norma del vector de pesos y minimizar el error de aproximación. Adicionalmente, este enfoque presenta la ventaja de que las ecuaciones matemáticas halladas dan lugar a la introducción de nociones de

funciones de kernel [41], [42]. La razón por la cual las funciones de kernel son incluidas los procesos de regresiones lineal es para llevar a cabo la regresión no lineal transportando las observaciones experimentales a un nuevo espacio o espacio característico, por lo cual realizar una regresión lineal es equivalente a realizar una regresión no lineal dentro del espacio original [42].

- **Representación dual del problema de regresión lineal**

Partiendo del supuesto que existe la matriz inversa $(X'X)^{-1}$, se puede realizar el cálculo del vector de ponderación como se plantea en las Ecuaciones α=$X'X(X'X)^{-2}X'Y = X'\beta$ (2.7) y (más adelante

$$\alpha = X'X(X'X)^{-2}X'Y = X'\beta \tag{2.7}$$

$$\beta = X(X'X)^{-2}X'Y \tag{2.8}$$

Estas ecuaciones anteriores son conocidas como la representación dual de un problema de regresión lineal [41], [42]. Esto representa el vector de peso α como una combinación lineal de los datos en la matriz X que se construye a partir de datos experimentales de las variables utilizadas para entrenar el sistema de inferencia. Este tipo de representación será útil más adelante cuando se considere agregar funciones de kernel en el proceso de regresión.

En particular, hay dos escenarios donde no es posible encontrar los factores de ponderación tales que se tenga el ajuste de los datos de prueba. Dichos escenarios son: insuficiencia en la cantidad de datos de prueba de tal forma que no asegure la ausencia de singularidad de la matriz $X'X$ o la presencia de datos que introduzcan mucho ruido. Para el caso donde se tiene alto ruido es posible la inclusión del concepto de regularización agregando un término de equilibro entre el nivel de minimización de errores (precisión) y la minimización de la norma de vector de peso (complejidad). Este concepto es incorporado en la regresión de crestas.

- **Regresión de Crestas: representación dual con regularización**

Los cálculos de regresión de crestas incluyen un término de compensación λ entre minimizar la norma del vector de peso y el error de aproximación. Teniendo en cuenta esto, la Ecuación más atrás se reescribe como la Ecuación más adelante, donde λ es un valor positivo, la no singularidad de la matriz $(X'X + \lambda I)$, por lo tanto, la existencia de su inversa está garantizada.

$$\beta = (X'X + \lambda I)^{-1}Y \tag{2.9}$$

En la Ecuación más atrás la matriz $X'X$ es también conocida como la matriz de Gram y define el producto escalar entre los datos en X.

En la regresión de cresta, se utiliza la Ecuación (más adelante) para realizar la inferencia basada en una nueva observación o conjunto de datos (x), es decir, $k_i = \langle x_i, x \rangle$, donde k es el vector que comprende el producto escalar entre cada vector l de datos que se tuvieron en cuenta durante el entrenamiento y los datos vectoriales que producirán la inferencia.

$$\hat{y} = \langle \alpha, x \rangle = y'(X'X + \lambda I)^{-1}k \tag{2.10}$$

- **Regresión de cresta no-lineal**

Los procesos de regresión lineal de mínimos cuadrados y regresión de cresta buscan resolver el problema de encontrar una relación lineal entre una variable dependiente y otras variables independientes, que son los genes de un individuo en regresión simbólica multigen. Aunque el uso de una relación lineal no conducirá necesariamente a los mejores resultados, existe la posibilidad de obtener buenos resultados al utilizar una relación no lineal dentro del proceso de inferencia. Esta es una posibilidad considerada por los métodos del kernel. Estos métodos llevan los datos a un espacio conocido como espacio característico por medio del uso de una función de mapeo $\varphi(\cdot)$. De esta manera, una relación lineal entre los datos experimentales en el espacio característico es equivalente a tener una relación no lineal en el espacio original [39], [42].

Si hay un conjunto de datos de entrenamiento u observaciones $((x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l))$ en el espacio original, una nueva base de datos, a saber $((\varphi(x_1), y_1), (\varphi(x_2), y_2), \ldots, (\varphi(x_l), y_l))$, es generada al aplicar la función de mapeo. El objetivo ahora es llevar a cabo la regresión lineal en la nueva base de datos mientras se mantiene al mínimo el error de aproximación.

En la regresión de cresta, la estimación en el nuevo espacio se está dada por la Ecuación (más adelante), donde $G_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$ y $k_i = \langle \varphi(x_i), \varphi(x) \rangle$.

$$\hat{y} = y'(G + \lambda I)^{-1} k \tag{2.11}$$

La inclusión de la matriz de Gram en el proceso de inferencia permite el uso de funciones de kernel para realizar el cálculo del producto escalar entre cada punto de datos de manera eficiente; lo cual en este caso, viene a significar que no es necesario conocer la función de mapeo $\varphi(\cdot)$ o la ubicación de los datos dentro del nuevo espacio después de su aplicación.

La función de kernel es una función que se encarga de calcular el producto interno dentro de un espacio de características directamente desde las entradas en el espacio original. Según la definición dada en [42], un kernel es una función $k(\quad)$ tal que, para todo $\mathrm{x}, \mathrm{z} \in X$, satisface la Ecuación (más adelante), donde $\varphi(\cdot)$ responde a un mapeo del espacio original X en un nuevo espacio característico F (Ecuación (más adelante)).

$$k(x, z) = \langle \varphi(x), \varphi(z) \rangle \tag{2.12}$$

$$\varphi(\cdot): x \rightarrow \varphi(x) \in F \tag{2.13}$$

El algoritmo de regresión simbólica no lineal híbrido implementado propone establecer una relación no lineal entre los genes utilizando funciones de kernel mediante la regresión de crestas no lineales.

Durante la evaluación del desempeño de cada individuo sobre cómo estos se ajustan a la información ingresada inicialmente para llevar a cabo entrenamiento, se lleva a cabo una regresión lineal por MCO en el algoritmo de regresión simbólica multigen, donde se halla el vector de peso que genera el valor de estimación con menor discrepancia con respecto al valor real de una variable de interés. Por el contrario, el algoritmo a implementar propone expresar el proceso de regresión que tiene lugar en un algoritmo de regresión simbólica multigen en términos de regresión de crestas. Esto permite la inclusión de funciones del kernel a través de la matriz Gram o matriz de producto interno (G). El funcionamiento general del algoritmo de regresión simbólica permanece sin cambios, excepto en las etapas donde se codifican los individuos y se miden sus desempeños. Esto se logra al incluir el cálculo de la matriz de Gram para el funciones kernel del tipo lineal, polinomial o gaussiano y calcular la inferencia aplicando la Ecuación (más atrás) [39].

En este sentido, el algoritmo a implementar ejecuta una nueva versión multigen del algoritmo de regresión simbólica que incluye el concepto de regresión no lineal a través de funciones del kernel.

Previo a la ejecución del algoritmo el usuario debe realizar la configuración de algunos parámetros e ingresar la base de datos con los datos para llevar a cabo el entrenamiento. A nivel general, el algoritmo está compuesto por tres partes principales: 1) configuración, 2) código del algoritmo de regresión simbólica no lineal híbrido y 3) graficar los resultados obtenidos. A continuación, se hará una descripción de cada una de las etapas planteadas:

1. **Configuración:** en esta etapa se realizan las actividades de cargue de la base de datos de datos con las observaciones experimentales, ingreso de información de la base de datos (cantidad del total de datos destinados al entrenamiento y a la predicción), configuración de parámetros numéricos de la programación genética y conjuntos de funciones tales como:
    - Número de individuos o cantidad de posibles soluciones dentro de la población.
    - Número de generaciones o cantidad de iteraciones que debe ejecutar el algoritmo.
    - Número máximo de genes por individuo.
    - Cantidad máxima de individuos en la primera generación.

- Cantidad máxima de individuos en cualquier generación diferente a la inicial.
- Probabilidad de cada operación genética.
- Maneras de construir estructura de árboles.
- Configuración de la operación de selección.
- Configuración de conjunto de funciones: definir las operaciones y funciones matemáticas utilizadas para la construcción las estructuras de árboles.

2. Regresión simbólica no lineal híbrida: En esta etapa se realizan las siguientes acciones:
   - Representación dual y configuración de la matriz kernel.
   - Paso 1: Construir la población inicial.
   - Paso 2: Modificar la población en cada nueva generación de manera iterativa.
   - Actividades iterativas: evaluar la puntuación de cada individuo de la población utilizando la configuración de kernel, seleccionar la mejor expresión de cada generación según los puntajes, creación de la nueva población al aplicar las operaciones genéticas según las probabilidades establecidas para estas y muestrear la base de datos ingresada con tal de evitar el sobre entrenamiento.

3. Graficar los resultados: en esta etapa se presenta de manera gráfica los resultados obtenidos para el entrenamiento y de inferencias realizadas por el algoritmo.

Con lo anterior se realiza el cumplimiento del segundo objetivo específico propuesto dado que se realizó el planteamiento del método de regresión simbólica con base al cual se realizan los pronósticos de velocidades de vientos promedio.

3. Implementación del algoritmo de regresión simbólica para pronósticos de viento

En el presente capítulo se realiza inicialmente el planteamiento de los casos de estudio; lo que incluye cuáles son las locaciones específicas que se seleccionaron, la cantidad de dato por locación y las variables meteorológicas contempladas para la elaboración de los pronósticos de viento. Posteriormente, se realizará la presentación las distintas simulaciones realizadas y los resultados de los pronósticos de vientos obtenidos a partir de la implementación del algoritmo de regresión simbólica planteado en el capítulo anterior.

3.1 Planteamiento de los casos de estudio

Como ya se mencionó al finalizar el capítulo 1, para el presente trabajo se seleccionaron cuatro locaciones costeras en diferentes regiones. En dichas locaciones se escogió una estación meteorológica que contase con datos continuos para varios años y que contuviesen diferentes variables meteorológicas con las cuales se pudiese alimentar el modelo de regresión simbólica que se presentó en el capítulo 2. Las locaciones seleccionadas fueron:

- Caso de estudio 1: Isla O‘ahu, en el estado estadounidense de Hawái.
- Caso de estudio 2: Isla Gran Caimán, capital de Islas Caimán.
- Caso de estudio 3: San Andrés, en el archipiélago de Colombia.
- Caso de estudio 4: Talcahuano, Chile.

En [7] se hace referencia a que las variables meteorológica más relevante para realizar pronósticos en la generación de potencia por parques eólicos que además de la velocidad del viento son: temperatura, presión y densidad de aire. Adicionalmente, en [20] se encontró que las variables más analizadas en pronósticos de viento son, además de las variables antes mencionadas, la potencia generada, radiación y precipitación. Por otra parte, en [43] se discute sobre la importancia o no de contemplar la dirección del viento para plantas de generación eólica, concluyendo que de la revisión de referencias realizada en dicho estudio esta variable no tiene influencia significativa en la generación de energía, sin embargo si debe ser considerada para el micro-emplazamiento del parque de generación.

Teniendo en cuenta lo anterior, para el presente trabajo se tuvieron en cuenta los valores máximos, promedios y mínimos por día de las siguientes variables meteorológicas: temperatura, humedad, presión, punto de rocío y velocidad del viento.

Se tomaron datos de promedios diarios de las diferentes variables meteorológicas con las que se contaban en la fuente de datos mencionada en la sección 1.2. Los conjuntos de datos utilizados tienen al menos tres años de datos históricos; una de las ventajas de la programación genética es su flexibilidad ante el caso de que la información con la que se cuente no se encuentre completa [32]. Para los casos de estudio 1 y 3 se cuenta con un conjunto de 1304 días de datos para cada uno, equivalente a aproximadamente tres años y medio de históricos meteorológicos consecutivos; por otra parte, para el caso de estudio 2 se cuenta con datos de 1155 días, equivalente a tres años y dos meses de históricos consecutivos; por otra parte, para el caso de estudio 4 se cuenta con 749 días de información repartidos en tres años, por lo cual este conjunto de datos contiene grandes vacíos de información histórica. A nivel general, los datos utilizados van desde el año 2017 hasta los primeros meses del año 2020. En la **Tabla 3.1** se presenta un resumen de lo anterior.

Tabla 3.1: Cantidad de datos por caso de estudio

Año	Cantidad de datos por caso de estudio			
	Caso 1	Caso 2	Caso 3	Caso 4
1: 2017	365	365	365	184
2: 2017-2018	730	730	730	383
3: 2017-2020	1304	1155	1304	749

3.2 Resultados obtenidos de los casos de estudio

Teniendo en cuenta que el tercer objetivo específico propuesto para el presente trabajo contempla la decisión de un número apropiado de datos de entrenamiento y prueba para el algoritmo propuesto, el primer paso previo a la implementación del algoritmo fue dividir los conjuntos de datos de cada ubicación y agruparlos variando la cantidad de datos para evaluar el funcionamiento del algoritmo en varios periodos de tiempo y para cada subgrupo de datos se realizó la variación de los datos destinados al entrenamiento y a la validación del algoritmo.

Si bien la segmentación de los datos por periodos de tiempo está fuera del alcance de los objetivos propuestos de la tesis, realizar este ejercicio da la posibilidad de evaluar el desempeño del algoritmo en la elaboración de pronósticos de vientos en un horizonte de aproximadamente tres años y de esta manera determinar si puede ser utilizado en la planeación y diseño de nuevas plantas eólicas y para realizar la programación de mantenimientos a mediano y largo plazo.

En este orden de ideas, se evaluó el desempeño del algoritmo para cada ubicación en conjuntos de datos de 1, 2 y 3 años. Luego, para cada conjunto de datos se varió la cantidad de datos ingresados para el entrenamiento y prueba. En la revisión del estado del arte realizada en [20] se encontró que más del 80% de los estudios realizados en pronósticos de viento toman más del 60% de los datos para entrenamiento de los diferentes algoritmos utilizados; teniendo en cuenta lo anterior, para el presente trabajo se

realizaron pruebas tomando como punto de inicio tomar el 30% de los datos de cada caso para el entrenamiento del algoritmo.

3.2.1 Resultados caso de estudio 1

- **Base de datos de un año**

A continuación, en la **Tabla 3.2** se presenta el consolidado de los resultados de los pronósticos realizados teniendo en cuenta una base de datos de un año, en el cual se varía la cantidad de datos utilizados para el entrenamiento y la validación de la inferencia obtenida por medio del algoritmo de regresión.

Tabla 3.2: Resultados de pronósticos del caso de estudio 1 para un año.

Ubicación 1							
Años	**Cantidad de Datos**	**% de datos para Entrenamiento**	**% de datos para Inferencia**	**Cantidad de Datos Entrenamiento**	**Cantidad de Datos Inferencia**	**Error Entrenamiento**	**Error Inferencia**
1	365	95%	5%	347	18	8,15%	10,07%
		90%	10%	329	36	8,12%	9,48%
		80%	20%	292	73	8,04%	9,08%
		70%	30%	256	109	8,10%	8,66%
		50%	50%	183	182	8,52%	8,33%
		30%	70%	110	255	8,82%	9,04%

En las **Figura 3-1**, **Figura 3-2** y **Figura 3-3** se presentan las gráficas obtenidas donde se puede observar cómo es el comportamiento de la inferencia obtenida contra los valores reales ingresados al algoritmo para los casos donde los porcentajes de datos destinados para el entrenamiento del algoritmo son del 90%, 50% y 30%, respectivamente.

Figura 3-1: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de un año de datos con un porcentaje de datos de entrenamiento del 90%.

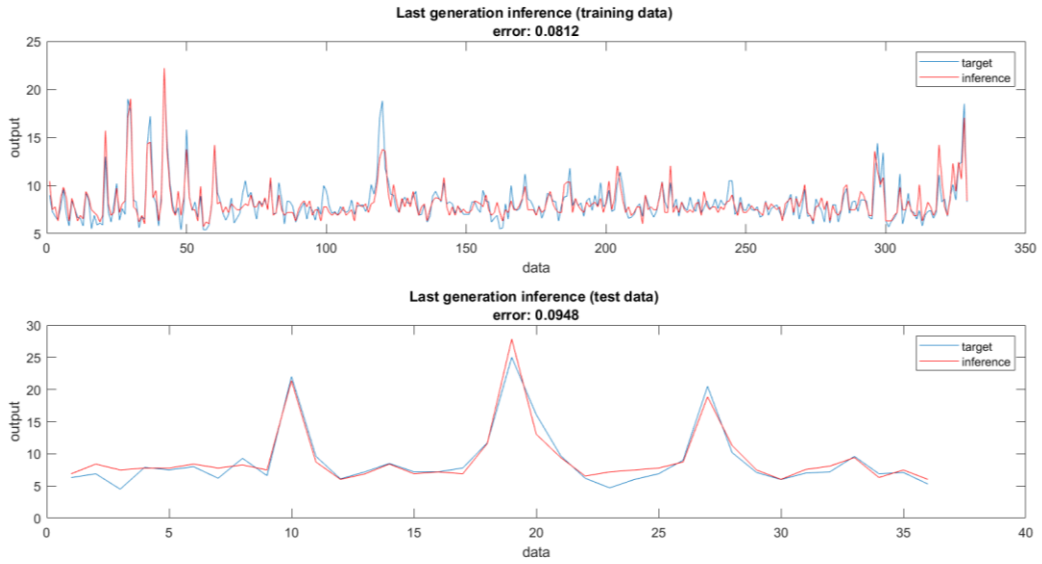
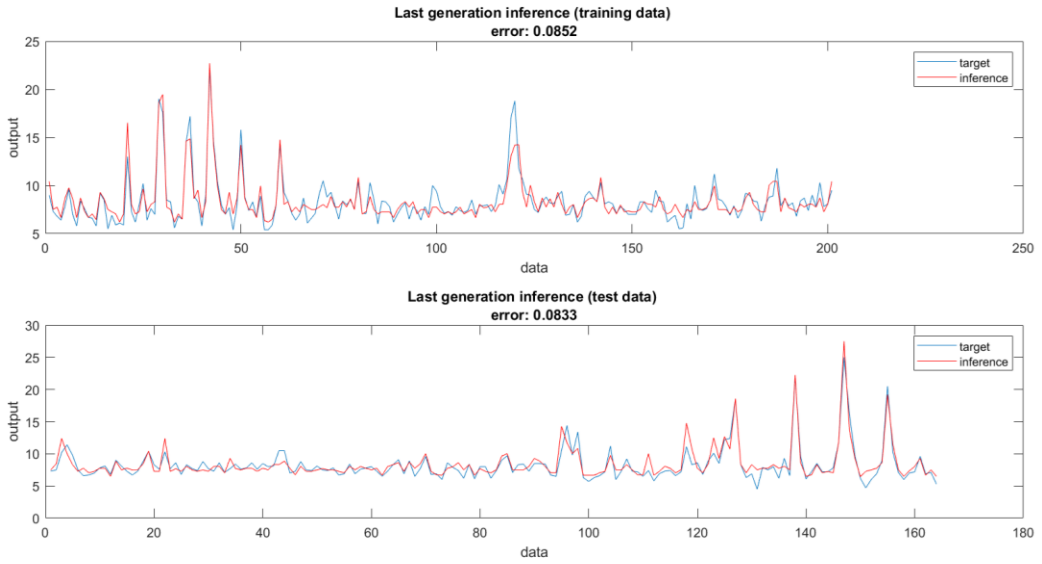
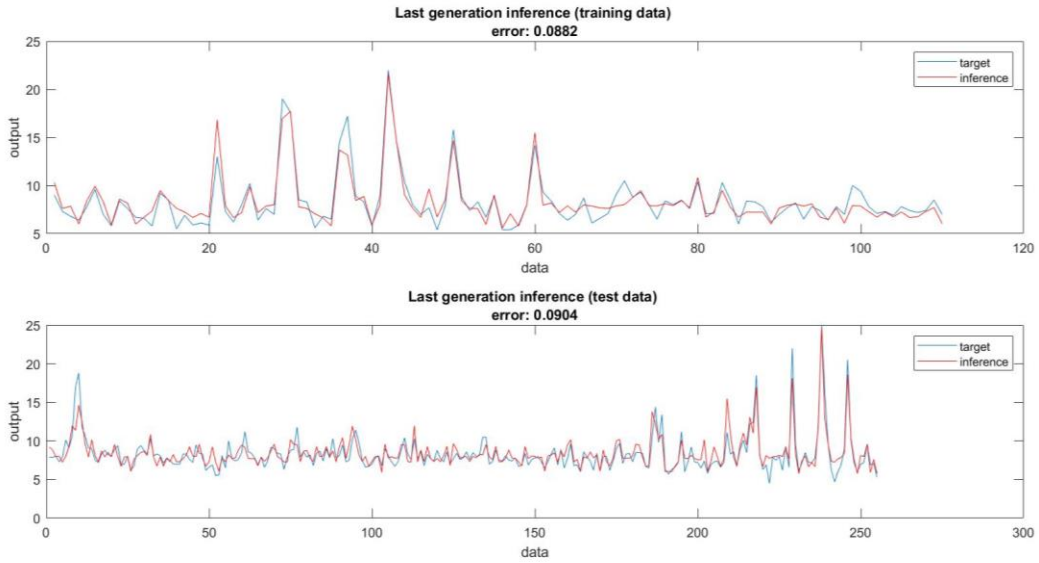


Figura 3-2: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de un año de datos con un porcentaje de datos de entrenamiento del 50%.



Nótese que para este para el caso de estar evaluando un año en el caso más extremo que es donde sólo se toma el 30% de los datos con los que se cuentan para el entrenamiento del algoritmo, se requieren 110 datos, equivalentes aproximadamente tres meses de datos históricos y con esto se puede hacer la proyección ocho meses.

Figura 3-3: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de un año de datos con un porcentaje de datos de entrenamiento del 30%.



En la **¡Error! No se encuentra el origen de la referencia.** se presentan los resultados de los pronósticos de viento realizados para los conjuntos de datos de la locación 2 variando las cantidades de datos para entrenamiento y prueba. Por su parte, en la **¡Error! No se encuentra el origen de la referencia.** se presenta el consolidado de las diferentes pruebas realizadas para cada conjunto de datos por año variando los valores de entrenamiento y prueba, relacionando los errores resultantes de cada uno.

- **Base de datos de dos años**

A continuación, en la **Tabla 3.3** se presenta el consolidado de los resultados de los pronósticos realizados teniendo en cuenta una base de datos de dos años, en el cual se varía la cantidad de datos utilizados para el entrenamiento y la validación de la inferencia obtenida por medio del algoritmo de regresión.

Tabla 3.3: Resultados de pronósticos del caso de estudio 1 para dos años.

Ubicación 1							
Años	**Cantidad de Datos**	**% de datos para Entrenamiento**	**% de datos para Inferencia**	**Cantidad de Datos Entrenamiento**	**Cantidad de Datos Inferencia**	**Error Entrenamiento**	**Error Inferencia**
2	730	95%	5%	694	36	8,81%	9,27%
		90%	10%	657	73	8,94%	8,64%
		80%	20%	584	146	9,79%	9,21%
		70%	30%	511	219	9,08%	8,99%
		50%	60%	292	438	8,61%	9,42%
		30%	70%	219	511	8,44%	9,41%

En las **Figura 3-4**, **Figura 3-5** y **Figura 3-6** se presentan las gráficas obtenidas donde se puede observar cómo es el comportamiento de la inferencia obtenida contra los valores reales ingresados al algoritmo para los casos donde los porcentajes de datos destinados para el entrenamiento del algoritmo son del 90%, 50% y 30%, respectivamente.

Figura 3-4: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de dos años de datos con un porcentaje de datos de entrenamiento del 90%.

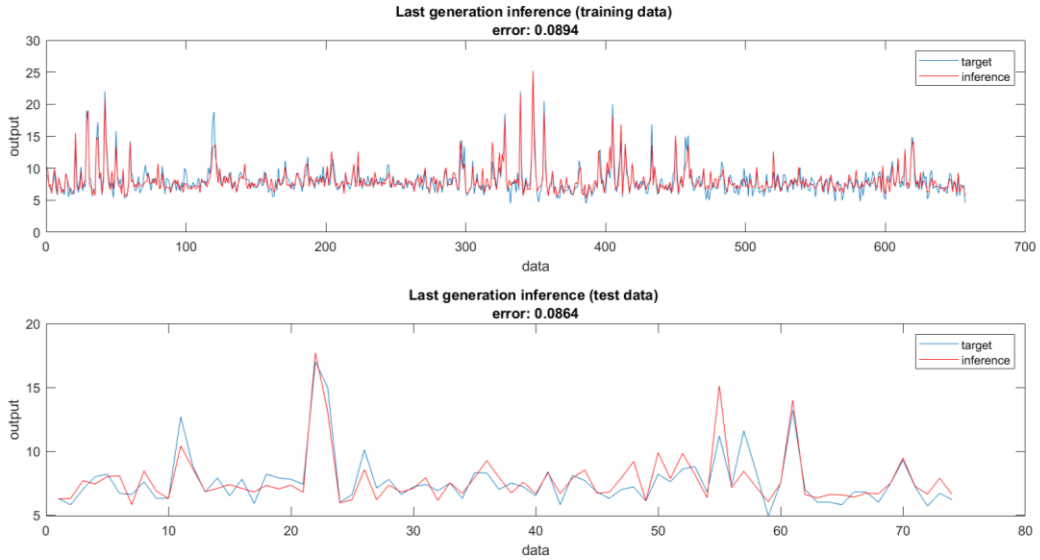


Figura 3-5: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de dos años de datos con un porcentaje de datos de entrenamiento del 50%.

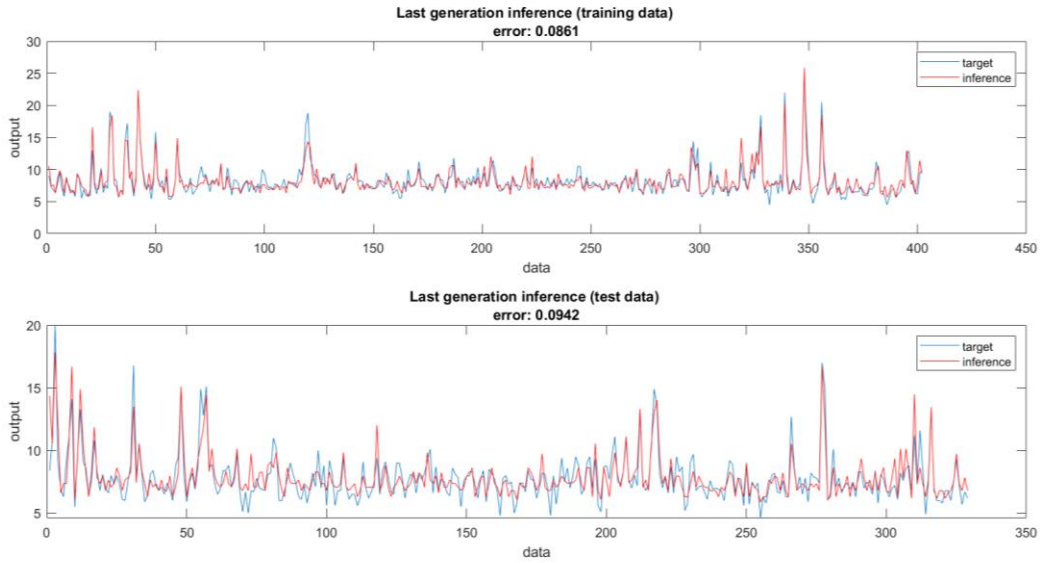
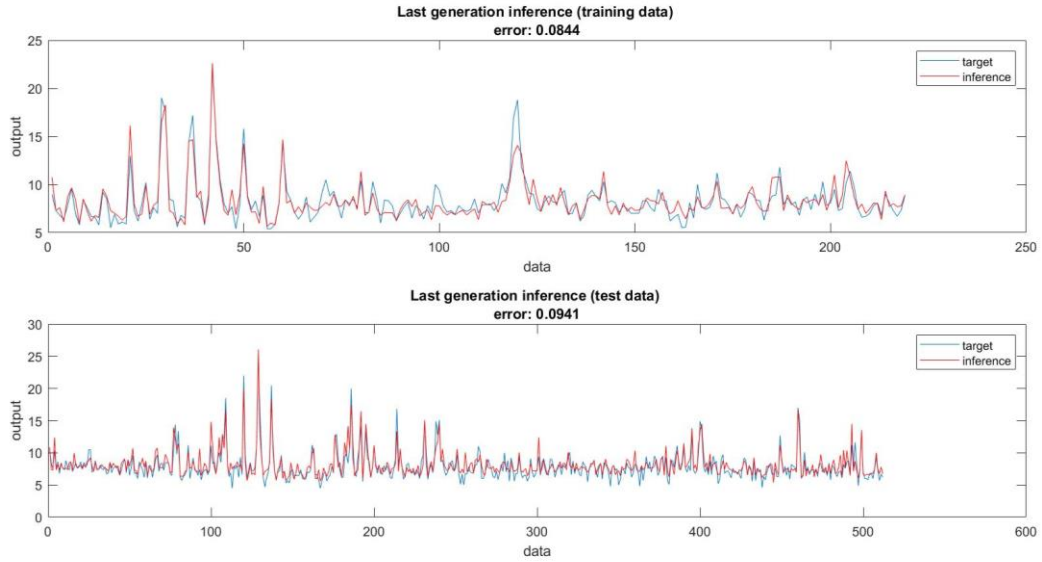


Figura 3-6: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de dos años de datos con un porcentaje de datos de entrenamiento del 30%.



- **Base de datos de tres años**

A continuación, en la **Tabla 3.4** se presenta el consolidado de los resultados de los pronósticos realizados teniendo en cuenta una base de datos de tres años, en el cual se varía la cantidad de datos utilizados para el entrenamiento y la validación de la inferencia obtenida por medio del algoritmo de regresión.

Tabla 3.4: Resultados de pronósticos del caso de estudio 1 para tres años.

Ubicación 1							
Años	**Cantidad de Datos**	**% de datos para Entrenamiento**	**% de datos para Inferencia**	**Cantidad de Datos Entrenamiento**	**Cantidad de Datos Inferencia**	**Error Entrenamiento**	**Error Inferencia**
3	1304	95%	5%	1239	65	8,74%	8,33%
		90%	10%	1174	130	8,74%	8,03%
		80%	20%	1043	260	8,89%	7,88%
		70%	30%	913	391	10,50%	9,61%
		56%	44%	730	574	8,87%	8,44%
		30%	70%	391	913	10,65%	10,98%

En las **Figura 3-7**, **Figura 3-8** y **Figura 3-9** se presentan las gráficas obtenidas donde se puede observar cómo es el comportamiento de la inferencia obtenida contra los valores reales ingresados al algoritmo para los casos donde los porcentajes de datos destinados para el entrenamiento del algoritmo son del 90%, 50% y 30%, respectivamente.

Figura 3-7: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de tres años de datos con un porcentaje de datos de entrenamiento del 90%.

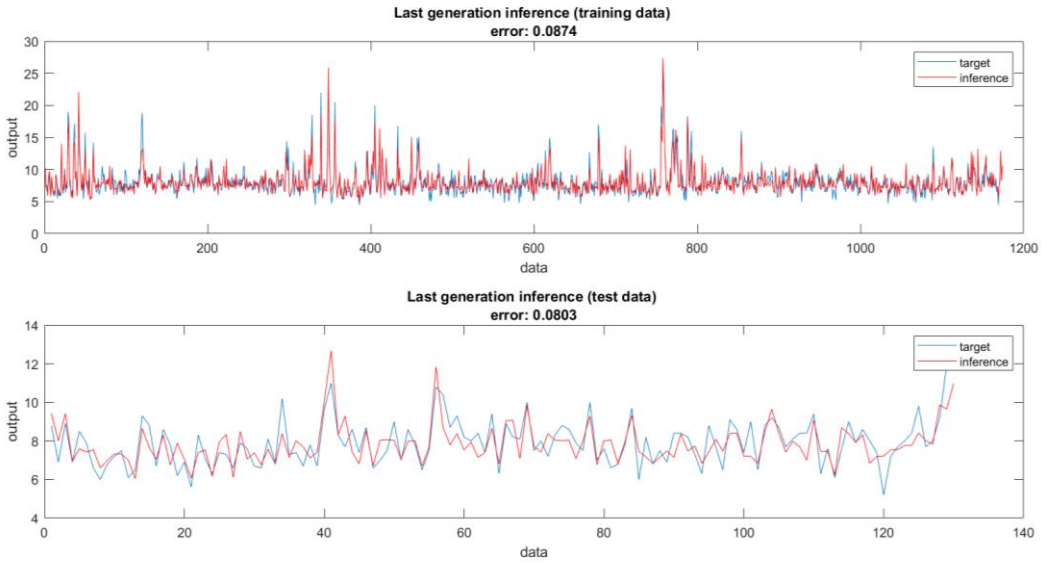


Figura 3-8: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de tres años de datos con un porcentaje de datos de entrenamiento del 50%.

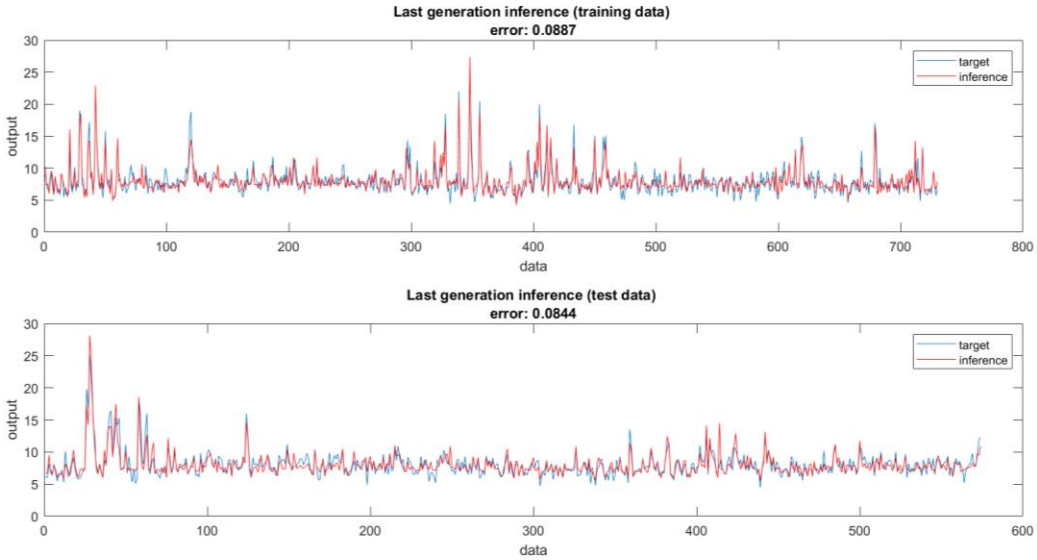
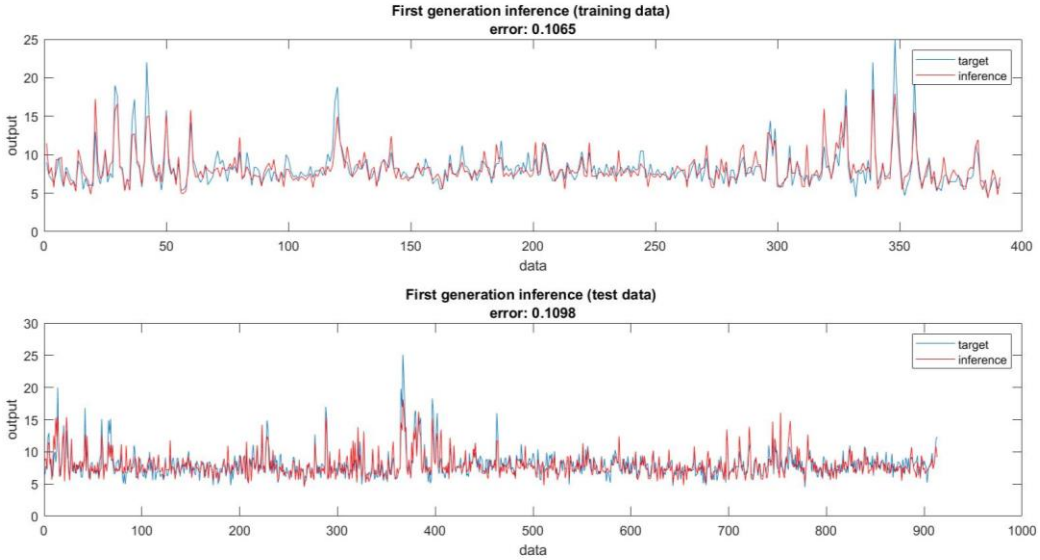


Figura 3-9: Pronósticos de velocidades de viento para el caso de estudio 1 para el caso de tres años de datos con un porcentaje de datos de entrenamiento del 30%.



3.2.2 Resultados caso de estudio 2

- **Base de datos de un año**

A continuación, en la **Tabla 3.5** se presenta el consolidado de los resultados de los pronósticos realizados teniendo en cuenta una base de datos de un año, en el cual se varía la cantidad de datos utilizados para el entrenamiento y la validación de la inferencia obtenida por medio del algoritmo de regresión.

Tabla 3.5: Resultados de pronósticos del caso de estudio 2 para un año.

Ubicación 2							
Años	**Cantidad de Datos**	**% de datos para Entrenamiento**	**% de datos para Inferencia**	**Cantidad de Datos Entrenamiento**	**Cantidad de Datos Inferencia**	**Error Entrenamiento**	**Error Inferencia**
1	365	95%	5%	347	18	10,97%	8,91%
		90%	10%	329	36	7,50%	7,22%
		80%	20%	292	73	8,01%	7,74%
		70%	30%	256	109	7,15%	8,42%
		50%	50%	183	182	6,41%	8,74%
		30%	70%	110	255	7,60%	9,22%

En las **Figura 3-10**, **Figura 3-11** y **Figura 3-12** se presentan las gráficas obtenidas donde se puede observar cómo es el comportamiento de la inferencia obtenida contra los valores reales ingresados al algoritmo para los casos donde los porcentajes de datos destinados para el entrenamiento del algoritmo son del 90%, 50% y 30%, respectivamente.

Figura 3-10: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de un año de datos con un porcentaje de datos de entrenamiento del 90%.

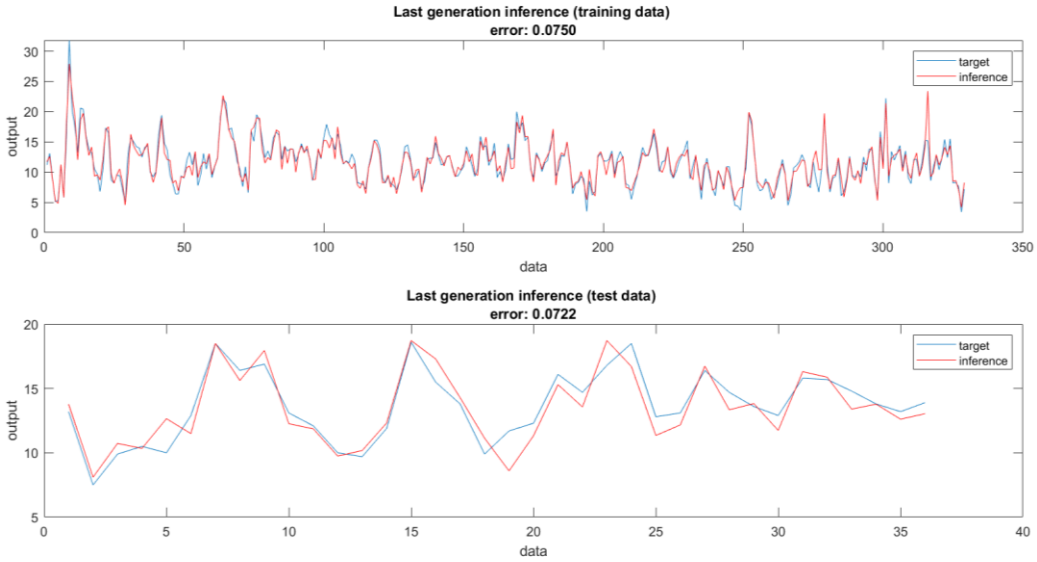


Figura 3-11: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de un año de datos con un porcentaje de datos de entrenamiento del 50%.

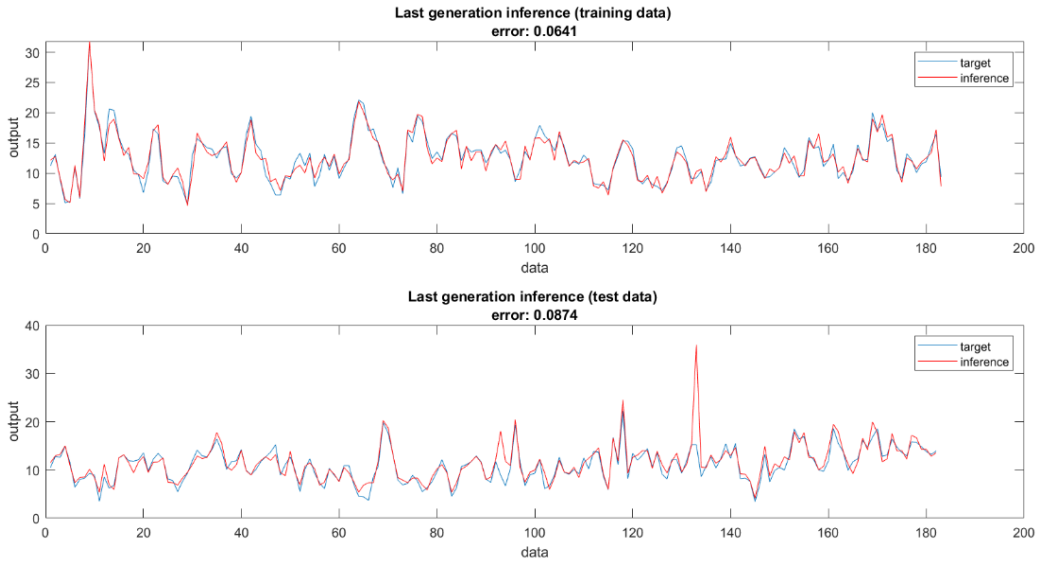
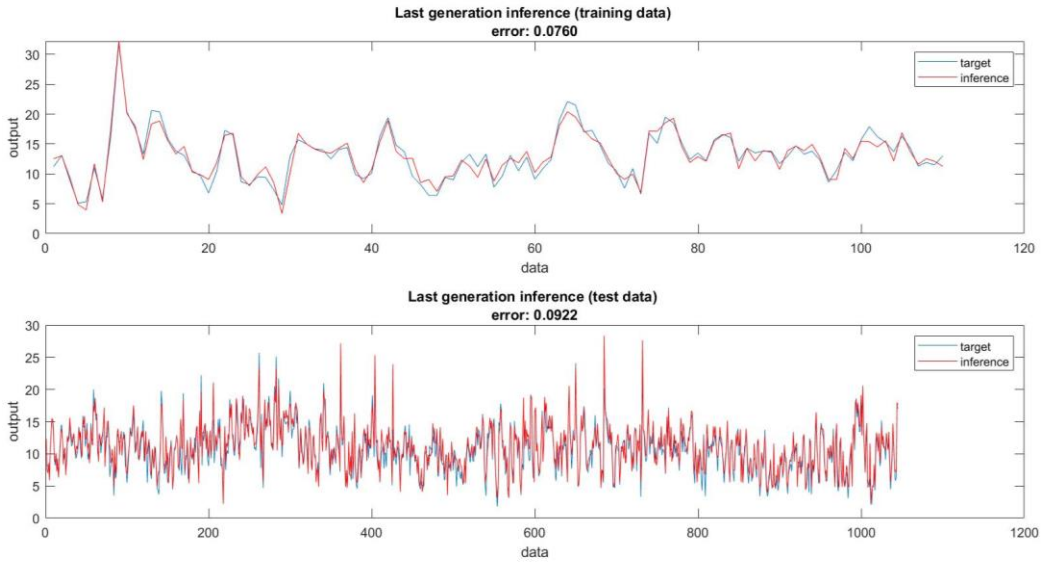


Figura 3-12: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de un año de datos con un porcentaje de datos de entrenamiento del 30%.



- **Base de datos de dos años**

A continuación, en la **Tabla 3.6** se presenta el consolidado de los resultados de los pronósticos realizados teniendo en cuenta una base de datos de dos años, en el cual se

varía la cantidad de datos utilizados para el entrenamiento y la validación de la inferencia obtenida por medio del algoritmo de regresión.

Tabla 3.6: Resultados de pronósticos del caso de estudio 2 para dos años.

Ubicación 2							
Años	**Cantidad de Datos**	**% de datos para Entrenamiento**	**% de datos para Inferencia**	**Cantidad de Datos Entrenamiento**	**Cantidad de Datos Inferencia**	**Error Entrenamiento**	**Error Inferencia**
2	730	95%	5%	694	36	8,49%	8,33%
		90%	10%	657	73	14,08%	15,53%
		80%	20%	584	146	8,86%	10,77%
		70%	30%	511	219	7,42%	13,96%
		50%	50%	365	365	7,67%	10,76%
		30%	70%	219	511	6,86%	9,14%

En las **Figura 3-13**, **Figura 3-14** y **Figura 3-15** se presentan las gráficas obtenidas donde se puede observar cómo es el comportamiento de la inferencia obtenida contra los valores reales ingresados al algoritmo para los casos donde los porcentajes de datos destinados para el entrenamiento del algoritmo son del 90%, 50% y 30%, respectivamente.

Figura 3-13: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de dos años de datos con un porcentaje de datos de entrenamiento del 90%.

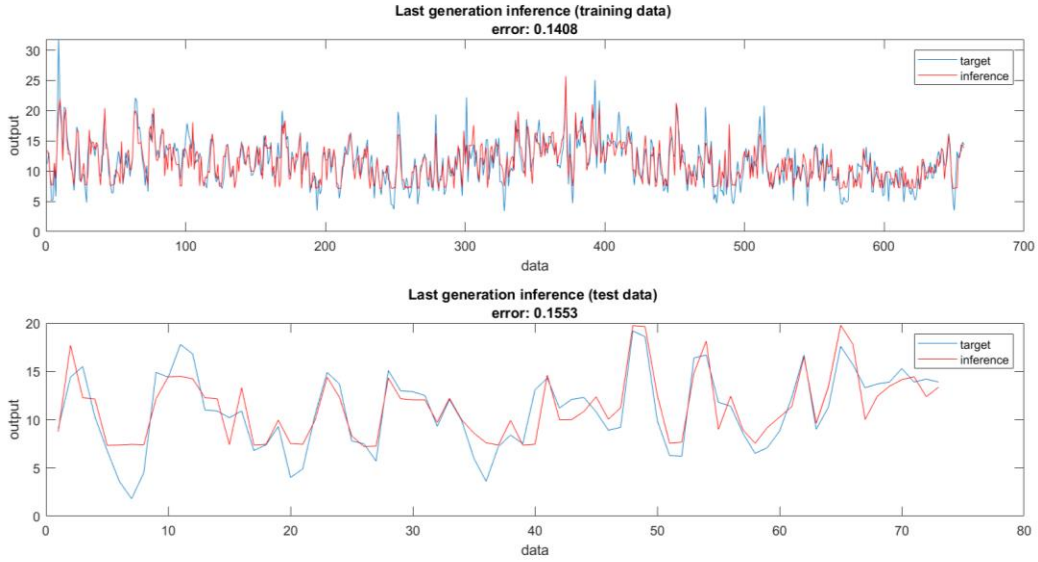


Figura 3-14: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de dos años de datos con un porcentaje de datos de entrenamiento del 50%.

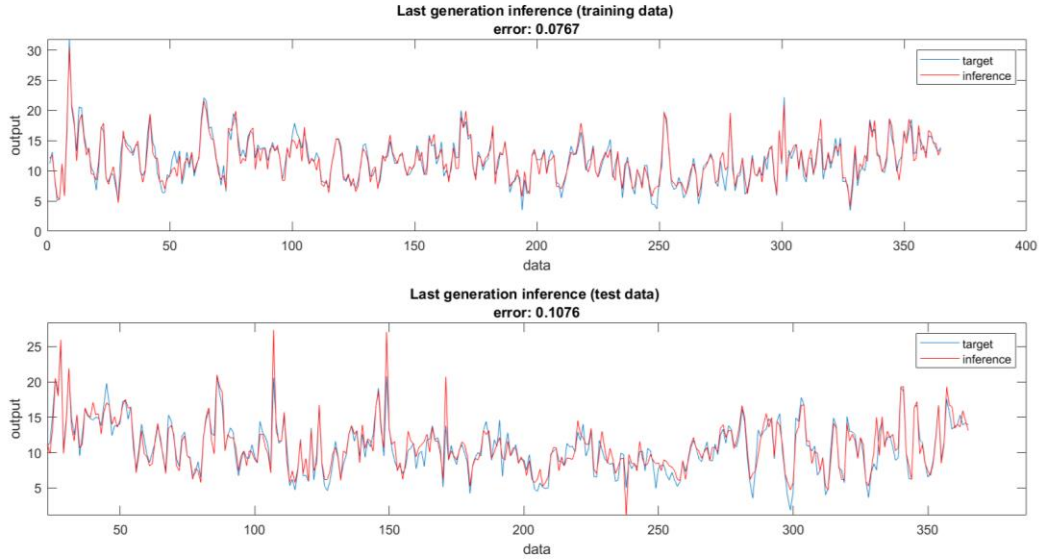
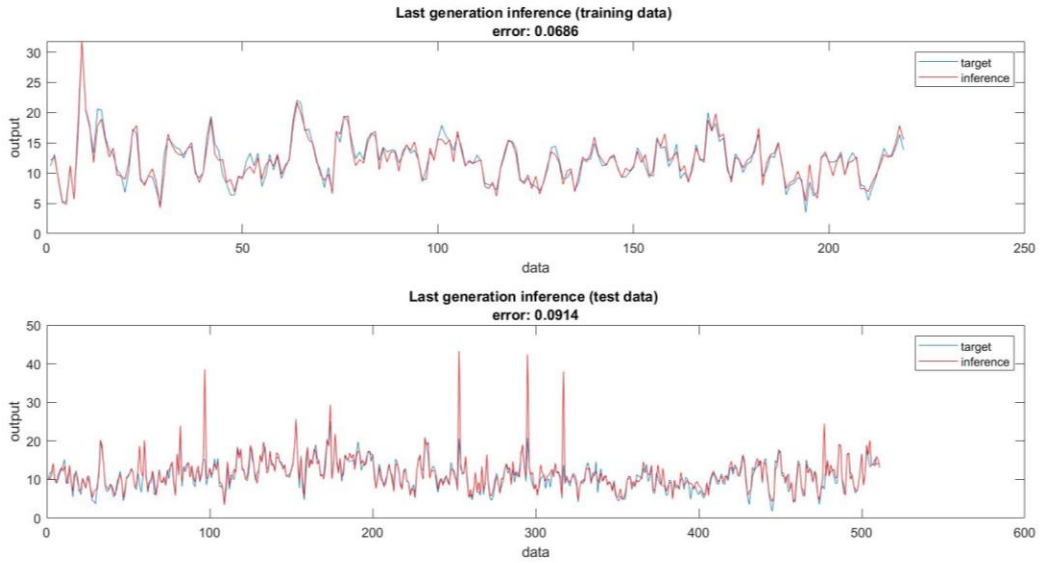


Figura 3-15: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de dos años de datos con un porcentaje de datos de entrenamiento del 30%.



- **Base de datos de tres años**

A continuación, en la

Tabla 3.7 se presenta el consolidado de los resultados de los pronósticos realizados teniendo en cuenta una base de datos de tres años, en el cual se varía la cantidad de datos utilizados para el entrenamiento y la validación de la inferencia obtenida por medio del algoritmo de regresión.

Tabla 3.7: Resultados de pronósticos del caso de estudio 2 para tres años.

Ubicación 2							
Años	**Cantidad de Datos**	**% de datos para Entrenamiento**	**% de datos para Inferencia**	**Cantidad de Datos Entrenamiento**	**Cantidad de Datos Inferencia**	**Error Entrenamiento**	**Error Inferencia**
3	1155	95%	5%	1097	58	8,63%	8,61%
		90%	10%	1040	115	14,79%	16,38%
		80%	20%	924	231	8,04%	8,91%
		70%	30%	809	346	8,58%	10,36%
		50%	50%	578	577	7,97%	10,06%
		30%	70%	347	808	7,59%	9,43%

En las **Figura 3-16**, **Figura 3-17** y **Figura 3-18** se presentan las gráficas obtenidas donde se puede observar cómo es el comportamiento de la inferencia obtenida contra los valores reales ingresados al algoritmo para los casos donde los porcentajes de datos destinados para el entrenamiento del algoritmo son del 90%, 50% y 30%, respectivamente.

Figura 3-16: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de tres años de datos con un porcentaje de datos de entrenamiento del 90%.

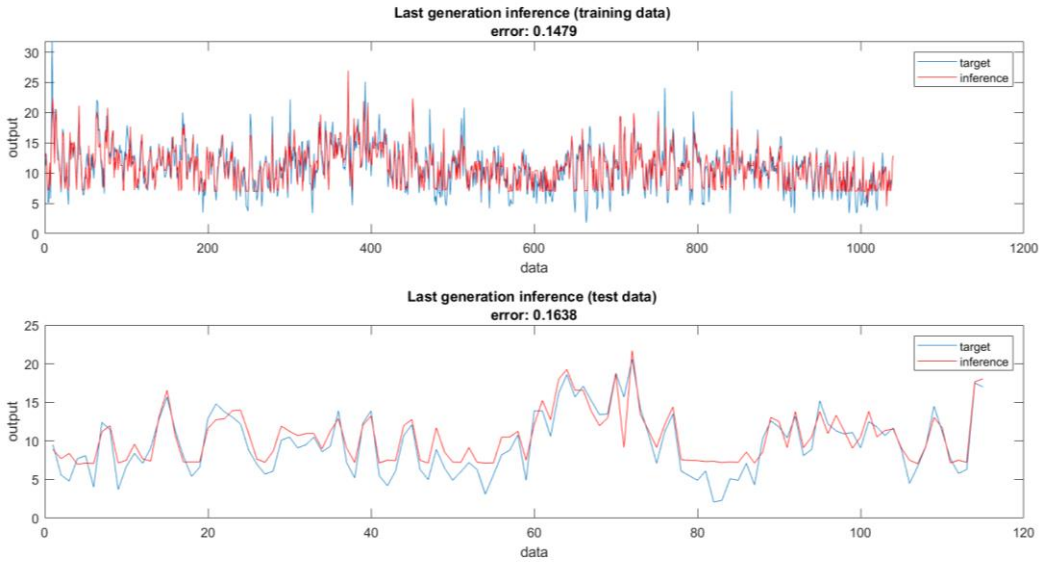


Figura 3-17: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de tres años de datos con un porcentaje de datos de entrenamiento del 50%.

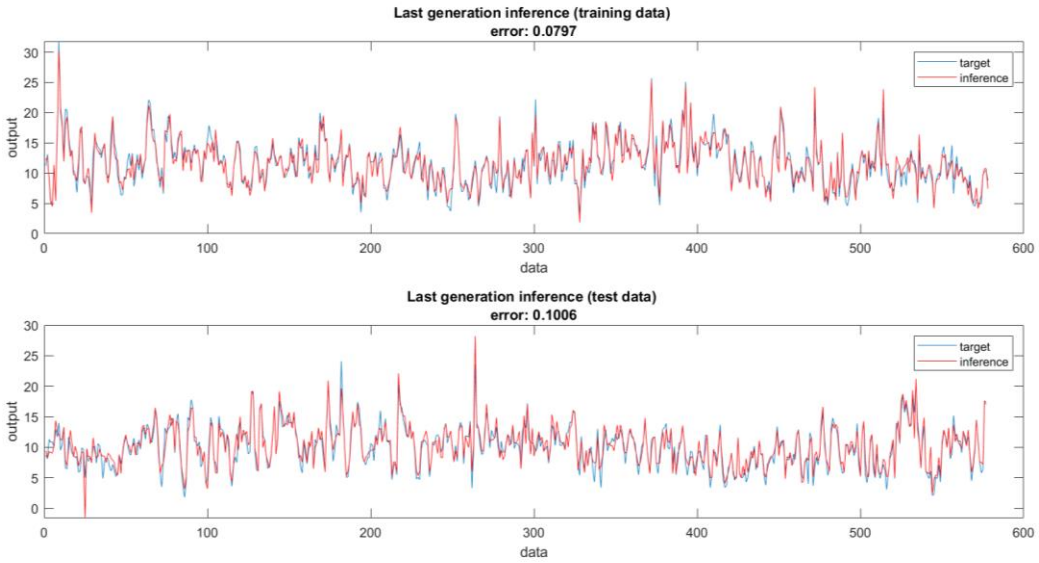
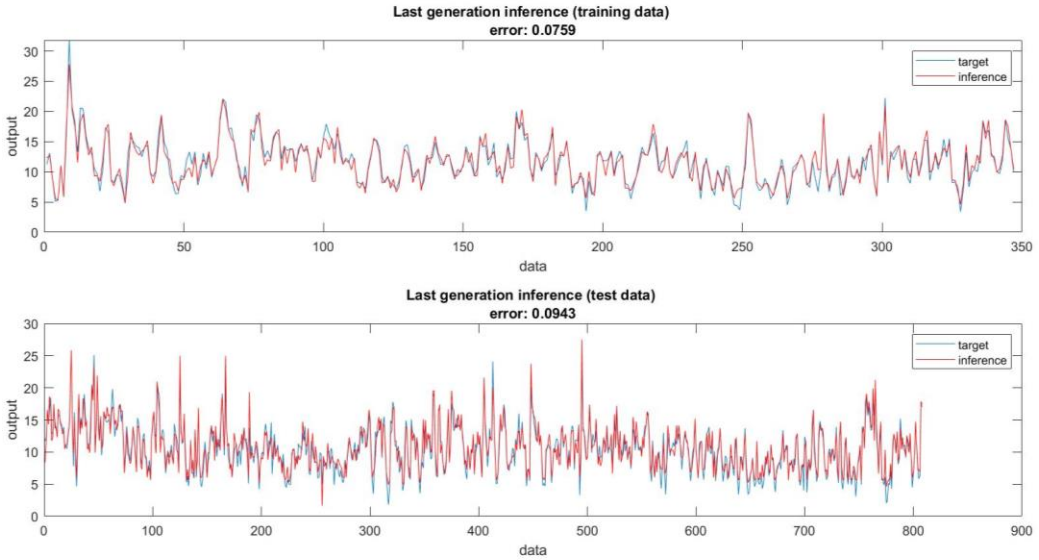


Figura 3-18: Pronósticos de velocidades de viento para el caso de estudio 2 para el caso de tres años de datos con un porcentaje de datos de entrenamiento del 30%.



3.2.3 Resultados caso de estudio 3

- **Base de datos de un año**

A continuación, en la **Tabla 3.8** se presenta el consolidado de los resultados de los pronósticos realizados teniendo en cuenta una base de datos de un año, en el cual se varía la cantidad de datos utilizados para el entrenamiento y la validación de la inferencia obtenida por medio del algoritmo de regresión.

Tabla 3.8: Resultados de pronósticos del caso de estudio 3 para un año.

Ubicación 3							
Años	**Cantidad de Datos**	**% de datos para Entrenamiento**	**% de datos para Inferencia**	**Cantidad de Datos Entrenamiento**	**Cantidad de Datos Inferencia**	**Error Entrenamiento**	**Error Inferencia**
1	365	95%	5%	347	18	20,07%	11,40%
		90%	10%	329	36	14,11%	14,47%
		80%	20%	292	73	13,80%	16,63%
		70%	30%	256	109	12,57%	17,55%
		50%	50%	183	182	14,21%	18,77%
		30%	70%	110	255	9,14%	18,71%

En las **Figura 3-19**, **Figura 3-20** y **Figura 3-21** se presentan las gráficas obtenidas donde se puede observar cómo es el comportamiento de la inferencia obtenida contra los valores reales ingresados al algoritmo para los casos donde los porcentajes de datos destinados para el entrenamiento del algoritmo son del 90%, 50% y 30%, respectivamente.

Figura 3-19: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de un año de datos con un porcentaje de datos de entrenamiento del 90%.

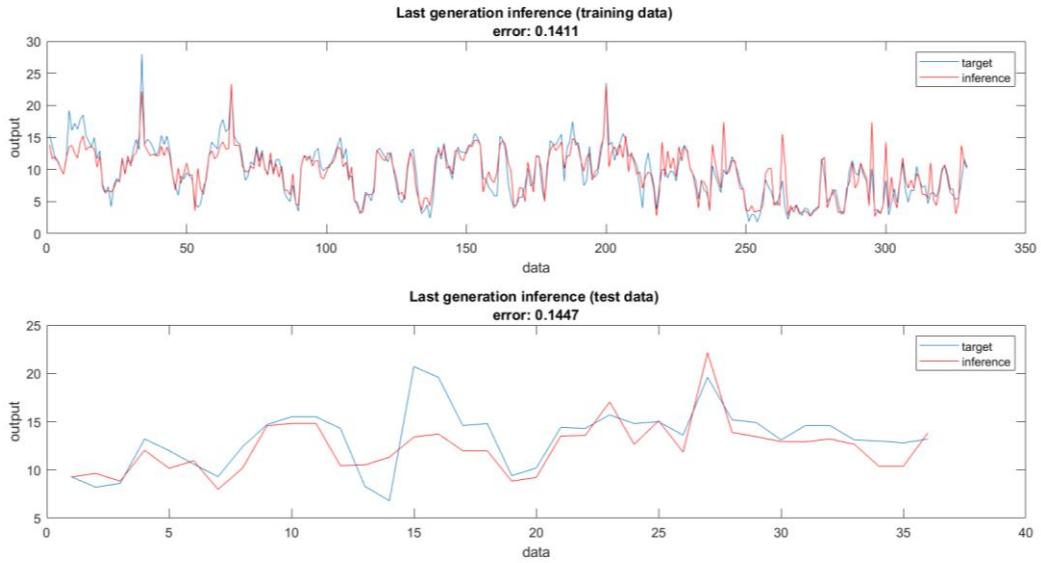


Figura 3-20: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de un año de datos con un porcentaje de datos de entrenamiento del 50%.

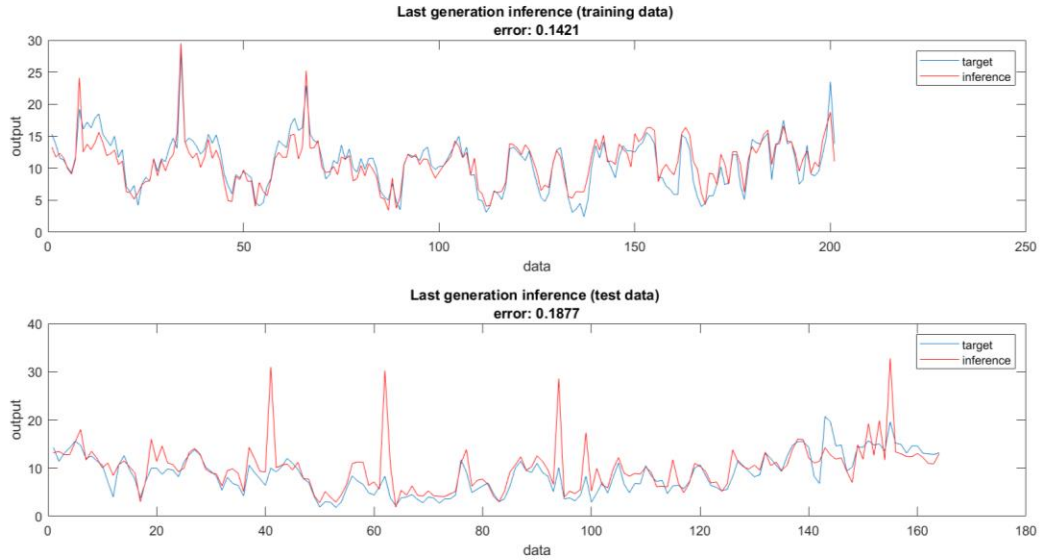
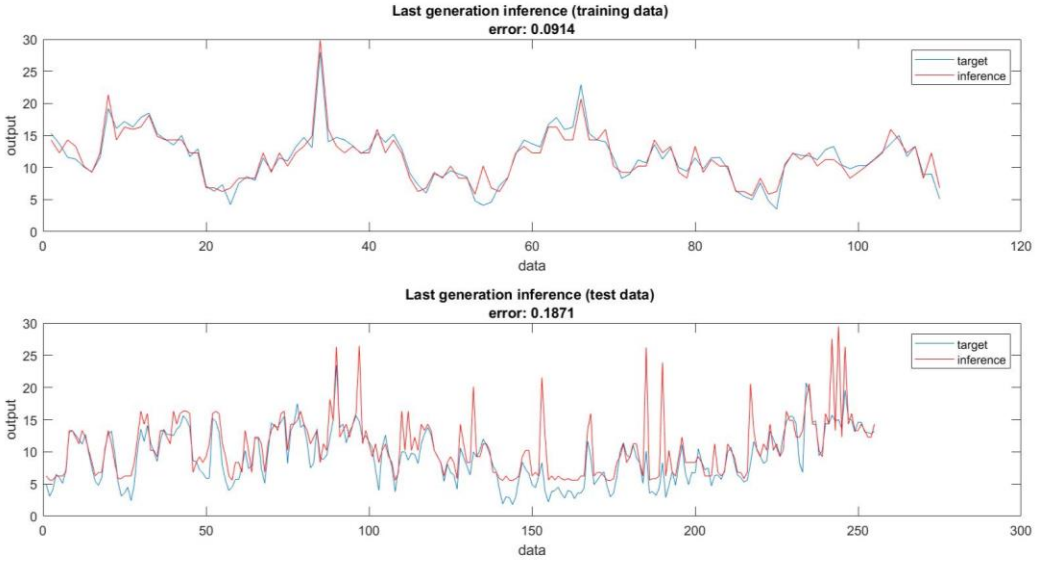


Figura 3-21: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de un año de datos con un porcentaje de datos de entrenamiento del 30%.



- **Base de datos de dos años**

A continuación, en la **Tabla 3.9** se presenta el consolidado de los resultados de los pronósticos realizados teniendo en cuenta una base de datos de dos años, en el cual se varía la cantidad de datos utilizados para el entrenamiento y la validación de la inferencia obtenida por medio del algoritmo de regresión.

Tabla 3.9: Resultados de pronósticos del caso de estudio 3 para dos años.

Ubicación 3							
Años	**Cantidad de Datos**	**% de datos para Entrenamiento**	**% de datos para Inferencia**	**Cantidad de Datos Entrenamiento**	**Cantidad de Datos Inferencia**	**Error Entrenamiento**	**Error Inferencia**
2	730	95%	5%	694	36	11,84%	8,05%
		90%	10%	657	73	13,10%	8,97%
		80%	20%	584	146	12,49%	12,15%
		70%	30%	511	219	11,96%	11,31%
		50%	50%	365	365	13,93%	11,34%
		30%	70%	219	511	11,58%	14,11%

En las **Figura 3-22**, **Figura 3-23** y **Figura 3-24** se presentan las gráficas obtenidas donde se puede observar cómo es el comportamiento de la inferencia obtenida contra los valores

reales ingresados al algoritmo para los casos donde los porcentajes de datos destinados para el entrenamiento del algoritmo son del 90%, 50% y 30%, respectivamente.

Figura 3-22: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de dos años de datos con un porcentaje de datos de entrenamiento del 90%.

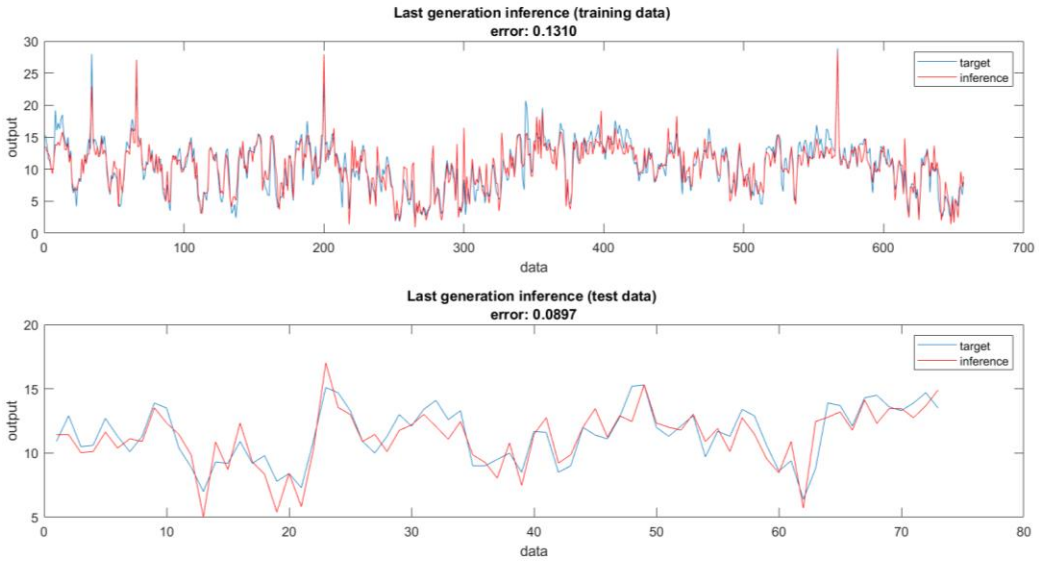


Figura 3-23: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de dos años de datos con un porcentaje de datos de entrenamiento del 50%.

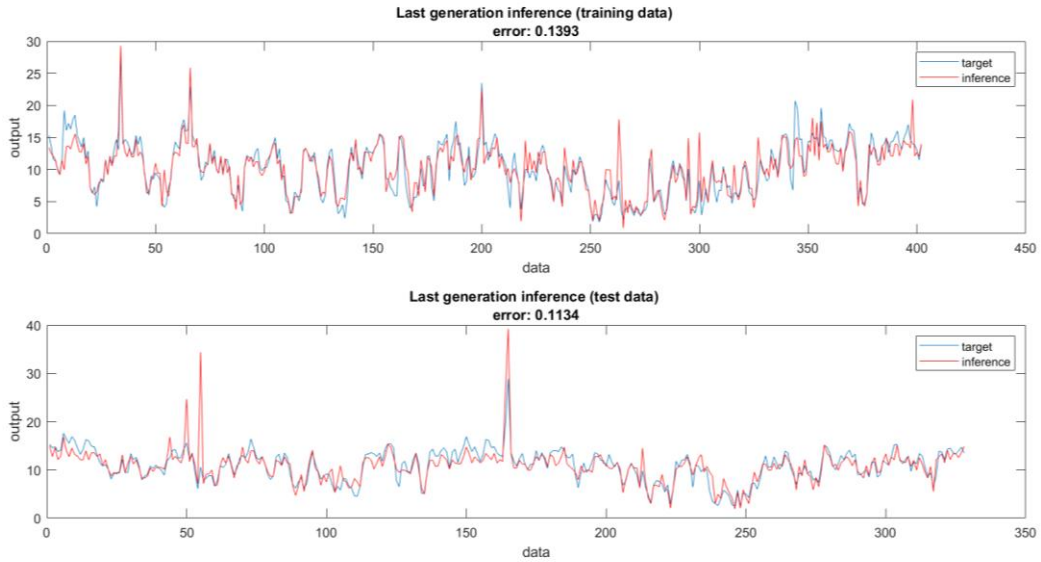
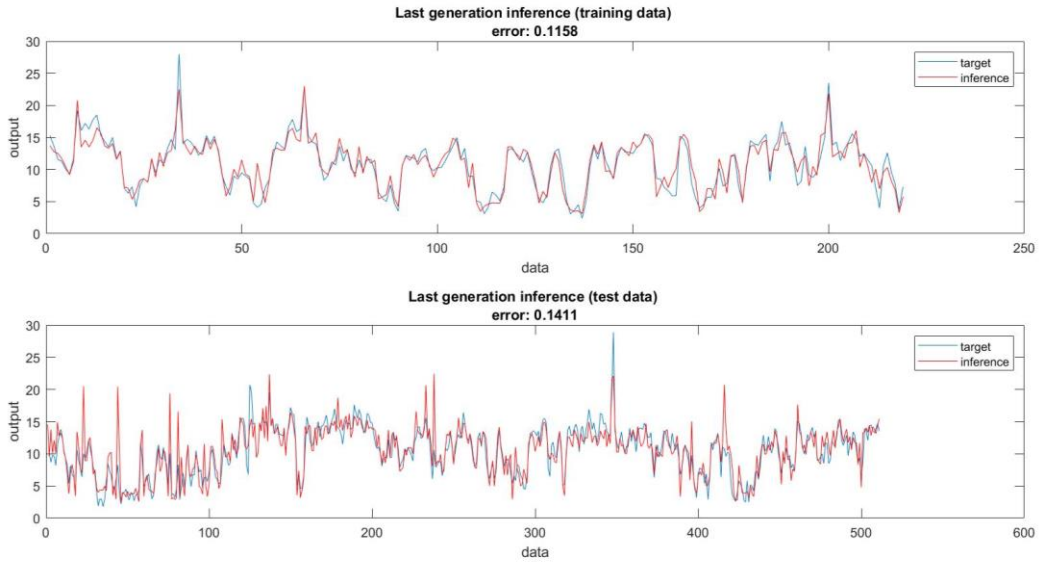


Figura 3-24: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de dos años de datos con un porcentaje de datos de entrenamiento del 30%.



- **Base de datos de tres años**

A continuación, en la **Tabla 3.10** se presenta el consolidado de los resultados de los pronósticos realizados teniendo en cuenta una base de datos de tres años, en el cual se varía la cantidad de datos utilizados para el entrenamiento y la validación de la inferencia obtenida por medio del algoritmo de regresión.

Tabla 3.10: Resultados de pronósticos del caso de estudio 3 para tres años.

Ubicación 3							
Años	**Cantidad de Datos**	**% de datos para Entrenamiento**	**% de datos para Inferencia**	**Cantidad de Datos Entrenamiento**	**Cantidad de Datos Inferencia**	**Error Entrenamiento**	**Error Inferencia**
3	1304	95%	5%	1239	65	12,44%	10,83%
		90%	10%	1174	130	11,23%	9,45%
		80%	20%	1043	261	11,47%	9,23%
		70%	30%	913	391	13,52%	14,83%
		50%	50%	652	652	11,56%	9,71%
		30%	70%	391	913	13,95%	12,11%

En las **Figura 3-25**, **Figura 3-26** y **Figura 3-27** se presentan las gráficas obtenidas donde se puede observar cómo es el comportamiento de la inferencia obtenida contra los valores

reales ingresados al algoritmo para los casos donde los porcentajes de datos destinados para el entrenamiento del algoritmo son del 90%, 50% y 30%, respectivamente.

Figura 3-25: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de tres años de datos con un porcentaje de datos de entrenamiento del 90%.

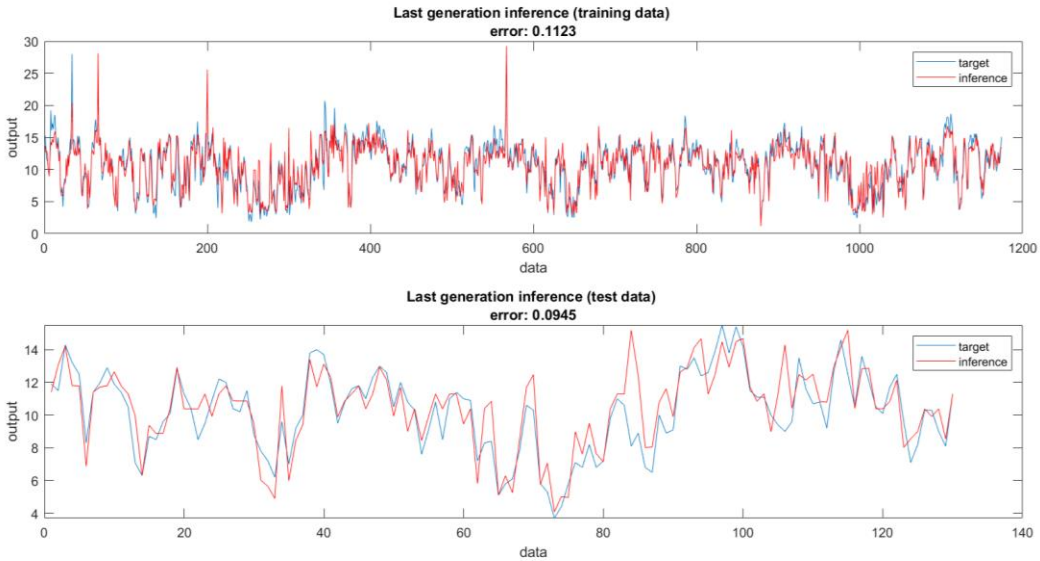


Figura 3-26: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de tres años de datos con un porcentaje de datos de entrenamiento del 50%.

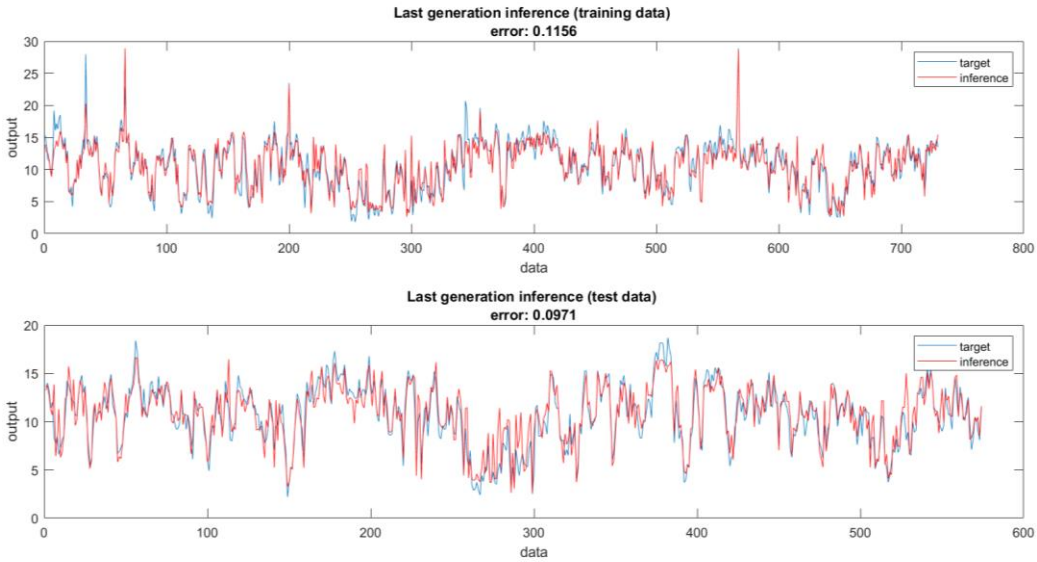
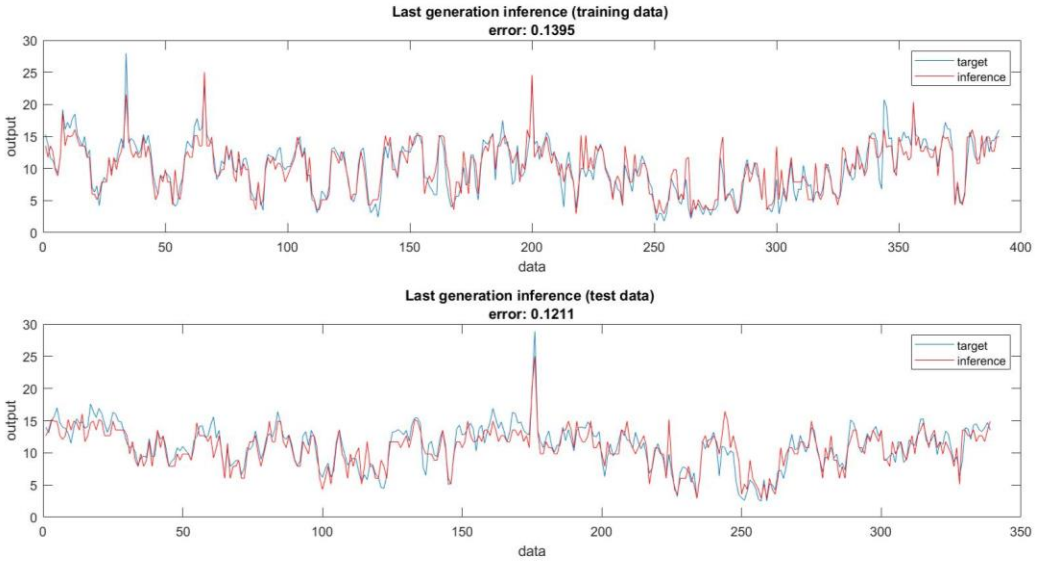


Figura 3-27: Pronósticos de velocidades de viento para el caso de estudio 3 para el escenario de tres años de datos con un porcentaje de datos de entrenamiento del 30%.



3.2.4 Resultados caso de estudio 4

- **Base de datos de un año**

A continuación, en la **Tabla 3.11** se presenta el consolidado de los resultados de los pronósticos realizados teniendo en cuenta una base de datos de un año, en el cual se varía la cantidad de datos utilizados para el entrenamiento y la validación de la inferencia obtenida por medio del algoritmo de regresión.

Tabla 3.11: Resultados de pronósticos del caso de estudio 4 para un año.

Ubicación 4							
Años	**Cantidad de Datos**	**% de datos para Entrenamiento**	**% de datos para Inferencia**	**Cantidad de Datos Entrenamiento**	**Cantidad de Datos Inferencia**	**Error Entrenamiento**	**Error Inferencia**
1	184	-	-	-	-	-	-
		90%	10%	166	18	12,11%	8,21%
		80%	20%	147	37	12,33%	11,04%
		70%	30%	129	55	12,86%	9,99%
		50%	50%	92	92	12,96%	11,30%
		30%	70%	55	129	9,92%	25,56%

En las **Figura 3-28**, **Figura 3-29** y **Figura 3-30** se presentan las gráficas obtenidas donde se puede observar cómo es el comportamiento de la inferencia obtenida contra los valores reales ingresados al algoritmo para los casos donde los porcentajes de datos destinados para el entrenamiento del algoritmo son del 90%, 50% y 30%, respectivamente.

Figura 3-28: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de un año de datos con un porcentaje de datos de entrenamiento del 90%.

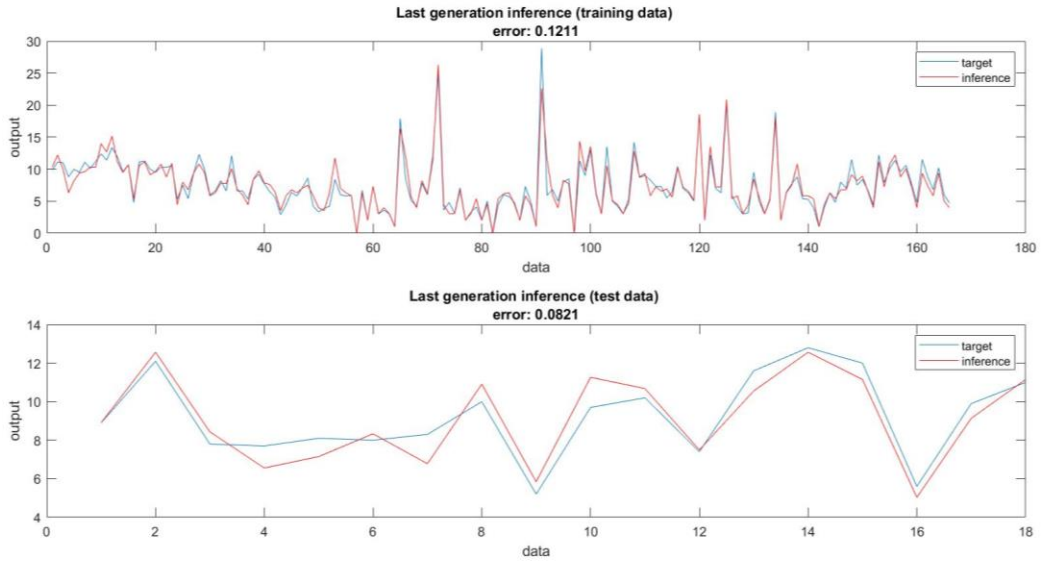


Figura 3-29: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de un año de datos con un porcentaje de datos de entrenamiento del 50%.

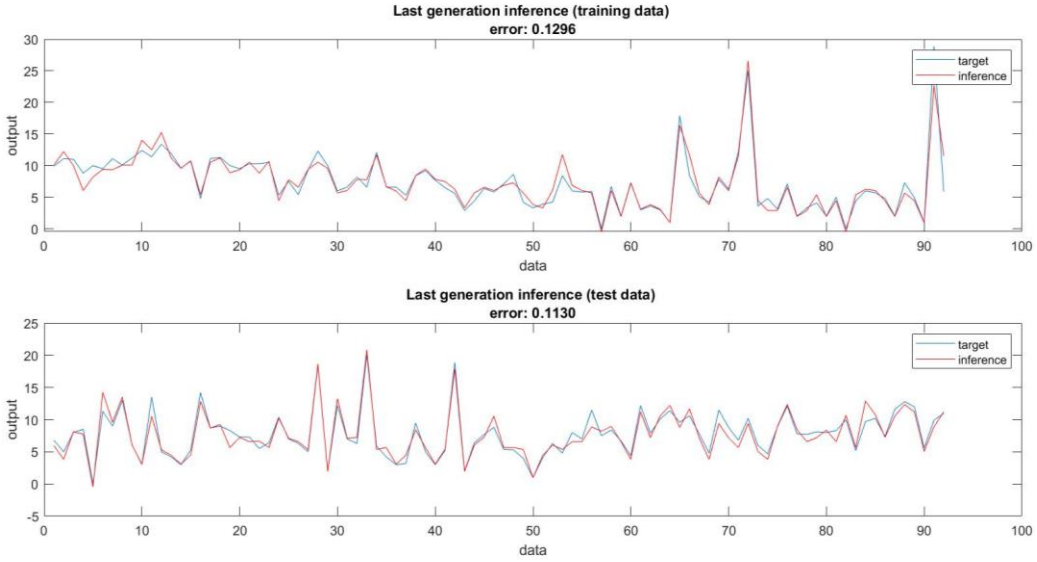
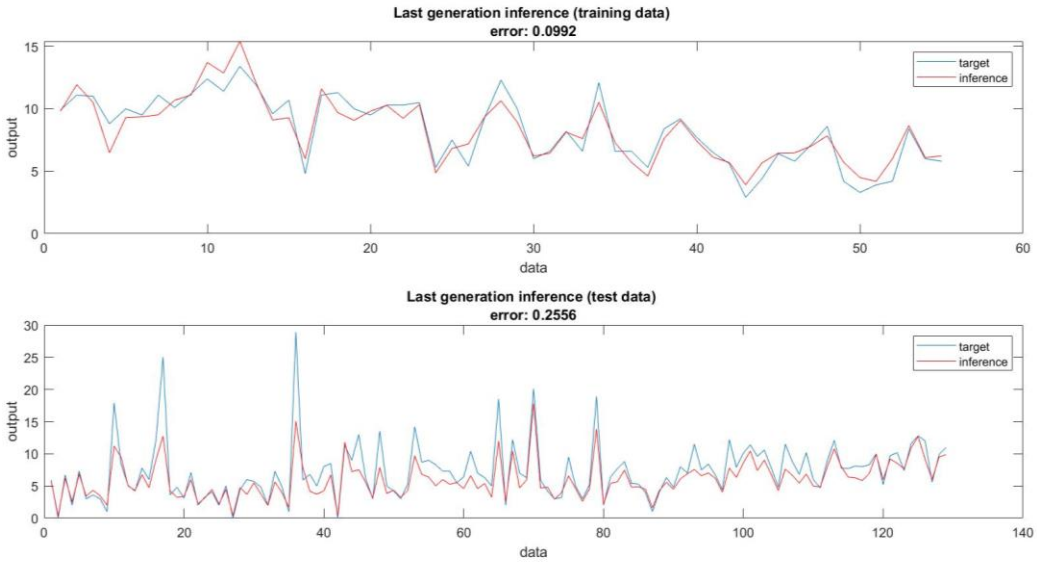


Figura 3-30: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de un año de datos con un porcentaje de datos de entrenamiento del 30%.



- **Base de datos de dos años**

A continuación, en la **Tabla 3.12** se presenta el consolidado de los resultados de los pronósticos realizados teniendo en cuenta una base de datos de dos años, en el cual se

varía la cantidad de datos utilizados para el entrenamiento y la validación de la inferencia obtenida por medio del algoritmo de regresión.

Tabla 3.12: Resultados de pronósticos del caso de estudio 4 para dos años.

Ubicación 4							
Años	**Cantidad de Datos**	**% de datos para Entrenamiento**	**% de datos para Inferencia**	**Cantidad de Datos Entrenamiento**	**Cantidad de Datos Inferencia**	**Error Entrenamiento**	**Error Inferencia**
2	383	95%	5%	364	19	10,59%	8,45%
		90%	10%	345	38	10,69%	8,48%
		80%	20%	306	77	11,47%	8,77%
		70%	30%	268	115	10,88%	9,97%
		50%	50%	192	191	11,64%	10,42%
		30%	70%	115	268	13,79%	10,76%

En las **Figura 3-31**, **Figura 3-32** y **Figura 3-33** se presentan las gráficas obtenidas donde se puede observar cómo es el comportamiento de la inferencia obtenida contra los valores reales ingresados al algoritmo para los casos donde los porcentajes de datos destinados para el entrenamiento del algoritmo son del 90%, 50% y 30%, respectivamente.

Figura 3-31: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de dos años de datos con un porcentaje de datos de entrenamiento del 90%.

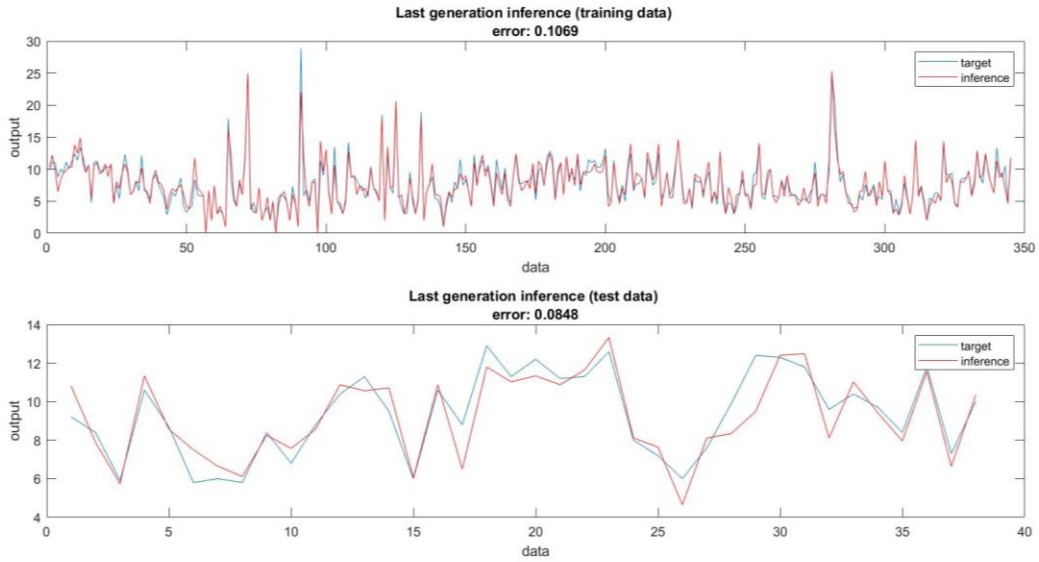


Figura 3-32: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de dos años de datos con un porcentaje de datos de entrenamiento del 50%.

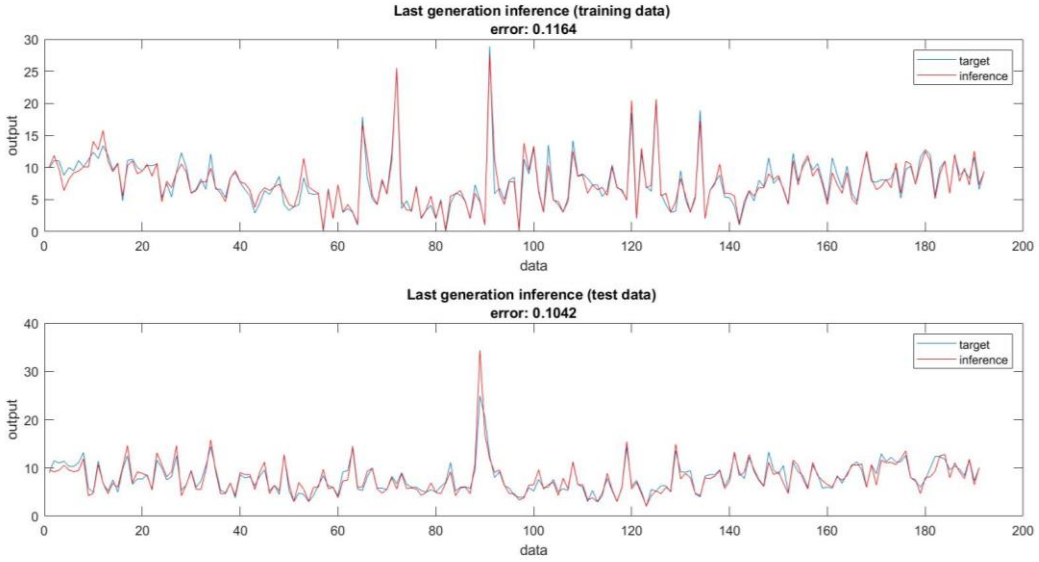
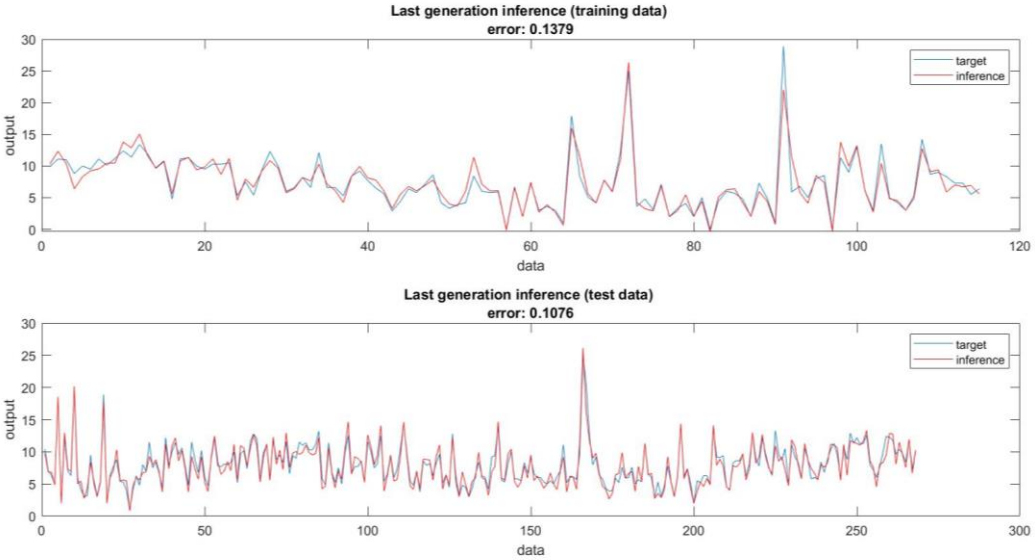


Figura 3-33: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de dos años de datos con un porcentaje de datos de entrenamiento del 30%.



- **Base de datos de tres años**

A continuación, en la **Tabla 3.13** se presenta el consolidado de los resultados de los pronósticos realizados teniendo en cuenta una base de datos de tres años, en el cual se

varía la cantidad de datos utilizados para el entrenamiento y la validación de la inferencia obtenida por medio del algoritmo de regresión.

Tabla 3.13: Resultados de pronósticos del caso de estudio 4 para tres años.

Ubicación 4							
Años	**Cantidad de Datos**	**% de datos para Entrenamiento**	**% de datos para Inferencia**	**Cantidad de Datos Entrenamiento**	**Cantidad de Datos Inferencia**	**Error Entrenamiento**	**Error Inferencia**
3	749	-	-	-	-	-	-
		90%	10%	674	75	9,46%	12,28%
		80%	20%	599	150	9,73%	9,88%
		70%	30%	524	225	15,66%	16,75%
		50%	50%	375	374	10,52%	9,07%
		30%	70%	225	524	11,12%	9,76%

En las **Figura 3-34**, **Figura 3-35** y **Figura 3-36** se presentan las gráficas obtenidas donde se puede observar cómo es el comportamiento de la inferencia obtenida contra los valores reales ingresados al algoritmo para los casos donde los porcentajes de datos destinados para el entrenamiento del algoritmo son del 90%, 50% y 30%, respectivamente.

Figura 3-34: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de tres años de datos con un porcentaje de datos de entrenamiento del 90%.

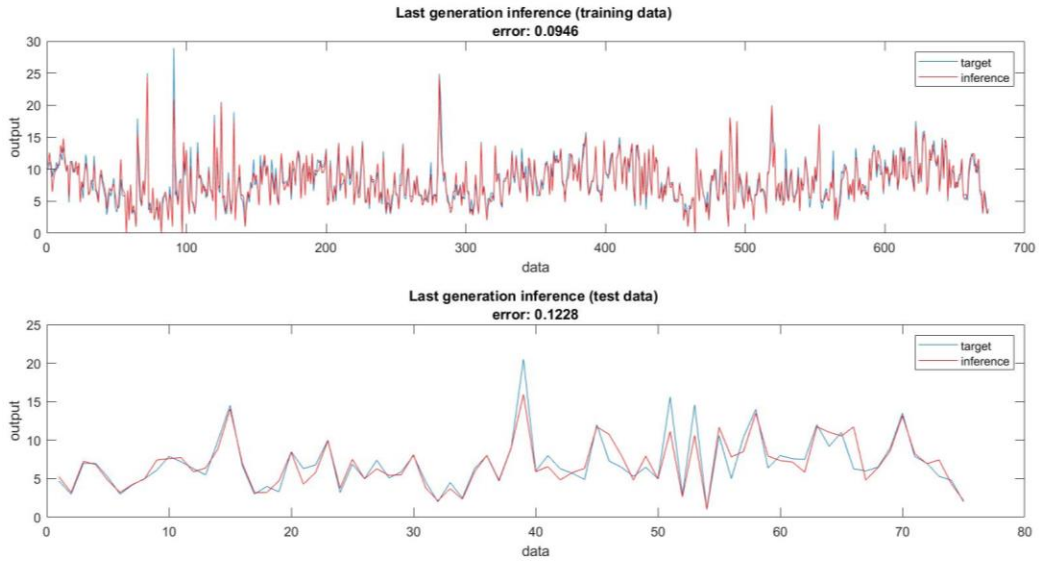


Figura 3-35: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de tres años de datos con un porcentaje de datos de entrenamiento del 50%.

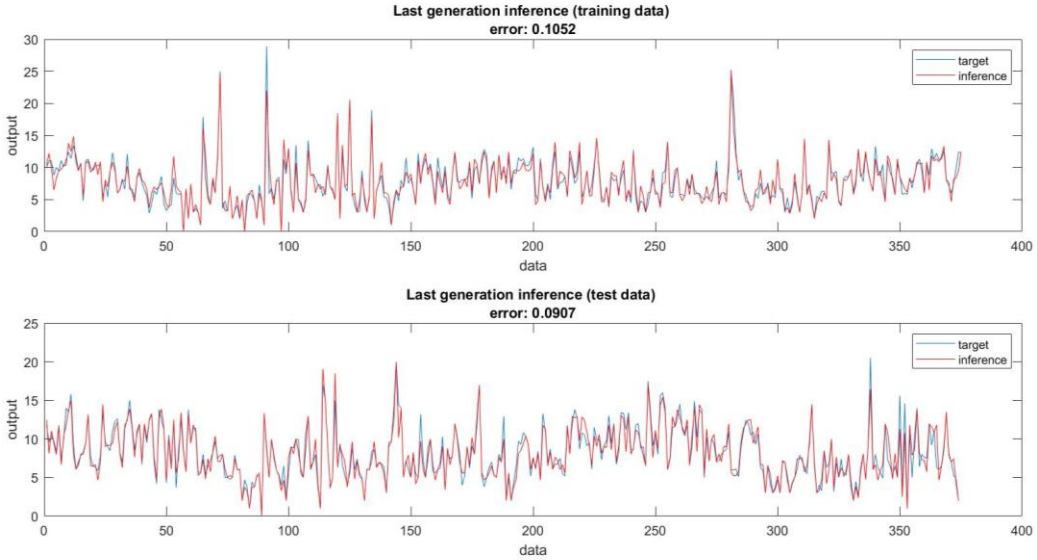
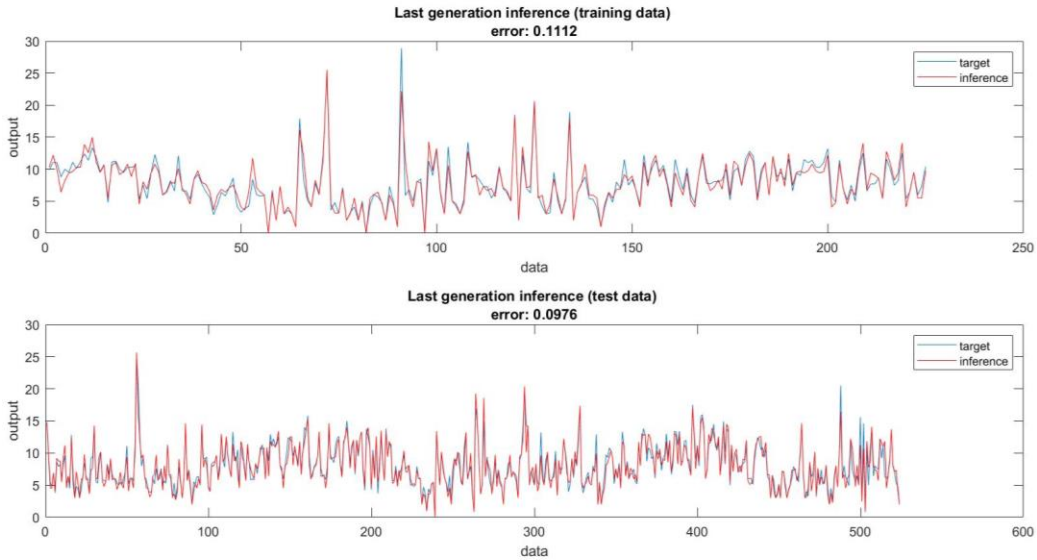


Figura 3-36: Pronósticos de velocidades de viento para el caso de estudio 4 para el escenario de tres años de datos con un porcentaje de datos de entrenamiento del 30%.



3.3 Análisis de resultados obtenidos

De las **Tabla 3.2**, **Tabla 3.3** y **Tabla 3.4**, y de las **Figura 3-1** a la **Figura 3-9** se puede observar que los errores obtenidos en las inferencias realizadas de los pronósticos realizados para el caso de estudio 1 son inferiores a 11%; estando los errores más altos en el escenario donde se consideran tres años de datos históricos. Por otra parte, los escenarios donde se tuvieron en cuenta datos históricos de uno y dos años tienen unos errores cercanos al 9% tanto para el caso de errores de entrenamiento como de la inferencia realizada.

Para el caso de estudio 2, de las **Tabla 3.5** a **Tabla 3.7** y de las **Figura 3-10** a la **Figura 3-18** se aprecia que los errores de este caso de estudio se encuentran en promedio en valores del orden de 8% para los errores de entrenamiento, mientras que los errores asociados a la inferencia se encuentran en promedio al 10%. Presentando un error máximo en el entrenamiento del algoritmo del 14,7% en el escenario donde se contempló la base de datos de dos años, y el error máximo en las inferencias obtenidas se presentó en el escenario donde se evaluó la base de datos de tres años, con un valor de 16,3%.

Por otra parte, de las **Tabla 3.8**, **Tabla 3.9** y **Tabla 3.10**, y de la **Figura 3-19** a la **Figura 3-27** se tiene el caso de estudio 3, donde los errores fueron mayores a los obtenidos que en el caso anterior, sin embargo el error máximo obtenido entre las inferencias realizadas se encuentra en un 18,7%, y el error máximo para el entrenamiento del algoritmo está en 20%. Para este caso de estudio se tiene que los errores más grandes tanto en inferencia como en entrenamiento se encuentran en el escenario donde se evaluó un año de datos experimentales o datos históricos. Sin embargo, en promedio los valores de los errores obtenidos para este caso de estudio se encuentran cercanos al 13%.

Finalmente, para el caso de estudio 4 se tienen las **Tabla 3.11**, **Tabla 3.12** y **Tabla 3.13**, y desde la **Figura 3-28** hasta la **Figura 3-36**, de donde se puede observar que en promedio los errores de entrenamiento y de inferencia se encuentran en valores cercanos al 11%; sin embargo, es en este caso de estudio donde se presentó el error más alto de inferencia entre todas las diferentes pruebas realizadas, con un valor de 25%; este valor se presentó en el escenario donde se evaluó una base de datos con históricos de un año.

En cualquier caso, aun cuando este fue el valor de error en inferencia más alto de todas las pruebas realizadas, en la **Figura 3-30** se puede observar que el comportamiento del gráfico de la inferencia sigue de cerca el gráfico con los valores históricos ingresados al algoritmo con base al cuál se realiza validación de dicha inferencia. Además, en esta misma figura se puede ver que el error asociado al entrenamiento es del 9,9%

Del caso de estudio 4 también es válido mencionar que este es el que menor cantidad de datos históricos contaba teniendo como referencia el mismo período de tiempo que los demás casos de estudio evaluados.

Los errores obtenidos en el presente trabajo con la implementación del algoritmo de regresión simbólica son mejores que los resultados obtenidos en [44] y [45], donde se implementaros metodologías de predicción numérica del clima (NWP- por sus siglas en inglés) y redes neuronales, en los cuales se tuvieron errores hasta del 33%.

Teniendo en cuenta lo anterior y con los valores de los errores obtenidos en las pruebas realizadas en los cuatro casos de estudio contemplados en el presente trabajo, se tiene que una adecuada cantidad de datos para el entrenamiento del algoritmo seleccionado está en el orden del 30%.

En este orden de ideas, en el presente trabajo se probó con los escenarios contemplados con bases de datos de un año y tres años que:

- Para el primer caso, con bases de datos de un año, si se cuenta con tres meses de datos históricos se puede realizar un pronóstico de las velocidades del viento offshore de hasta 8 meses;
- Por otra parte, para el segundo escenario mencionado con bases de datos de tres años, si se cuenta con ceca de año datos históricos es posible con el algoritmo utilizado realizar un pronóstico de velocidades de viento offshore de hasta dos años.

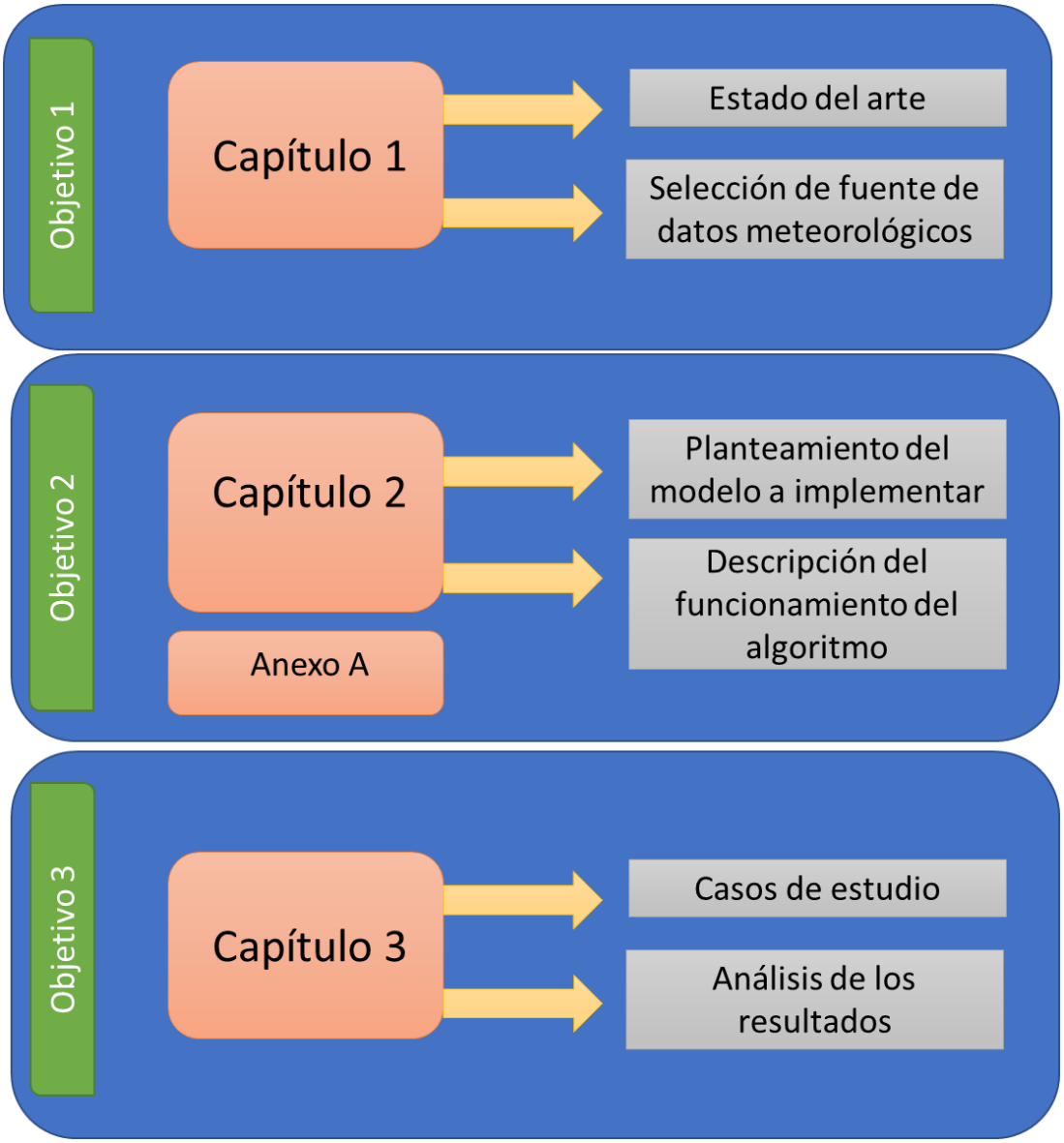
Con lo anteriormente planteado, se tiene que se cumplió el tercer objetivo específico dado que se implementó el modelo de regresión simbólica propuesto en diferentes conjuntos de

datos para la generación de pronósticos de vientos, decidiendo además un número apropiado de datos de entrenamiento y prueba.

3.4 Cumplimiento de los objetivos

A continuación, en la **Figura 3-37:** Logro de objetivos por secciones del documento. se hace una síntesis de cómo se alcanzó el cumplimiento de los objetivos propuestos a lo largo del presente documento:

Figura 3-37: Logro de objetivos por secciones del documento.



4. Conclusiones y recomendaciones

4.1 Conclusiones

- El algoritmo genético de regresión simbólica hibrida implementado contempla métodos matemáticos de kernel que lo hacen versátil en la implementación de regresiones lineales evitando que se presenten matrices singulares que no permitan el uso de este en determinados escenarios.

- Se obtuvieron cuatro conjuntos de datos meteorológicos históricos con base a los cuales se pudieron realizar pronósticos de velocidades de vientos promedios.

- Se realizó de manera satisfactoria el planteamiento de un método de regresión simbólica con el cual se pueden realizar pronósticos de velocidades de vientos promedio correlacionados a las variables meteorológicas con las que se disponía.

- Se implementó exitosamente el modelo de regresión simbólica híbrida en cuatro conjuntos de datos meteorológicos históricos y se seleccionó un número apropiado de datos de entrenamiento y prueba para la elaboración de pronósticos de velocidades de viento promedio con el modelo planteado.

- Se probó la eficacia del algoritmo implementado tanto con bases de datos que contenían información histórica completa como también en el caso donde se presentan bases de datos con las cuales no se cuenta con un histórico completo

- El algoritmo planteado puede ser implementado para realizar la planeación a mediano y largo plazo de las plantas de generación eólica offshore: desde un planeamiento en la etapa de diseño del proyecto hasta la planeación de los

mantenimientos siempre que se cuente con un conjunto de datos históricos suficiente según los objetivos a alcanzar.

4.2 Trabajos futuros

- Integrar una segunda etapa de optimización para minimizar los errores en las inferencias elaboradas con el algoritmo de regresión simbólica hibrida planteado en el presente trabajo.

- Realizar la integración de la velocidad del viento para seleccionar las mejores ubicaciones donde instalar los aerogeneradores dentro de una planta eólica offshore.

- Evaluar el potencial eólico offshore en las zonas costeras de Colombia y Latinoamérica

- Realizar la integración de generación offshore a una micro-red para mejorar las condiciones de calidad del servicio en zonas costeras o islas que presenten déficit en este aspecto del servicio.

A. Anexo: Algoritmo de regresión simbólica híbrido en Matlab

```
%% Symbolic_regression_algorithm (HYBRID VERSION)

% Description: this code executes a new multigene version of symbolic
regression algorithm which includes some concepts of no-linear
regression by means of
%               kernel functions. The algorithm is executed on one of two
experimental data bases of industrial variables.
%               It is made to build a model which estimates the value of
some variable from auxiliar ones.
%               The first data base consists in a pH neutralization
process data and the second one in a concrete quality data.
%               You need to configure some parameters and choice the data
base you want to work with.
%               The final result is a plot of the estimation given by the
best models found in the first, the last and the best generation.

clear all; close all; clc;

%% ---------------------------------------------------------------
CONFIGURATION -------------------------------------------------------------
----------------
%% ---------------------------------------------------------------------
-----------------------------------------------------------------------
------------

%% ----------------------------------------------------------------
Loading data --------------------------------------------------------------
---------------

data = 40
%If data = 1, data base 1 is loaded

if data == 39
M = xlsread('WindCayman_3.xlsx');
nx= M(293:end,1:end-1);
ny= M(293:end,end);
x = M(1:292,1:end-1);
y = M(1:292,end);
end
```

```
if data == 40
M = xlsread('WindSanAndres_2.xlsx');
nx= M(220:end,1:end-1);
ny= M(220:end,end);
x = M(1:219,1:end-1);
y = M(1:219,end);
end

if data == 41
M = xlsread('WindKailuaHawuai_1.xlsx');
nx= M(106:end,1:end-1);
ny= M(106:end,end);
x = M(1:105,1:end-1);
y = M(1:105,end);
end

if data == 42
M = xlsread('WindTalcahuano');
nx= M(76:end,1:end-1);
ny= M(76:end,end);
x = M(1:75,1:end-1);
y = M(1:75,end);
end


configuration.data.test_data_input = nx;
configuration.data.test_data_output = ny;
configuration.data.train_data_input = x;
configuration.data.train_data_output = y;

%% ----------------------------------------------------- Taking
experimental data matrix information -----------------------------------
-----------------------


%Knowing number of experimental observations and variables we are
working with.
configuration.data.num_cases =
size(configuration.data.train_data_input,1);
configuration.data.num_independent_vars =
size(configuration.data.train_data_input,2);

%% ------------------------------------------------------- Numerical
parameters configuration ----------------------------------------------
----------------

configuration.population_size =  50  %50
%Amount of individuals or potential solutions in the population.
configuration.num_generations =  150 %36
%Number of generations or iterations we want the algorithm to run.
configuration.num_max_genes = 4;
%Maximum number of genes per individual.
configuration.depth_max_initial = 5;
%Individuals' maximum length at the beginning of the process.
configuration.depth_max_created = 6;
%Individuals' maximum length after any genetic operation.
```

```
%Genetic operations configuration
%Genetic operations probabilities. The sum of all probabilities MUST be
1!
configuration.crossover.probability = 0.9;
configuration.reproduction.probability = 0.1;
configuration.mutation.probability = 0;
configuration.permutation.probability = 0;
configuration.edition.frequency = 0;
%Edition frequency must be less than num_generations. If it is greater,
edition

%never happens.
configuration.encapsulation.probability = 0;
configuration.initial_decimation = 0;
%1 = active, 0 = inactive. If active, decimation is performed in
generation 0.
configuration.decimation.percentage = 10;
%[0 100%] percentage of individuals to be conserved after fitness
calculation in

%generation 0.
configuration.elitist.percentage = 10;
%[0-100%] if it is greater than cero elitist strategy is active.

%% ---------------------------------------------------- Cualitative
parameters configuration ------------------------------------------------
----------------

configuration.generative_mode = 2;
%Way to build tree structures. 1='full', 2='grow', 3='ramped half and
half'.

% Selection operation configuration
%1='fitness proportional', 2='tournament', 3='ranking'
configuration.selection.mode_for_reproduction = 1;
configuration.selection.mode_for_crossover_p1 = 1;
configuration.selection.mode_for_crossover_p2 = 1;
configuration.selection.mode_for_mutation = 1;
configuration.selection.mode_for_permutation = 1;
configuration.selection.mode_for_encapsulation = 1;
configuration.selection.mode_for_decimation = 1;

configuration.selection.tournament_candidats = 2;
%It is important only if selection mode is 'tournament'.
configuration.selection.rank_pressure = 2;
%Values in the range [1,2]. It is importat only if selection mode is
'ranking'.
configuration.selection.mating_pool = 1;
%1=active 0=inactive (when proportional fitness selection) if it is
active all

%selection methods are the same.
```

```
%% ------------------------------------------------------- Functions
set configuration -------------------------------------------------------
-------------


%Define functions and mathematical operations used to build the
structures.
configuration.functions.names =
{'times','minus','plus','rdivide','tanh','sin','cos','exp','pdivision','
psqrt','plog'};
configuration.functions.symbols =
char('A','B','C','D','E','F','G','H','I','J','K');
configuration.functions.active = logical([1 1 1 1 1 1 1 1 1 1 1]);
%Only those which are '1' are taken into account.
configuration.functions.num_active =
numel(find(configuration.functions.active));


%Looking for the number of input arguments for every function.
for fun = 1:length(configuration.functions.names)
    configuration.functions.num_arguments(fun) =
nargin(configuration.functions.names{fun});
end

%% ---------------------------------------------------- Dual
representation and kernel configuration ---------------------------------
-------------------------

configuration.kernel.type = 'linear';
%Kernel to use in inference calculations. It can be 'linear', 'fast
gaussian' or

%'Polynomial'.
configuration.kernel.lambda = 0.1;
%Compensation parameter in the dual representation of linear regressión
with

%regularization.

%% ------------------------------------------------------- Final
solution configuration --------------------------------------------------
------------------

configuration.solution_mode = 1;
%1=best solution in the last generation, 2=best solution among all
generations.

%% ------------------------------------------------------------- END
CONFIGURATION -----------------------------------------------------------
--------------
%% ---------------------------------------------------------------------
---------------------------------------------------------------------
------------
```

```
%% ---------------------------------------------------- HYBRID MULTIGEN
SYMBOLIC REGRESSION ALGORITHM -------------------------------------------
--------------------
%% ---------------------------------------------------------------------
---------------------------------------------------------------------
------------

edition_control = 0;

%% ----------------------------------------------------- STEP 1. Building
the initial population ------------------------------------------------
-----------

initial_population =
initial_population_creation_multigene(configuration);  %Creating an
initial population without identical individuals.

if any(cellfun(@isempty,initial_population)) == 0
%If every individual was initialized

    population.individuals = initial_population;

%% ------------------------------------------------ STEP 2. Modifying the
population every generation -------------------------------------------
-----------


%Scoring each individual of the initial population.
    population = scoring_population_multigene(population,configuration);


%If decimation option is active only some percentage of the population
is kept.
    if configuration.initial_decimation == 1
        [population,configuration] =
decimation_operation(population,configuration);
        population =
scoring_population_multigene(population,configuration);
    end

    instance = configuration.data.num_cases;

%It is executed iteratively
    for generation=1:configuration.num_generations

        generation
        edition_control = edition_control + 1;


%Keeping in memory the best expression by generation with its scores.
        best_score = max([population.scores.normalized]);
        best_id = find([population.scores.normalized] == best_score);
        register(generation).best_individual =
population.individuals{best_id(1)};
        register(generation).coefficients =
population.coefficients{best_id(1)};
```

```
        register(generation).best_scores =
population.scores(best_id(1));
        register(generation).mean_score = mean([population.scores.raw]);



%Sampling data base to avoid over-fitting and reduce computational cost.
        if mod(generation,2)==1
            instance = ceil(configuration.data.num_cases*rand);
            configuration.data.train_data_input = x(1:instance,:);
            configuration.data.train_data_output = y(1:instance,:);
        else
            configuration.data.train_data_input = x(instance:end,:);
            configuration.data.train_data_output = y(instance:end,:);
        end


%Creating a new population by applying genetic operations.
        [population,configuration] =
create_new_population_multigene(population,configuration);

%Scoring the individuals of the new population.
        population =
scoring_population_multigene(population,configuration);


%If edition option is active, individuals are evaluated and simplified.
        if edition_control == configuration.edition.frequency
           population =
editing_population_multigene(population,configuration);
           edition_control = 0;
        end
    end


configuration.data.train_data_input = x;
configuration.data.train_data_output = y;


%% ----------------------------------------------------- STEP 3. Giving
the final solution -----------------------------------------------------
-----------

generation_scores=[register(:).best_scores];
[~,generation_id]=min([generation_scores.raw]);


%Evaluating the solution with the whole database.
[~,~,register(generation).coefficients] =
scoring_individual_multigene(register(generation).best_individual,config
uration.data.train_data_input,configuration.data.train_data_output,confi
guration);
[~,~,register(generation_id).coefficients] =
scoring_individual_multigene(register(generation_id).best_individual,con
figuration.data.train_data_input,configuration.data.train_data_output,co
nfiguration);
```

```
if configuration.solution_mode == 1

%Finding the best solution in the last generation.
    solution =
convert_sym2str(register(generation).best_individual,configuration);
    coefficients = register(generation).coefficients;

else

%Finding the best solution among all the generations.
    solution =
convert_sym2str(register(generation_id).best_individual,configuration);
    coefficients = register(generation_id).coefficients;
end


%% ---------------------------------------------------------------- END
ALGORITHM ----------------------------------------------------------------
-----------
%% ---------------------------------------------------------------------
---------------------------------------------------------------------------
------------


%% ------------------------------------------------------------------
PLOTTING -----------------------------------------------------------------
-------------
%% ---------------------------------------------------------------------
---------------------------------------------------------------------------
------------


    %% ------------------------------------------------ Fist generation
best individual inference ------------------------------------------------
-----------
    figure4=figure('Color',[1 1 1],'Name','Results','Position',[50 170
400 400]);


%Inference with training data
    prediction =
making_inference(register(1).best_individual,register(1).coefficients,co
nfiguration.data.train_data_input,configuration);
    error = mean(abs((configuration.data.train_data_output-
prediction)./prediction));

    subplot(2,1,1)
    plot(configuration.data.train_data_output)
    hold on;
    plot(prediction,'r')
    xlabel('data'), ylabel('output'), title(sprintf('First generation
inference (training data) \n error: %1.4f % ', error))
    legend('target','inference')
```

```
%Inference with test data
    prediction =
making_inference(register(1).best_individual,register(1).coefficients,co
nfiguration.data.test_data_input,configuration);
    error = mean(abs((configuration.data.test_data_output-
prediction)./prediction));

    subplot(2,1,2)
    plot(configuration.data.test_data_output)
    hold on;
    plot(prediction,'r')
    xlabel('data'), ylabel('output'), title(sprintf('First generation
inference (test data) \n error: %1.4f % ', error))
    legend('target','inference')

    %% ------------------------------------------------- Last generation
best individual inference ---------------------------------------------
-----------

    figure5=figure('Color',[1 1 1],'Name','Results','Position',[480 170
400 400]);


%Inference with training data
    prediction =
making_inference(register(generation).best_individual,register(generatio
n).coefficients,configuration.data.train_data_input,configuration);
    error = mean(abs((configuration.data.train_data_output-
prediction)./prediction));

    subplot(2,1,1)
    plot(configuration.data.train_data_output)
    hold on;
    plot(prediction,'r')
    xlabel('data'), ylabel('output'), title(sprintf('Last generation
inference (training data) \n error: %1.4f % ', error))
    legend('target','inference')


%Inference with test data
    prediction =
making_inference(register(generation).best_individual,register(generatio
n).coefficients,configuration.data.test_data_input,configuration);
    error = mean(abs((configuration.data.test_data_output-
prediction)./prediction));

    subplot(2,1,2)
    plot(configuration.data.test_data_output)
    hold on;
    plot(prediction,'r')
    xlabel('data'), ylabel('output'), title(sprintf('Last generation
inference (test data) \n error: %1.4f % ', error))
    legend('target','inference')
```

```
    %% ------------------------------------------------- Best generation
best individual inference ---------------------------------------------
-----------

    figure6=figure('Color',[1 1 1],'Name','Results','Position',[910 170
400 400]);


%Inference with training data
    prediction =
making_inference(register(generation_id).best_individual,register(genera
tion_id).coefficients,configuration.data.train_data_input,configuration)
;
    error = mean(abs((configuration.data.train_data_output-
prediction)./prediction));

    subplot(2,1,1)
    plot(configuration.data.train_data_output)
    hold on;
    plot(prediction,'r')
    xlabel('data'), ylabel('output'), title(sprintf('%i^{th} generation
inference (training data) \n error: %1.4f % ', generation_id, error))
    legend('target','inference')


%Inference with testing data
    prediction =
making_inference(register(generation_id).best_individual,register(genera
tion_id).coefficients,configuration.data.test_data_input,configuration);
    error = mean(abs((configuration.data.test_data_output-
prediction)./prediction));

    subplot(2,1,2)
    plot(configuration.data.test_data_output)
    hold on;
    plot(prediction,'r')
    xlabel('data'), ylabel('output'), title(sprintf('%i^{th} generation
inference (test data) \n error: %1.4f % ', generation_id, error))
    legend('target','inference')

%% ---------------------------------------------------------------- END
PLOTTING ---------------------------------------------------------------
------------
%% ---------------------------------------------------------------------
---------------------------------------------------------------------
------------

else
  error('There are problems building the initial population. Try with
other parameters')
end
```

Bibliografía

[1] C. Arriagada, “ANALISIS COMPARATIVO DE LA GOBERNABILIDAD DE MERCADOS DE GENERACION ELECTRICA,” PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE ESCUELA DE INGENIERIA, 1998.

[2] M. Deveci, E. Ozcan, and R. John, “Offshore wind farms: A fuzzy approach to site selection in a black sea region,” *2020 IEEE Texas Power Energy Conf. TPEC 2020*, 2020, doi: 10.1109/TPEC48276.2020.9042530.

[3] U. Cali, N. Erdogan, S. Kucuksari, and M. Argin, “TECHNO-ECONOMIC analysis of high potential offshore wind farm locations in Turkey,” *Energy Strateg. Rev.*, vol. 22, no. November 2017, pp. 325–336, 2018, doi: 10.1016/j.esr.2018.10.007.

[4] L. A. Barroso and A. J. Conejo, *Decision making under uncertainty in electricity markets*. 2006.

[5] Y. Zhao, L. Ye, Z. Li, X. Song, Y. Lang, and J. Su, “A novel bidirectional mechanism based on time series model for wind power forecasting,” *Appl. Energy*, vol. 177, pp. 793–803, 2016, doi: https://doi.org/10.1016/j.apenergy.2016.03.096.

[6] GWEC, “Global wind energy council report 2018,” *Wind Glob. Counc. Energy*, no. April, pp. 1–61, 2019, [Online]. Available: www.gwec.net.

[7] A. E. Saleh, M. S. Moustafa, K. M. Abo-Al-Ez, and A. A. Abdullah, “A hybrid neuro-fuzzy power prediction system for wind energy generation,” *Int. J. Electr. Power Energy Syst.*, vol. 74, pp. 384–395, 2016, doi: 10.1016/j.ijepes.2015.07.039.

[8] A. M. Foley, P. G. Leahy, A. Marvuglia, and E. J. McKeogh, “Current methods and advances in forecasting of wind power generation,” *Renew.*

Energy, vol. 37, no. 1, pp. 1–8, 2012, doi: 10.1016/j.renene.2011.05.033.

[9] M. Á. Vanegas Ramos, “Implementación de modelos locales en el espacio de fase para el pronóstico de variables hidrometeorológicas a partir de series de tiempo,” p. 398, 2011.

[10] G. Riahy and M. Abedi, “Short term wind speed forecasting for wind turbine applications using linear prediction method,” *Renew. Energy*, vol. 33, pp. 35–41, Jan. 2008, doi: 10.1016/j.renene.2007.01.014.

[11] Z. Lin and X. Liu, “Wind power forecasting of an offshore wind turbine based on high-frequency SCADA data and deep learning neural network,” *Energy*, vol. 201, p. 117693, 2020, doi: https://doi.org/10.1016/j.energy.2020.117693.

[12] C. Wan, Y. Song, Z. Xu, G. Yang, and A. H. Nielsen, “Probabilistic Wind Power Forecasting with Hybrid Artificial Neural Networks,” *Electr. Power Components Syst.*, vol. 44, no. 15, pp. 1656–1668, 2016, doi: 10.1080/15325008.2016.1198437.

[13] S. Salcedo-Sanz, E. G. Ortiz-García, Á. M. Pérez-Bellido, A. Portilla-Figueras, and L. Prieto, “Short term wind speed prediction based on evolutionary support vector regression algorithms,” *Expert Syst. Appl.*, vol. 38, no. 4, pp. 4052–4057, 2011, doi: 10.1016/j.eswa.2010.09.067.

[14] A. Marvuglia and A. Messineo, “Monitoring of wind farms’ power curves using machine learning techniques,” *Appl. Energy*, vol. 98, no. May, pp. 574–583, 2012, doi: 10.1016/j.apenergy.2012.04.037.

[15] V. Veloso de Melo and W. Banzhaf, “Automatic feature engineering for regression models with machine learning: An evolutionary computation and statistics hybrid,” *Inf. Sci. (Ny).*, vol. 430–431, no. November, pp. 287–313, 2018, doi: 10.1016/j.ins.2017.11.041.

[16] C. Wan, Z. Xu, P. Pinson, Z. Y. Dong, and K. P. Wong, “Probabilistic forecasting of wind power generation using extreme learning machine,” *IEEE Trans. Power Syst.*, vol. 29, no. 3, pp. 1033–1044, 2014, doi: 10.1109/TPWRS.2013.2287871.

[17] Y. Zhang, K. Liu, L. Qin, and X. An, “Deterministic and probabilistic interval

prediction for short-term wind power generation based on variational mode decomposition and machine learning methods," *Energy Convers. Manag.*, vol. 112, pp. 208–219, 2016, doi: 10.1016/j.enconman.2016.01.023.

[18] S. An, H. Shi, Q. Hu, X. Li, and J. Dang, "Fuzzy rough regression with application to wind speed prediction," *Inf. Sci. (Ny).*, vol. 282, pp. 388–400, 2014, doi: 10.1016/j.ins.2014.03.090.

[19] G. Santamaría-Bonfil, A. Reyes-Ballesteros, and C. Gershenson, "Wind speed forecasting for wind farms: A method based on support vector regression," *Renew. Energy*, vol. 85, pp. 790–809, 2016, doi: 10.1016/j.renene.2015.07.004.

[20] F. Andrés and A. Rodríguez, "Tendencias recientes en el pronóstico de velocidad de viento para generación eólica," Universidad Nacional de Colombia, 2017.

[21] C. Wan, Z. Xu, P. Pinson, Z. Y. Dong, and K. P. Wong, "Optimal prediction intervals of wind power generation," *IEEE Trans. Power Syst.*, vol. 29, no. 3, pp. 1166–1174, 2014, doi: 10.1109/TPWRS.2013.2288100.

[22] H. Demolli, A. S. Dokuz, A. Ecemis, and M. Gokcek, "Wind power forecasting based on daily wind speed data using machine learning algorithms," *Energy Convers. Manag.*, vol. 198, no. July, p. 111823, 2019, doi: 10.1016/j.enconman.2019.111823.

[23] T. G. Barbounis, J. B. Theocharis, M. C. Alexiadis, and P. S. Dokopoulos, "Long-term wind speed and power forecasting using local recurrent neural network models," *IEEE Trans. Energy Convers.*, vol. 21, no. 1, pp. 273–284, 2006, doi: 10.1109/TEC.2005.847954.

[24] C. Potter and M. Negnevitsky, "Very short-term wind forecasting for tasmanian power generation," *2006 IEEE Power Eng. Soc. Gen. Meet. PES*, vol. 21, no. 2, pp. 965–972, 2006, doi: 10.1109/pes.2006.1709044.

[25] G. Giebel and E. Al., "The State of the Art in Short-Term Prediction of Wind Power," *ANEMOS.plus*, no. January, pp. 1–110, 2011, doi: 10.13140/RG.2.1.2581.4485.

[26] IEEE PES, “Open Data Sets « IEEE PES Intelligent Systems Subcommittee.” [Online]. Available: https://site.ieee.org/pes-iss/data-sets/.

[27] NASA, “NASA POWER - Prediction Of Worldwide Energy Resources.” 2019, [Online]. Available: https://power.larc.nasa.gov/data-access-viewer/.

[28] “Solcast, 2019,” *Global solar irradiance data and PV system power output data*. https://solcast.com/.

[29] “WunderMap® | Interactive Weather Map and Radar | Weather Underground.” 2019, [Online]. Available: https://www.wunderground.com.

[30] W. B. Langdon, R. Poli, N. F. McPhee, and J. R. Koza, “Genetic programming: An introduction and tutorial, with a survey of techniques and applications,” in *Studies in Computational Intelligence*, vol. 115, 2008, pp. 927–1028.

[31] S. Nguyen, M. Zhang, D. Alahakoon, and K. C. Tan, “Visualizing the evolution of computer programs for genetic programming [Research Frontier],” *IEEE Comput. Intell. Mag.*, vol. 13, no. 4, pp. 77–94, 2018, doi: 10.1109/MCI.2018.2866731.

[32] I. Icke and J. C. Bongard, “Improving genetic programming based symbolic regression using deterministic machine learning,” *2013 IEEE Congr. Evol. Comput. CEC 2013*, no. June, pp. 1763–1770, 2013, doi: 10.1109/CEC.2013.6557774.

[33] J. Žegklitz and P. Pošík, “Symbolic regression in dynamic scenarios with gradually changing targets,” *Appl. Soft Comput. J.*, vol. 83, p. 105621, 2019, doi: 10.1016/j.asoc.2019.105621.

[34] S. Chatterjee and A. S. Hadi, *Regression Analysis by Example*, 5th Editio. Wiley, 2013.

[35] E. Cortés Pérez, A. Nuñez Rodríguez, R. E. Moreno De La Torre, O. Lastres Danguillecourt, and J. R. Dorrego Portela, “Forecast of Wind Speed with a Backpropagation Artificial Neural Network in the Isthmus of Tehuantepec Region in the State of Oaxaca, Mexico.,” *Acta Univ.*, vol. 22, pp. 7–14, Dec. 2012, [Online]. Available: https://www.redalyc.org/articulo.oa?id=41623190001.

[36] C. A. Martínez, “Problemas abiertos en la aplicación de la Regresión Simbólica en el pronóstico de series de tiempo,” Nacional de Colombia Sede Medellín, 2011.

[37] E. Pérez, D. Bautista, F. Acevedo, and J. Pimentel, “Programación Genética Aplicada al Pronóstico de Viento en la Región del Istmo de Tehuantepec.,” *2° Congr. Int. Energías Renov.*, vol. 1, pp. 1–11, 2016, [Online]. Available: http://www.unistmo.edu.mx/~neto_144/papers/CIER_2016_Articulo.pdf.

[38] J. Garcia, D. Alvarez, and S. Rivera, “Ensemble based optimization for electric demand forecast: Genetic programming and heuristic algorithms,” *Rev. Int. Metod. Numer. para Calc. y Disen. en Ing.*, vol. 36, no. 3, pp. 1–12, 2020, doi: 10.23967/j.rimni.2020.07.001.

[39] S. Manrique-Naranjo, M. Guzman, and S. Rodriguez, “Hybrid inference algorithm by combining genetic programming methods and nonlinear regression techniques,” no. June, 2018, doi: 10.17654/ETAI070010001.

[40] A. Patelli, *Genetic programming techniques for nonlinear systems identification*, Rum Ed. 2011.

[41] J. R. Koza, “Genetic programming as a means for programming computers by natural selection,” *Stat. Comput.*, vol. 4, no. 2, pp. 87–112, 1994, doi: 10.1007/BF00175355.

[42] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press, 2004.

[43] S. Lumbreras and A. Ramos, “Offshore wind farm electrical design: a review,” *Wind Energy*, vol. 16, no. 3, pp. 459–473, Apr. 2013, doi: https://doi.org/10.1002/we.1498.

[44] J. Tambke, M. Lange, U. Focken, J. O. Wolff, and J. A. T. Bye, “Forecasting offshore wind speeds above the North Sea,” *Wind Energy*, vol. 8, no. 1, pp. 3–16, 2005, doi: 10.1002/we.140.

[45] J. Liu, X. Wang, and Y. Lu, “A novel hybrid methodology for short-term wind power forecasting based on adaptive neuro-fuzzy inference system,” *Renew. Energy*, vol. 103, pp. 620–629, 2017, doi:

10.1016/j.renene.2016.10.074.