Universidad Andina Simón Bolívar

Sede Ecuador

Área de Estudios Sociales y Globales

Maestría en Cambio Climático y Negociación Ambiental

Análisis del desempeño de redes neuronales artificiales en la reconstrucción de datos pluviométricos de la ciudad de Quito

Gustavo Ricardo Egüez Dávila

Tutor: Henry Paz

Quito, 2020

Cláusula de cesión de derecho de publicación

Yo, Gustavo Ricardo Egüez Dávila, autor de la tesis intitulada “Análisis del desempeño de redes neuronales artificiales en la reconstrucción de datos pluviométricos de la ciudad de Quito”, mediante el presente documento dejo constancia de que la obra es de mi exclusiva autoría y producción, que la he elaborado para cumplir con uno de los requisitos previos para la obtención del título de Magíster en Cambio Climático y Negociación Ambiental en la Universidad Andina Simón Bolívar, Sede Ecuador.

1. Cedo a la Universidad Andina Simón Bolívar, Sede Ecuador, los derechos exclusivos de reproducción, comunicación pública, distribución y divulgación, durante 36 meses a partir de mi graduación, pudiendo por lo tanto la Universidad, utilizar y usar esta obra por cualquier medio conocido o por conocer, siempre y cuando no se lo haga para obtener beneficio económico. Esta autorización incluye la reproducción total o parcial en los formatos virtual, electrónico, digital, óptico, como usos en red local y en internet.
2. Declaro que en caso de presentarse cualquier reclamación de parte de terceros respecto de los derechos de autor/a de la obra antes referida, yo asumiré toda responsabilidad frente a terceros y a la Universidad.
3. En esta fecha entrego a la Secretaría General, el ejemplar respectivo y sus anexos en formato impreso y digital o electrónico.

31 de enero de 2020

Firma: _____

Resumen

Los impactos del cambio climático se determinan utilizando *modelos de impacto*. Estos modelos numéricos se alimentan, entre otras fuentes, con información climática de los últimos treinta años. Idealmente la información ha sido medida con la ayuda de estaciones meteorológicas y registrada a lo largo del tiempo, sin embargo, existen muchas localidades donde no se ha podido registrar la información desde hace treinta años o no se ha medido y registrado del todo.

Para estos lugares sin información es necesario utilizar otros mecanismos, entre los más comunes están: bases de datos que han pasado por un proceso de aumento de escala, salidas de modelos de circulación general e información obtenida por sensores remotos. El presente trabajo propone otro mecanismo, el uso de redes neuronales artificiales para reconstruir los datos del pasado tomando como referencia los datos de otras estaciones.

Se crearon dos redes neuronales, una red profunda y una red concurrente y dos modelos estadísticos, un modelo SARIMA y una regresión lineal. Se generaron datos aproximados de las estaciones M0003 y M0025, estaciones que pertenecen al Instituto Nacional de Meteorología e Hidrología. Se evaluó el desempeño de las redes neuronales y la pertinencia del uso de los datos reconstruidos en modelos de impacto del cambio climático.

La red neuronal concurrente y el modelo estadístico SARIMA tuvieron unos indicadores de desempeño más bajos que el modelo de regresión lineal y la red neuronal profunda. Se concluyó que el proceso de reconstrucción de datos se asemeja más a un proceso de regresión que a un proceso de series temporales. La correlación para el modelo SARIMA y para la red neuronal concurrente fue menor a 0.8, por lo tanto, esta información no es apta para el uso en modelos de impacto al cambio climático.

El modelo de regresión y la red neuronal profunda cumplen los requerimientos que permiten incorporar los datos reconstruidos en un modelo de impacto al cambio climático.

Palabras Clave:

Redes neuronales, inteligencia artificial, reconstrucción de datos, modelos de impacto, lluvia, LSTM.

A mi familia, mi soporte, mi apoyo, mi todo.

Agradecimientos

Mis más sinceros agradecimientos a mi tutor, Henry Paz, sin su orientación en una materia tan intricada este trabajo no hubiera sido posible. Al Instituto Nacional de Meteorología e Hidrología por haber facilitado los datos de lluvia de varias estaciones a lo largo del territorio nacional, su aporte fue insumo fundamental para el desarrollo de esta tesis. A Jorge Núñez, Luis Maisincho y Guillermo Armenta, por haber compartido su experiencia en temas relevantes para este trabajo, su aporte ayudó a complementar y explicar los resultados de este trabajo.

Tabla de contenidos

Figuras y tablas

Introducción

Tradicionalmente, al desarrollar un sistema, un programador escribe instrucciones que implementan un algoritmo en un ordenador con el objetivo de automatizar una tarea de la que se tiene pleno conocimiento.

Existen tareas complejas de las cuales es difícil o imposible obtener un algoritmo que permita automatizarlas, por ejemplo, diferenciar si una imagen corresponde a un animal o a un objeto, pronosticar enfermedades en base a un historial de tomografías o predecir los eventos de lluvia, etc. (Barr 1982, xi)

El principal potencial de la inteligencia artificial es la automatización de estos problemas complejos. En lugar de desarrollar un algoritmo, un programador escribe instrucciones que permiten a un ordenador aprender de datos existentes, para luego, en base a lo aprendido, automatizar una tarea con un porcentaje de exactitud (Barr 1982, xi). Esta técnica se denomina *aprendizaje automático*.

Las redes neuronales, que son la herramienta utilizada en el presente trabajo, son una de las implementaciones de algoritmos que utilizan el aprendizaje automático para gestionar problemas complejos.

Un caso de éxito de la aplicación de redes neuronales es el reconocimiento de cáncer de piel. Se utilizó una base de datos de imágenes de pacientes de los cuales un subconjunto desarrolló este tipo de cáncer. Se aplicaron algoritmos de aprendizaje automático, implementados por redes neuronales, y al final se logró predecir una condición de cáncer futuro con un 95% de exactitud (Ansari y Sarode 2017, 2880).

Dentro del cambio climático, otros problemas complejos, difíciles de caracterizar desde el punto de vista de las ciencias de la computación, han sido tratados con la ayuda del aprendizaje automático y redes neuronales. La mayor parte de estos problemas se relaciona con la reducción de escala de simulaciones realizadas por modelos matemáticos, la predicción de variables hidrometeorológicas y el consumo de energía. Se han obtenido exactitudes mayores al 90% (Luk, Ball, y Sharma 2001, 692), (Abhishek et al. 2012, 317), (Tapoglou et al. 2014), (Snell, Gopal, y Kaufmann 2000).

Otro problema complejo presentado en el marco del cambio climático es la determinación del impacto de este fenómeno en una población. El impacto generalmente se determina a través de la aplicación de modelos numéricos que se alimentan, entre otras fuentes, de escenarios futuros del clima y tienen como resultado el impacto del cambio

climático en sistemas naturales y en una serie de sectores económicos. A estos modelos se los llama modelos de impacto del cambio en el clima o simplemente modelos de impacto (University of Edinburgh 2020, párr. 3). Es importante recalcar la diferencia entre los modelos de impacto y los modelos de circulación general. Los resultados de la modelación del cambio en el clima futuro producido por los gases de infecto invernadero son utilizados como información de entrada para los primeros.

Los escenarios futuros del clima, se han establecido, desde los años 80, mediante el uso de modelos matemáticos que simulan el comportamiento de la atmósfera, océanos, criosfera y superficie terrestre (IPCC 2020, párr. 2), (Weart 2009, 270). A pesar de varias críticas a estos modelos (modelos de circulación global o *GCM*, por sus siglas en inglés) su constante mejora a lo largo del tiempo ha hecho que sean aceptados como el mejor mecanismo para determinar el clima futuro (IPCC 2020, párr. 2).

Los GCM, a pesar de ser el mejor mecanismo para simular el clima futuro –al momento–, tienen una resolución espacial baja, en el orden de los cientos de kilómetros. La determinación del impacto del cambio climático, a través del uso de modelos de impacto, requiere información con una resolución espacial de aproximadamente 50 kilómetros, por lo tanto, la aplicación directa de los GCM en modelos de impacto no es útil (UNFCCC 2020, 1).

Para sobrellevar este problema se han desarrollado varias metodologías de reducción de escala que permiten transformar los datos de salida de los GCM, que tienen una resolución en el orden de cientos de kilómetros, a datos con una resolución decenas de kilómetros. Existen dos tipos de metodologías de reducción de escala: dinámicas y estadísticas. Las últimas se basan en relaciones estadísticas entre las variables atmosféricas globales de los GCM y variables meteorológicas locales (University of Edinburgh 2020). Para determinar estas relaciones es necesario tener un registro de observaciones de variables meteorológicas, capaces de caracterizar los fenómenos locales, en el periodo de 1960 a 1990 con un ideal de información reciente (IPCC 2017, 15).

Las regiones montañosas a través de todo el mundo, incluida la región en la que se encuentra la ciudad de Quito, se caracterizan por tener una alta variabilidad espacio temporal de precipitación, por lo tanto, para implementar una metodología de reducción de escala de los datos de salida de los GCM, se necesita observaciones que puedan caracterizar esta variabilidad. Las estaciones que conforman la red meteorológica del Ecuador fueron ubicadas antes de que el cambio climático fuera estudiado y expuesto a

la comunidad (MAE 2020, 24), es únicamente a partir del año 2004 que se empiezan a instalar estaciones automáticas en el país con objetivos de estudio de cambio climático (Maisincho 2019, entrevista personal). La alta variación de lluvia y la situación actual de la red de estaciones dificultan la caracterización de los procesos locales y por lo tanto la ejecución de metodologías de reducción de escala (Campozano et al. 2016, 1).

En casos en los que no hay datos locales disponibles se han usado bases de datos que ya han pasado por un proceso de reanálisis, en el que se reduce la escala y se interpolan los datos mediante la aplicación de una combinación de modelos y observaciones de estaciones metereológicas y satelitales. Ejemplos de estas bases de datos son WorldClim, NCEP Reanalysis o NCEP Reanalysis 2. WorldClim utiliza una metodología de interpolación de alta resolución (Hijmans et al. 2005, 1971) y por su lado Reanalysis utiliza su propio sistema de análisis y predicción (NOAA 2020, párr. 1).

Sin embargo, todo modelo debe validarse con información generada in situ, sobre todo en geografías tan complejas como la ecuatoriana. Lastimosamente, en el país, la cantidad de información meteorológica representativa y completa es escasa, por lo que es necesario buscar alternativas para completar estos datos.

Una alternativa a las bases de datos existentes es la utilización de redes neuronales para generar datos aproximados de lluvia del pasado. Como se describió anteriormente, una serie de estudios han utilizado redes neuronales en sistemas climáticos, demostrando que el uso de esta herramienta produce resultados con una exactitud mayor al 90%.

El presente trabajo propone el análisis del desempeño de redes neuronales artificiales (*ANN*, por sus siglas en inglés) para la generación de datos pluviométricos aproximados del pasado de la ciudad de Quito tomando como base mediciones en el rango de 1970 a 2018 de pluviómetros existentes. De esta manera, se podría disponer de información que permita ejecutar procesos estadísticos de reducción de escala de los resultados de los modelos globales de circulación para poder utilizar sus resultados como insumo para los modelos de impacto.

Inicialmente la investigación se centró en la ciudad de Quito debido a que existen estaciones pluviométricas con un periodo de observaciones que permiten la comparación de las mediciones reconstruidas y las reales y al mismo tiempo la construcción de una red neuronal artificial. Luego de una inspección a los datos recolectados se determinó que es posible realizar el mismo ejercicio para otro par de estaciones ubicadas en la costa del país.

El objetivo principal del estudio es determinar el desempeño del uso de redes neuronales en la generación aproximada de datos pluviométricos en la ciudad de Quito. Cuatro objetivos secundarios se desprenden del objetivo principal: el primero, crear y entrenar una red neuronal capaz de reproducir los datos de una estación pluviométrica en base a los datos de otra, el segundo, generar los datos del pasado de una de las estaciones pluviométricas, el tercero, calcular la exactitud de los datos reconstruidos para analizar el desempeño de la red neuronal y compararlos con los datos disponibles en WorldClim y, el cuarto, evaluar el desempeño del uso de redes neuronales para la reconstrucción de datos en los modelos de impacto.

El trabajo realizado por Tran Anh Duong et. al. (2018) establece una metodología para el trabajo con redes neuronales y datos meteorológicos. Tomando como base esta metodología se crearon dos tipos diferentes de redes neuronales para reconstruir los datos con el fin de poder comparar su desempeño, de igual manera se reconstruyeron los datos con dos métodos estadísticos para disponer de información para comparar los esfuerzos de reconstrucción de datos. Posteriormente se compararon los esfuerzos de reconstrucción con los datos existentes en WorldClim.

Existe una serie de artículos que abordan el uso de redes neuronales en procesos climáticos, entre ellos se ha seleccionado los trabajos realizados por Tran Anh Duong et. al. (2018), Kumar Abhishek et. al. (2012) y (Snell, Gopal, y Kaufmann 2000) como fundamento para este escrito. De igual manera, se han realizado entrevistas a expertos, involucrados en el desarrollo de proyectos relacionados con el cambio climático, para analizar la pertinencia del esfuerzo de generación de datos aproximados.

La tesis está dividida en cuatro capítulos. En el primero, el marco teórico, se establecen las perspectivas desde las cuales aborda el problema de investigación, se enuncian una serie de conceptos relevantes al tema y se resume la producción científica actual relevante. En el segundo capítulo, generación de datos de pasado, se crean dos redes neuronales y dos modelos estadísticos para generar datos aproximados del pasado y se determina el desempeño de estos para una estación de Quito y otra de la costa. En el tercer capítulo, análisis, se comparan los resultados obtenidos con bases de datos existentes y se caracteriza la pertinencia del uso de redes neuronales para la reconstrucción de datos. Finalmente, en el apartado de conclusiones y recomendaciones, se contrastan los objetivos del presente estudio con los resultados analizados, de igual manera, se enuncian posibles caminos que extiendan el trabajo realizado en la tesis.

Capítulo primero
Marco Teórico

1. Perspectiva

El presente trabajo de investigación se abordará desde dos disciplinas: inteligencia artificial y estadística y se lo contextualizará en la adaptación al cambio climático. La Inteligencia Artificial permitirá guiar la creación de una red neuronal con el fin de reconstruir los datos y la Estadística, por su lado, se utilizará para crear dos modelos que permitan contrastar los datos reconstruidos y para determinar la pertinencia del uso de redes neuronales.

2. Conceptos relevantes

Inteligencia Artificial

La inteligencia artificial es una disciplina que puede ser analizada desde diferentes ciencias como la filosofía, matemáticas, economía, neurociencia, ingeniería computacional, etc. Cada una de las disciplinas mencionadas analiza un sistema artificial, lo categoriza y propone diferentes perspectivas y cuestiones (Ceccaroni 2007, 4). Desde el punto de vista de la ingeniería computacional, la inteligencia artificial es el conjunto de hardware y software dispuesto de tal manera que imite el pensamiento o la manera de actuar del ser humano (Ceccaroni 2007, 7).

Alan Turing, quien es considerado el padre de la computación y la inteligencia artificial, creó una prueba que permite catalogar a cualquier sistema artificial como inteligente. En la prueba de Turing, si existen dos sistemas, uno manipulado por un humano y otro por un sistema artificial y un tercer humano interactúa, a través de un teclado, con los dos sistemas y no es capaz de distinguir cuál de los dos sistemas es operado por un humano entonces se puede catalogar al sistema artificial como inteligencia artificial (Gomez 2001, 5).

Interpretar las preguntas hechas por un humano a través de un teclado y responder en concordancia es un proceso profundo e intrincado, por ejemplo, existen varias oraciones y peticiones diferentes que tienen el mismo objetivo, es decir, expresan el

mismo deseo. Un sistema artificial debe ser capaz de discernirlas y procesarlas. A este tipo de proceso profundos e intrincados se los llama procesos complejos y se podría decir que son aquellos que son imposibles o muy difíciles de modelar a través de un algoritmo (Matich 2001, 4).

La inteligencia artificial ha demostrado ser una herramienta potente para modelar sistemas complejos y pronosticar o replicar su comportamiento (Ceccaroni 2007). Uno de los procesos más complicados en la naturaleza, el cáncer, no ha sido caracterizado y no puede ser explicado, por lo tanto, es difícil generar una cura o detectarlo de manera temprana. Existen estudios que hacen uso de la inteligencia artificial para detectar el cáncer de mama en etapas tempranas, estos estudios no se han preocupado por caracterizar o estudiar el funcionamiento de esta enfermedad, en su lugar, usan metodologías de inteligencia artificial para determinar si un paciente tiene o no y en qué probabilidad un cáncer temprano de mama.

Aprendizaje de máquina

El aprendizaje de máquina es el mecanismo con el que se logra que un sistema tenga inteligencia artificial. Los algoritmos de aprendizaje de máquina se auto configuran sin intervención del humano. Ejemplos de aprendizaje de máquina son algoritmos que configuran automáticamente una regresión lineal, algoritmos genéticos o algoritmos que implementan redes neuronales (Alpaydin 2014, 1). Al aprendizaje de máquina se lo puede categorizar, respecto de la información provista al sistema en el momento de entrenamiento, en tres categorías: aprendizaje no supervisado, aprendizaje por refuerzos y aprendizaje supervisado.

Aprendizaje no supervisado

Existen procesos de aprendizaje del ser humano en los que se categorizan y agrupan una serie de objetos dependiendo sus cualidades, por ejemplo, si a un niño se le entregan figuras geométricas él puede aprender a categorizarlas y agruparlas por su forma, su color, material, etc. También es posible configurar una red neuronal artificial para que aprenda a categorizar y agrupar los datos con los que se la alimenta. A este proceso se lo llama aprendizaje no supervisado (Matich 2001).

El aprendizaje no supervisado no tiene una salida esperada, su objetivo es encontrar características y agrupar los datos por las características descubiertas (Hinton y Sejnowski 1999, 3). Por ejemplo, si a una red se la entrena con imágenes de perros y

caballos aprende que ha visto animales diferentes, no sabe cómo se llaman, pero sabe que son diferentes y los puede categorizar. La próxima vez que a la red se le presente la imagen de un caballo que no ha visto antes esta será capaz de ponerla en la categoría de estos animales.

Aprendizaje por refuerzo

En el aprendizaje por refuerzo trabaja bajo un esquema de estados y acciones. Un sujeto podría cambiar de estado ejecutando una acción, por ejemplo, un gato podría estar sentado y cambiar de estado a caminando al ejecutar la acción de caminar. El cambio de estado supone una recompensa que puede ser positiva o negativa, si es positiva, la próxima vez que el sujeto se encuentre en una situación parecida tenderá a replicar la misma acción, por el contrario, si la recompensa es negativa el sujeto tenderá a ejecutar otras acciones para así maximizar la recompensa global (Alpaydin 2014, 518).

Aprendizaje supervisado

Como se mencionó anteriormente, la disposición de neuronas conectadas entre sí permite al cerebro entender el comportamiento de un sistema complejo sin conocer los detalles a fondo. En lugar de entender el sistema, el cerebro aprende, en base a la experiencia, como dicho sistema se comporta. Por ejemplo, cuando un niño está aprendiendo a atrapar un balón, tiene que interpretar la trayectoria del objeto y la posición de sus manos, al inicio su interpretación es torpe y los resultados erróneos, pero a medida que practica, intento tras intento, corrigiendo los errores del pasado, su interpretación mejora hasta lograr un buen resultado.

De igual manera, un sistema de inteligencia artificial necesita aprender de la experiencia, para lograrlo, se ejecuta un procedimiento de entrenamiento en el que se presentan valores conocidos para las entradas y sus respectivas salidas. Al inicio la respuesta del sistema es torpe, pero a medida que se va ajustando la salida se vuelve más precisa (Matich 2001), (Alpaydin 2014, 21).

El proceso de aprendizaje es cíclico, se presentan ejemplos de salidas para un arreglo de entradas (Alpaydin 2014, 22) y se repite el proceso, múltiples veces. Por ejemplo, si se desea crear un sistema artificial capaz distinguir perros de gatos, secuencialmente se ingresan al sistema imágenes de perros y gatos y se le instruye el tipo de animal que representa cada imagen.

Redes Neuronales

Las redes neuronales son un tipo de implementación de la inteligencia artificial que busca imitar el pensamiento humano a través de la replicación de la anatomía del cerebro (Matich 2001, 3). Según la anatomía, el cerebro está compuesto por células llamadas neuronas que están interconectadas entre sí. Cada una de ellas recibe impulsos a través de las dendritas y, de acuerdo con la intensidad de los impulsos, produce una señal química hacia otras neuronas a través del axón. Se observó que esta configuración permite a los humanos caracterizar sistemas complejos tomando como referencia la experiencia en lugar de la explicación técnica del fenómeno (Matich 2001, 5).

Las redes neuronales artificiales, (*ANN*, por sus siglas en inglés) son modelos matemáticos de una neurona que pueden ser implementados a través de software en dispositivos de hardware (Matich 2001, 5). En la Figura 1, se puede observar una representación gráfica de una neurona artificial.

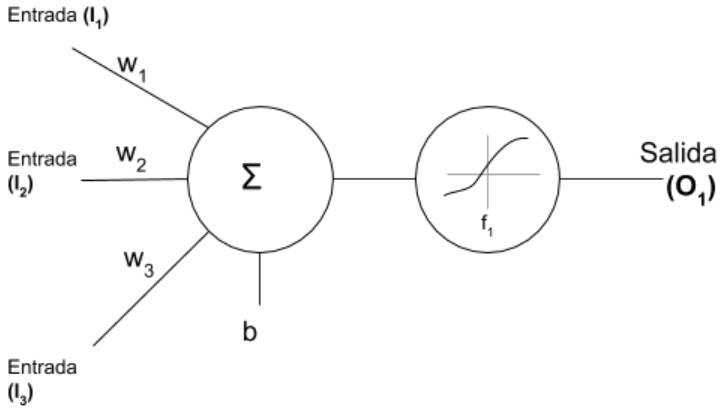


Figura 1. Neurona artificial

El modelo matemático asigna un peso a cada entrada de la neurona, ejecuta una sumatoria de las señales y su peso y al final produce una salida cuyo nivel depende de una función de activación especificada también al momento de implementar la neurona (Alpaydin 2014, 36). En la ecuación (1) se puede observar la salida para la neurona de la Figura 1. Si se generaliza la ecuación (1), para una neurona con un número diverso de entradas (*n*) se obtiene la ecuación (2).

$$O_1 = f_1(w_1.I_1 + w_2.I_2 + w_3.I_3 + b) \quad (1)$$

$$o = f(\sum_{i=1}^{n} w_i.I_i + b) \quad (2)$$

El mecanismo de entrenamiento más común en redes neuronales es el aprendizaje supervisado. Si se observa la red neuronal de la Figura 2, se puede apreciar que las entradas se propagan a través de las diferentes capaz de izquierda a derecha, por esta razón la red neuronal se denomina *red neuronal de avance*. De manera inversa, cuando se entrena la red neuronal, el error se propaga de derecha a izquierda, este mecanismo de propagación se llama *propagación hacia atrás.*

El mecanismo de propagación hacia atrás funciona de la siguiente manera: 1) Se presentan las entradas a la red neuronal y se asignan valores aleatorios para los pesos de cada conexión entre neuronas, por lo que se obtiene un valor para la salida. 2) Se compara el valor obtenido con el valor real, se obtiene un error. 3) El error se distribuye a las neuronas de derecha a izquierda. 4) se ajustan los pesos tomando en cuenta como información el error y 5) se repite el proceso para la siguiente muestra de datos (Hecht-Nielsen 2014, 70).

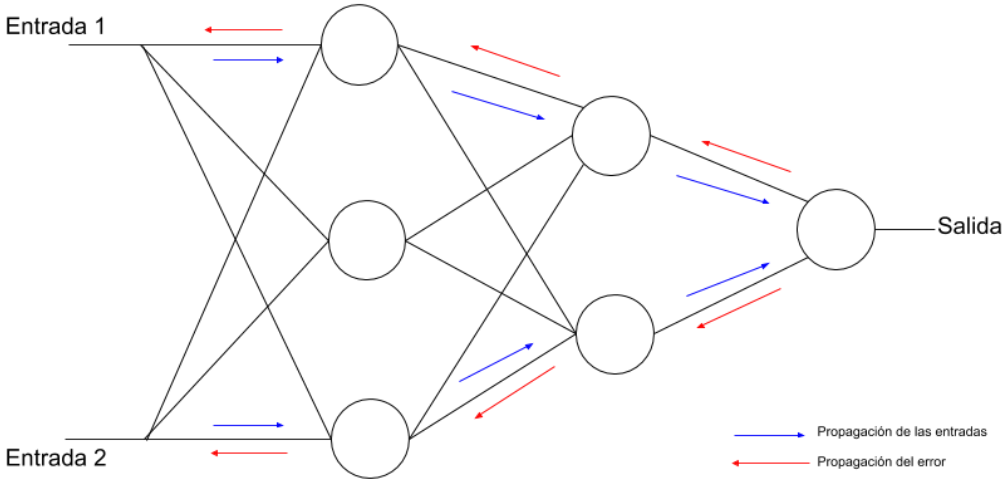


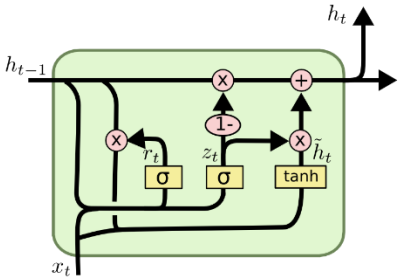
Figura 2. Red neuronal profunda (DNN)

Las redes neuronales que trabajan con aprendizaje supervisado a su vez se pueden clasificar en: redes neuronales profundas (*DNN*, por sus siglas en inglés) redes neuronales recurrentes y redes neuronales convolucionales.

Las redes neuronales profundas, como ya se ha explicado, son redes que tienen múltiples capas, los datos fluyen desde las entradas a través de las capas hasta la salida, de igual manera, la propagación del error sucede desde la salida hacia la entrada (Information Resources Management Association 2019, 31). Para que una red neuronal sea considerada profunda debe tener al menos dos capas intermedias, es decir más de tres contando entradas y salidas (Alpaydin 2014, 307).

En cada capa la red aprende a reconocer una característica de las entradas, por ejemplo, si se dispone una red, con tres capas internas, y se la entrena con imágenes de rostros humanos la primera capa será capaz de reconocer líneas, la segunda capa será capaz de reconocer ojos, narices y boca y la tercera capa será capaz de reconocer un rostro humano. Esta capacidad se llama jerarquía de atributos y es el mecanismo a través del cual las redes neuronales son capaces de manejar gran cantidad de atributos multidimensionales que pasan pon funciones no lineales (Alpaydin 2014, 294).

Por otro lado, existen otro tipo de procesos en los que el resultado es una combinación de las entradas actuales y lo sucedido en el pasado, en este caso las redes neuronales tradicionales pierden su efectividad porque no tienen manera de saber lo que sucedió en el pasado. Las redes neuronales recurrentes se caracterizan por tener memoria, logran este efecto al retroalimentar la salida de una neurona a la entrada de esta, en la Figura 3 se puede apreciar la configuración de una neurona recurrente.



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Figura 3. Neurona de una red recurrente
Elaboración: https://colah.github.io

Estudios realizados por Bengio et. al. (“Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies” 2009, 6) establecieron un fenómeno en el que la memoria se va perdiendo a medida que pasa por las capas de una red neuronal recurrente, el fenómeno se denominó el problema de las dependencias de largo plazo. Para resolver este problema, en 1997, Hochreiter & Schmidhuber (“Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies” 2009, 6) introdujeron el concepto de redes neuronales de memoria larga a corto plazo o LSTM (*Long short-term memory*, en inglés).

Una unidad de una red LSTM no es una neurona, en su lugar es un conjunto de redes neuronales interconectadas de manera especial para sobrellevar el problema de las dependencias de largo plazo. Una red LSTM no lucha con el problema de memoria de

largo plazo, fue diseñada para recordar las dependencias de algún tiempo en el pasado (Olah 2015, párr. 16).

Aplicaciones de las redes neuronales

Las redes neuronales han sido aplicadas prácticamente en cada proceso complejo que se ha detallado. El ejemplo más sencillo de aplicación de las redes neuronales artificiales es la clasificación entre imágenes de gatos y perros. Puede parecer que clasificar estas imágenes sea un proceso sencillo, pero no lo es, un sistema automático tiene que aprender diferenciar un perro de un gato cuando ambos tienen una cabeza, orejas, cuatro patas, una cola, etc. (Parkhi et al. 2020, 1).

Otro ejemplo práctico de aplicación de redes neuronales es la detección temprana de cáncer de mama. El cáncer es una enfermedad de la que la humanidad aún no tiene el conocimiento completo por lo tanto su detección temprana o cura es un proceso complejo (AlShehri y Khatun 2009, 79). Los mecanismos actuales para detección de cáncer tienen un porcentaje de falla de alrededor de 30% (AlShehri y Khatun 2009, 79), un porcentaje bastante alto. Un estudio realizado por el Departamento de Computación de la Facultad de Ingeniería de la Universidad de Malasia utilizó redes neuronales para determinar si un paciente tiene cáncer de mama en etapa temprana. El estudio logró determinar la presencia y localización de los tumores con un 94% de exactitud para un modelo matemático de mama (AlShehri y Khatun 2009, 88).

El cambio climático es considerado también un problema complejo ya que requiere la caracterización de varios sistemas de los cuales es difícil generar algoritmos o modelos numéricos (Flato, Marotzke, y Abiodun 2013, 746). Era obvia entonces la proliferación en el uso de redes neuronales dentro de esta disciplina.

En un estudio desarrollado en Portugal (Trigo y Palutikof 1999), se buscaba reproducir la tendencia de las anomalías de temperatura para una área específica. En un primer intento se utilizaron las salidas de modelos de circulación general y se determinó que estos eran incapaces de reproducir los fenómenos locales. Era entonces necesario reducir la escala de las salidas de los modelos de circulación general para describir los fenómenos locales. Este proceso se llevó a cabo usando una regresión lineal y a la par con redes neuronales. Al final se pudo probar que las redes neuronales obtuvieron mayor desempeño relativo a los mecanismos lineales utilizados.

El pronóstico del tiempo es también un proceso complejo ya que modelar el comportamiento de la atmósfera requiere de gran conocimiento a nivel local y de gran

capacidad de procesamiento (Hirani y Mishra 2016, 1). Se han utilizado diferentes configuraciones de redes neuronales para pronóstico de algunas variables meteorológicas con diversa exactitud en los resultados. Uno de los esfuerzos más importantes es el estudio realizado por la Universidad de Munich sobre la predicción de lluvia en la ciudad de Camu en Vietnam (Tran Anh, Bui, y Rutschmann 2018). Se demostró que era posible pronosticar la lluvia con un coeficiente de correlación de 0.98 para 80 días en el futuro.

Generalización y overfitting

Las redes neuronales que son entrenadas con aprendizaje supervisado y aprenden las características de un proceso a partir de la información de los datos son un ejemplo de aprendizaje inductivo. Al contrario del aprendizaje deductivo, donde un concepto general es aplicado a un caso particular, en el aprendizaje inductivo se intentan extraer generalidades a partir de ejemplos o casos específicos (S. Lawrence y C. L. Giles 2000, 114).

El objetivo de una red neuronal es lograr la generalización, una vez que se ha entrenado la red con datos representativos, esta debe ser capaz de producir resultados satisfactorios para datos que no ha visto antes (S. Lawrence y C. L. Giles 2000, 114).

Existen configuraciones de parámetros de una red neuronal que hacen que la misma aprenda con demasiado detalle, incluso hasta el ruido, de los datos de entrenamiento (S. Lawrence y C. L. Giles 2000, 116). Al momento de evaluar la red neuronal la misma se presenta con un alto porcentaje de efectividad, pero es incapaz de producir resultados satisfactorios para datos que no ha visto antes, es incapaz de generalizar. Este problema se conoce como *overfitting* o sobre entrenamiento.

Conceptos relevantes de la Estadística

Debido a que el presente trabajo pretende reconstruir los datos de lluvia de una estación tomando como referencia los datos de otra es importante analizar dos conceptos estadísticos: la correlación y la causalidad.

Correlación

El concepto de correlación cruzada o correlación entre dos variables no es un enunciado complicado. En su forma más básica la correlación es una medición del grado de interrelación de dos variables (Kenny 1979, 23). La correlación es un coeficiente que puede ser positivo, mostrando que cuando una variable crece en valor la otra también,

puede ser negativo, mostrando que cuando una variable crece la otra decrece o puede tender a cero estableciendo entonces que no existe correlación entre dos variables. En la Figura 4 se muestra como lucen los diferentes tipos de correlación.

Figura 4. Tipos de correlación.

Elaboración: https://blogs.ugto.mx/enfermeriaenlinea/unidad-didactica-5-correlacion-y-regresion/

Causalidad

La causalidad es un concepto un poco más profundo que la correlación. El ser humano siempre tiende a querer expresar lo que le pasa o pasará con oraciones tipo: me pasó X porque Y o si X ha cambiado entonces Y cambiará. Por lo tanto la causalidad, como su nombre lo indica es el hecho que un evento sea producido por una causa (Brady 2011, párr. 2). En términos generales implica que unos cambios en una variable afecten directamente a la otra o a su vez, que cambios en una tercera variable, global a las otras dos, produzca cambios en ellas. Para que dos variables sean causales tienen que cumplir cuatro requisitos: tiene que observarse correlación entre las dos variables, tiene que existir una ausencia de efecto cuando se suspende la causa, tiene que haber un cambio en el efecto cuando hay un cambio en la causa y tiene que existir un mecanismo o efecto global que relacione la causa con el efecto (Kenny 1979, 13).

Autocorrelación

La autocorrelación, de igual manera que la correlación, es un coeficiente que expresa el grado de interrelación entre dos medidas. En el caso de la correlación, se expresa la relación entre dos variables diferentes. En el caso de la autocorrelación expresa la relación entre una variable medida en el presente y la misma variable medida en el pasado (N.I.S.T 2020).

Modelo estadístico

Un modelo estadístico es una representación matemática de un sistema. Cumple con ciertos requerimientos respecto de sus datos, por ejemplo, estos tienen que representar un porcentaje de toda la población (McCullagh 2002, 1225). Su función principal es interpretar o estimar predicciones de una variable en función de un segmento de datos de la población. El modelo SARIMA y de regresión lineal son dos modelos relevantes al presente estudio que se detallarán a continuación.

Una regresión lineal es la representación lineal matemática entre dos o más variables. El objetivo es predecir un valor para una variable que no se encuentra en la muestra (Yale 2020, párr. 2). Cabe recalcar que antes de realizar una regresión lineal es determinante establecer si las variables tienen un grado de correlación. De no existir correlación la regresión no será determinante.

El modelo SARIMA permite caracterizar procesos llamados series temporales. Una serie temporal es un conjunto de datos que se encuentran auto correlacionados en el tiempo, es decir, los valores de la variable en el pasado tienen un coeficiente de correlación significante con los valores presentes y de igual manera los valores presentes tienen correlación con los valores futuros (NIST 2020, párr. 1).

En su forma más sencilla un modelo SARIMA se transforma en un modelo ARMA (*Auto Regresive Moving Averge*, en inglés). Un modelo arma expresa una serie temporal como una función de sus retardos o valores en el pasado (NIST 2020). Los procesos que son series temporales frecuentemente presentan una tendencia, sea al crecimiento o decrecimiento. En estos casos se añade una componente de integración *I* al modelo arma para poder modelar esta tendencia, el modelo se convierte en un modelo ARIMA. Existen procesos que también presentan estacionalidad, es decir se repiten cada cierto tiempo. Para poder modelar estos procesos se introducen componentes estacionales componiendo así el modelo SARIMA (De la Fuente, s. f., párr. 2). Un modelo SARIMA se expresa como en la ecuación 3, donde *AR* es el índice del componente de auto regresión, *I* es el índice del componente de integración, *MA* es el componente de media móvil y *ARs*, *Is* y *MAs* son los componentes estacionales del modelo.

$$SARIMA = (AR, I, MA)(AR_s, I_s, MA_s) \quad (3)$$

3. Producción científica relevante en redes neuronales aplicadas a la climatología

En la literatura existe una serie de estudios en diferentes ámbitos de la inteligencia artificial y los sistemas climáticos: pronóstico de lluvia, pronóstico de temperatura, mejoramiento de escala de salidas de modelos de circulación general, etc. A continuación, se describen los análisis más relevantes para el presente trabajo de investigación.

Se puede considerar como punto de partida al trabajo realizado por Ch. Jyosthna Devi et. al. (2012). Ellos ejecutaron un esfuerzo de comparación del pronóstico de temperatura utilizando una red neuronal profunda y una red con retroalimentación. Obtuvieron datos del portal *https://www.wunderground.com/*. Los datos contenían mediciones de presión atmosférica, humedad relativa, velocidad y dirección de viento, etc. Lo interesante de este trabajo es que se determina que el error, al utilizar una red neuronal retroalimentada, es menor que el obtenido al usar de una que no tiene retroalimentación. Concretamente, el error de los datos normalizados, es decir llevados a una escala de cero a uno, se redujo de 0.19 a 0.182. Este artículo constituye una base de conocimiento que apoya la idea de la generación de datos del pasado utilizando redes neuronales con retroalimentación en lugar de redes que no tienen un mecanismo similar.

Otro trabajo relevante para el desarrollo de este estudio es el llevado a cabo por Kumar Abhishek et. al. (2012) titulado *Weather forecasting model using Artificial Neural Network*. Ellos ejecutan un esfuerzo de predicción de temperatura para la estación Toronto Lester B. Pearson Int'l A, Ontario, Canadá. El objetivo principal del artículo es determinar la manera en que la variación de los parámetros de una red neuronal afecta el rendimiento de esta al predecir 365 días de temperatura. Luego de evaluar varias configuraciones se concluye que el aumentar el número de neuronas por cada capa de la red neuronal aumenta el desempeño (reduce el error cuadrático medio). También se concluye que el número de muestras afecta el desempeño de la red, si se entrena la red con más muestras, menor es el error cuadrático medio, por ende, mayor el desempeño. Finalmente, se establece la metodología para determinar si la ANN sufre de *overfitting*. Las conclusiones sobre la variación en los parámetros de una red y la metodología de análisis del error cuadrático medio respecto del *overfitting* inspiran la metodología seguida en este trabajo.

Otro trabajo que merece la pena mencionar es el realizado por Tran Anh Duong, et. al. (2018), titulado *Long Short-Term Memory for Monthly Rainfall Prediction in*

Camu, Vietnam. En este artículo se realiza una comparación de tres configuraciones de redes neuronales para la predicción de lluvia. Para llevar a cabo la predicción se utilizaron datos mensuales de los últimos 30 años.

Las redes neuronales profundas que se compraron son: red neuronal profunda clásica, una variación de esta, la red neuronal profunda con estacionalidad y una red LSTM. En el estudio se varía la cantidad de neuronas o elementos de memoria según sea el caso y la cantidad de veces que se ha presentado los datos de entrenamiento a las distintas redes neuronales (*epochs*).

Los autores del artículo, tras realizar un análisis bibliográfico, establecen que para medir el desempeño estadístico de una red neuronal es necesario contar con una medición de la calidad del modelo y una de error. Ellos utilizaron, como medición de calidad, el coeficiente de determinación y como mediciones de error el error cuadrático medio y el error absoluto medio. Los resultados del estudio demostraron que el desempeño de la red neuronal LSTM (R = 0.9651) fue superior a las otras dos configuraciones propuestas (0.7340 y 0.7248). El mecanismo de evaluación de desempeño antes mencionado es el utilizado en esta tesis.

Discusión

Las redes neuronales han crecido rápidamente en los últimos años en gran parte debido al aumento de estudios realizados en torno al tema. Al existir más estudio del tema, los autores generan diversas opiniones, unas a favor del uso de estas redes y otras enfatizando sus puntos débiles. La crítica más fuerte, en cuanto al uso de redes neuronales en el pronóstico climático, es la incapacidad de estas para reproducir fielmente procesos no lineales (Tealab et. al. 2017). La precipitación es considerada un proceso complejo no lineal (Roux et. al. 2009) y, por ende, según esta línea de pensamiento, la aplicación de las redes neuronales para el pronóstico de precipitación no tendría mucha pertinencia. Sin embargo, la administración nacional oceanográfica y atmosférica de Estados Unidos (NOAA por sus siglas en inglés), entre otros institutos de diferentes países, ha venido explorando y aplicando las redes neuronales en pronóstico de la precipitación desde 1997 (NOAA 2020). En general, no se utiliza a las redes neuronales como una herramienta de remplazo de las técnicas actuales sino como un complemento a las mismas (NOAA 2020).

Capítulo segundo
Generación de datos aproximados del pasado

1. Procesamiento de los datos

Como se ha mencionado en párrafos anteriores el principal objetivo del presente trabajo es generar datos aproximados de lluvia del pasado para la ciudad de Quito mediante el uso de redes neuronales artificiales. Los procesos estadísticos que pueden apoyar de mejor manera al presente estudio son los pronósticos. En un pronóstico se genera un valor aproximado para una variable teniendo como referencia información del pasado. En este trabajo se invertirá en forma de espejo el eje x (eje del tiempo) de tal manera que se puedan aplicar los mismos procesos estadísticos del pronóstico para la generación de datos aproximados del pasado. En la Figura 5 se explica la manera en la que el proceso de generación de datos del pasado se lleva de la misma forma que un proceso de pronóstico.

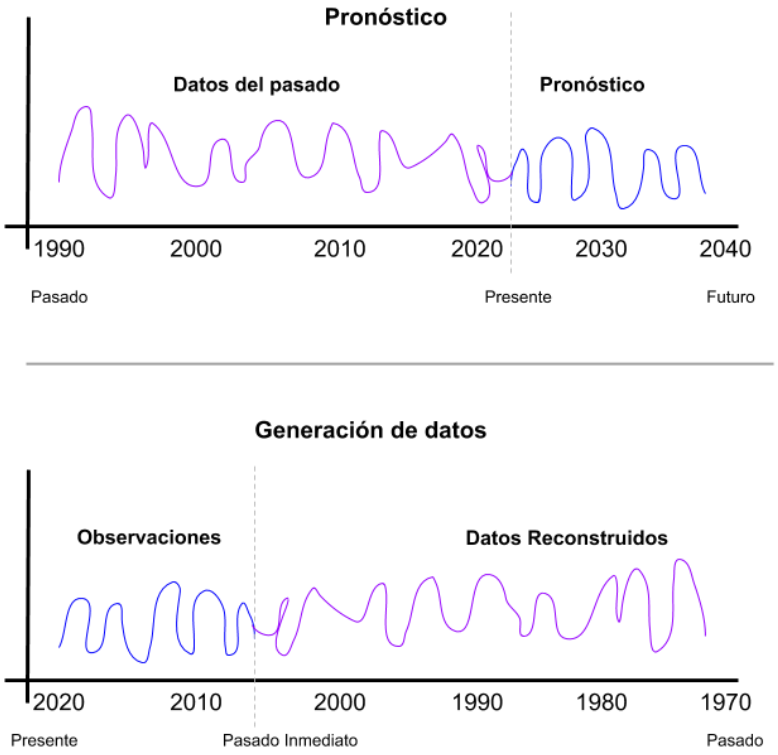


Figura 5. Proceso de reconstrucción vs proceso de pronóstico (Imagen referencial con fines didácticos. Una serie temporal de lluvia luce como la de la Figura 19)

En esta tesis se aplican diferentes modelos para generar datos aproximados del pasado, dos de estos modelos son estadísticos y dos usan redes neuronales. Para cada modelo los pasos a seguir son: entrenamiento del modelo, generación de datos aproximados del pasado y evaluación del desempeño del modelo.

Para el primero paso, entrenamiento del modelo, se utilizan las observaciones desde el pasado inmediato hasta el presente para entrenar el modelo (Figura 6). En el segundo paso, generación de datos aproximados, se utilizan como entrada al modelo las observaciones del pasado (datos desde 1970 hasta el 2008) de la estación de referencia y se obtienen datos aproximados, en el mismo periodo, para la estación objetivo (Figura 7). En el tercer paso, evaluación de desempeño del modelo, se utilizan los datos aproximados generados en el paso 2 y las observaciones existentes de la estación objetivo para calcular la correlación entre las observaciones, el error absoluto promedio y el sesgo de la generación de datos del pasado (*bias*, en inglés), (Figura 8). En la Figura 9 se describe este proceso y los datos utilizados.

Figura 6. Paso 1, entrenamiento del modelo

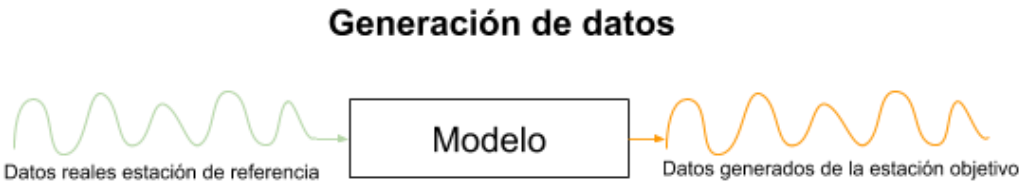


Figura 7. Paso 2, generación de datos aproximados del pasado

Figura 8. Paso 3, evaluación del modelo

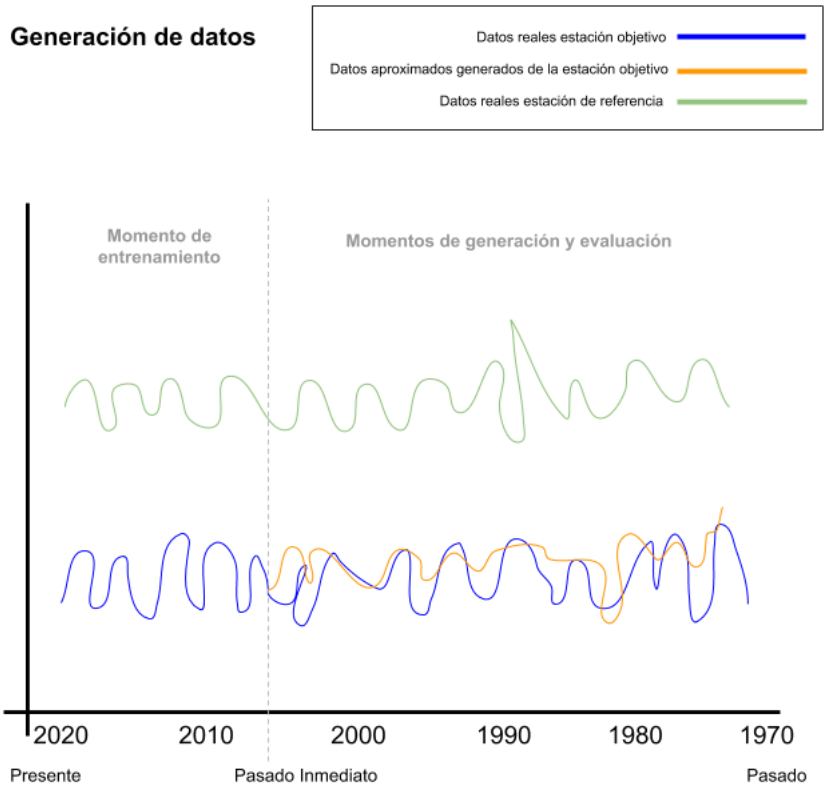


Figura 9. Datos utilizados en cada paso del proceso de generación de datos aproximados

Acopio de información

Se decidió recopilar datos de estaciones meteorológicas de distintas regiones del país con el fin de analizar la correlación entre ellas y contrastar con la distancia existente entre ellas. El Instituto Nacional de Meteorología e Hidrología (INAMHI) facilitó, con fines investigativos, los datos que se encuentran descritos en la

Tabla 1.

La distribución geográfica aproximada de las estaciones se puede apreciar en la Figura 10.

Tabla 1.
Detalle de las estaciones analizadas

Identificador	Nombre	Longitud	Latitud	Elevación
M0003	Izobamba	78°33'18.46"W	0°01'29.20"S	3058 m.s.n.m
M0007	Nuevo Rocafuerte	75°24'10.90"W	0°55'12.10"S	185 m.s.n.m
M0008	Puyo	77°57'29.80"W	1°30'20.40"S	956 m.s.n.m
M0024	Inamhi Iñaquito	78°29'15.83"W	0°10'41.89"S	2789 m.s.n.m
M0025	La Concordia	79°22'49.00"W	0°21'57.33"S	554 m.s.n.m
M0026	Puerto Ila	79°20'56.10"W	0°29'34.80"S	319 m.s.n.m
M0027	Santo Domingo Aeropuerto	79°12'22.34"W	0°14'54.95"S	379 m.s.n.m

Fuente: INAMI
Elaboración: Propia

Figura 10. Distribución geográfica de las estaciones estudiadas

Los datos obtenidos están en el periodo de 1970 al 2018, con totales diarios y mensuales, fueron provistos en dos archivos planos de texto, uno para totales diarios y otro para totales mensuales. Los registros presentan datos faltantes en algunos periodos.

Procesamiento de la información

Las librerías de software utilizadas para el desarrollo y evaluación de las redes neuronales requieren que los datos estén dispuestos en formato *CSV*, con un archivo

existente por cada estación a procesar. Se realizaron *scripts* en lenguaje *Python* que permitieron transformar los archivos de texto en archivos con el formato esperado para cada estación.

De igual manera, las librerías utilizadas requieren que no exista información faltante en periodos intermedios de estudio. Algunas de las estaciones presentaron vacíos de datos para una serie de periodos. Se rellenaron los datos copiando el valor de la observación anterior. A continuación, en la Tabla 2, se muestra un detalle de la información faltante y reconstruida para cada estación.

Tabla 2
Detalle de información faltante

Estación	Información faltante
M0003	Ninguna
M0007	1970 - 1975 (todo el periodo) 1976 (agosto, septiembre, octubre y noviembre) 1977 (enero) 1981 (octubre y noviembre) 1983 (enero y febrero) 1986 (octubre) 2018 (noviembre y diciembre)
M0008	1983 (febrero) 2002 (mayo) 2018 (todo el año)
M0024	1971 - 1974 (todo el periodo) 1985 (marzo, junio, julio, agosto, septiembre, noviembre y diciembre) 2003 (agosto) 2007 (diciembre) 2010 (mayo)
M0025	1970 (junio) 1971 (julio, agosto, septiembre, octubre, noviembre y diciembre) 1972 (enero, abril, junio, julio, agosto, septiembre, octubre y noviembre) 1985 (diciembre) 1988 (mayo)

Estación	Información faltante
M0026	1979 (agosto y septiembre) 1980 (febrero) 1981 (enero, febrero, marzo, mayo, junio, julio, agosto) 1990 (febrero)
M0027	1975 (febrero) 1985 (diciembre) 1986 (septiembre y octubre) 1987 (julio, agosto, septiembre, octubre, noviembre y diciembre) 1988 (enero, abril, mayo, agosto y septiembre) 1989 (octubre) 1990 (julio, agosto y septiembre) 1991 (junio) 1992 (marzo) 1999 - 2018 (todo el periodo)

Fuente: INAMI
Elaboración: Propia

Al inspeccionar los totales de lluvia diarios se pudo observar que el nivel ruido-señal era elevado por lo tanto se decidió calcular los totales semanales. Luego de una inspección a la señal de totales semanales se obtuvo una relación señal-ruido igualmente alta, por lo tanto, se decidió usar los totales mensuales de lluvia.

Análisis de correlación y causalidad

El análisis de correlación permite determinar qué estaciones están relacionadas y tienen mediciones similares y por lo tanto en qué estaciones están gobernadas por los mismos procesos causales. En la Tabla 3, se muestra la correlación entre las estaciones estudiadas.

Se puede observar que las estaciones más cercanas presentan una relación fuerte mientras que las estaciones más lejanas tienen una correlación muy baja, por ejemplo, se observa que la correlación entre la estación M0007 y la estación M0026 es muy baja y entre la estación M0024 y M0023 es alta.

Tabla 3
Correlación entre estaciones

Estaciones	Correlación	Periodo
M0003 - M0007	0.016	1976 - 2018
M0003 - M0008	0.27	1970 - 2017
M0003 - M0024	0.83	1975 - 2018
M0003 - M0025	0.50	1970 - 2018
M0003 - M0026	0.54	1970 - 2018
M0003 - M0027	0.55	1970 - 1998
M0007 - M0008	0.40	1976 - 2017
M0007 - M0024	-0.017	1976 - 2018
M0007 - M0025	0.056	1976 - 2018
M0007 - M0026	-0.012	1976 - 2018
M0007 - M0027	-0.038	1976 - 1998
M0008 - M0024	0.21	1975 - 2017
M0008 - M0025	0.11	1970 - 2017
M0008 - M0026	0.11	1970 - 2017
M0008 - M0027	0.05	1970 - 1998
M0024 - M0025	0.43	1975 - 2018
M0024 - M0026	0.47	1975 - 2018
M0024 - M0027	0.46	1975 - 1998
M0025 - M0026	0.84	1970 - 2018
M0025 - M0027	0.86	1970 - 1998
M0026 - M0027	0.87	1970 - 1998

Fuente y elaboración: propias

Tomando como referencia los datos de correlación, el presente estudio se enfocará en realizar la reconstrucción de datos para la estación M0003 en tomando como referencia la estación M0024 y para la estación M0025 tomando como referencia la estación M0026. Se descarta la alta correlación existente entre las estaciones M0025, M0026 y M0027 porque el rango de datos existentes es muy bajo como para entrenar la red neuronal.

2. Generación de datos aproximados del pasado

Como se detalló anteriormente, en este trabajo de investigación se invertirá el eje del tiempo para generar datos aproximados del pasado a través del uso las mismas herramientas de pronóstico. Estas herramientas hacen uso de la autocorrelación de una variable para pronosticar la misma.

Al invertir el eje del tiempo se produce un fenómeno no contemplado en el pronóstico estadístico, la existencia de datos del futuro. En la Figura 11 se puede observar este fenómeno. En resumen, al invertir el eje del tiempo existen, de manera figurativa, datos del futuro de la estación de referencia. Por esta razón, la generación de datos del pasado se podría efectuar como un proceso de pronóstico basado en la autocorrelación o como un proceso de pronóstico basado en la correlación entre la estación de referencia y la estación objetivo.

Se utilizará un modelo estadístico SARIMA y un modelo de redes neuronales LSTM para generar datos aproximados del pasado utilizando la autocorrelación, por otro lado, para el pronóstico basado en la correlación se utilizará una regresión lineal estadística y una regresión lineal basada en redes neuronales.

Generación de datos del pasado utilizando el modelo SARIMA

Las redes neuronales son una herramienta poderosa y por ello en ocasiones se las utiliza en procesos que podrían haber sido gestionados con herramientas más sencillas como regresiones lineales u otros modelos estadísticos. El presente trabajo busca comparar el comportamiento de diferentes modelos, incluidos entre ellos las redes neuronales, con el objetivo de determinar la pertinencia del uso de estas en procesos de reconstrucción de datos.

Un modelo SARIMA es aplicable a un proceso estacionario auto correlacionado (Aladag y Eğrioğlu 2012, 58). La estacionaridad se representa en un conjunto de datos cuando no existen tendencia al crecimiento o decrecimiento. La autocorrelación de se

determina a través del diagrama de la función de autocorrelación (*ACF*, por sus siglas en inglés).

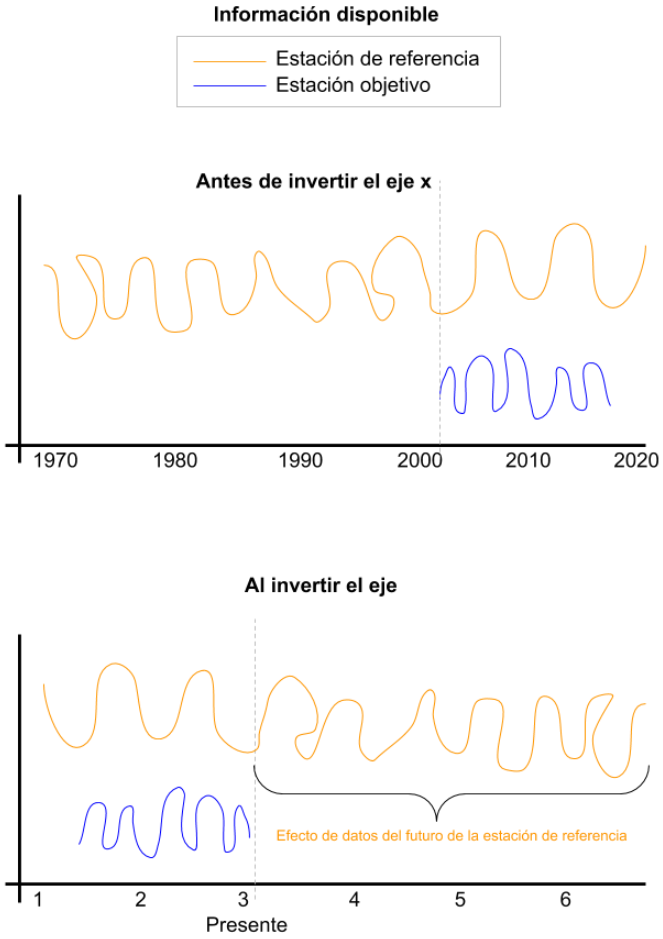


Figura 11. Efecto de datos futuros disponibles

Antes de determinar la estacionaridad y la autocorrelación para el proceso de reconstrucción es preciso tratar los datos de las estaciones ya que el modelo SARIMA trabaja con una sola variable y en el caso de estudio se tienen dos: los datos de la estación de referencia y los datos de la estación objetivo (estación de la que se desea generar datos aproximados del pasado). Para conservar la información de las dos estaciones se realizó el promedio de estos y sobre tal promedio se aplicó el modelo SARIMA y se generaron datos aproximados del pasado. En la Figura 12 se puede observar un detalle del proceso realizado.

En la Figura 13 se puede observar la tendencia de la señal promedio entre M0003 y M0024. Cabe recalcar que la tendencia a pesar de no ser constante no muestra un

crecimiento o decrecimiento de su promedio a lo largo del rango de mediciones por lo que se puede concluir que el proceso es estacionario.

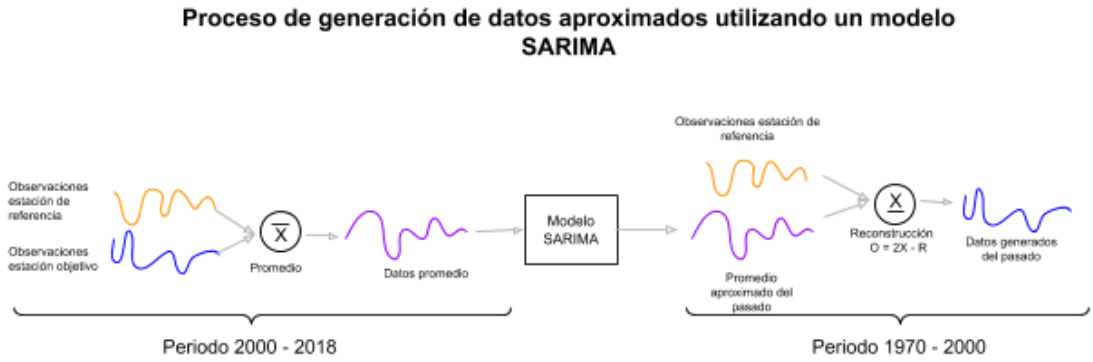


Figura 12. Proceso de generación de datos con un modelo SARIMA

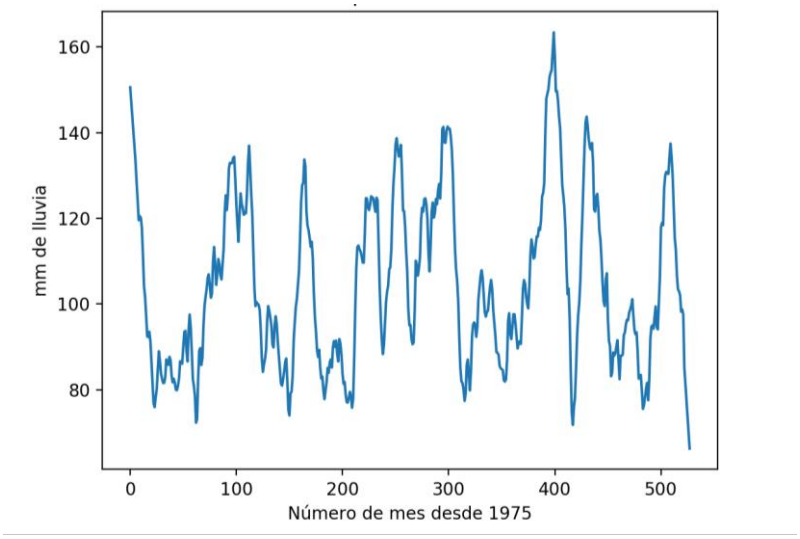


Figura 13. Tendencia del promedio entre las estaciones M0003 y M0024

De igual manera, para el promedio de las estaciones M0025 y M0026 se puede observar, en la Figura 14, una tendencia variable pero que no crece en el tiempo, se puede concluir entonces que el promedio es un proceso estacionario.

En la Figura 15 se muestra la función de autocorrelación para el promedio de las estaciones M0003 y M0024. Se puede observar que existe una autocorrelación media, entre el 40% y 50% para la mayoría de los retardos.

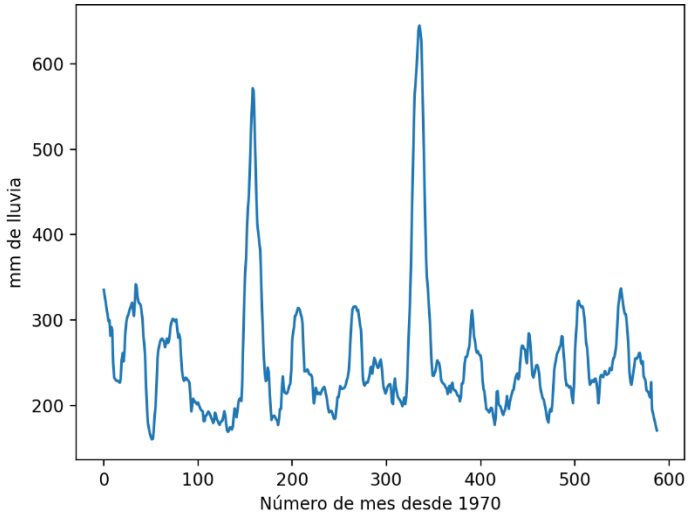


Figura 14. Tendencia del promedio entre las estaciones M0025 y M0026

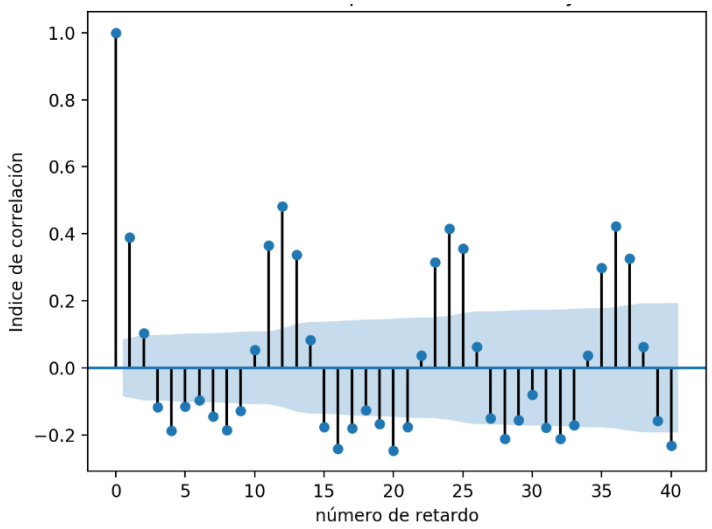


Figura 15. Autocorrelación del promedio de M0003 y M0024

Para las estaciones M0025 y M0026 se puede apreciar una autocorrelación alta, entre el 70% y el 80%. Dados los valores de autocorrelación para las cuatro estaciones se puede estimar que un modelo SARIMA tendrá resultados con mejor aproximación para las estaciones M0025 y M0026 que para M0003 y M0024.

Para crear el modelo SARIMA se debe encontrar la mejor combinación de parámetros estacionales y no estacionales. En el presente trabajo se utilizaron librerías que permiten la auto calibración del modelo a partir de los datos existentes. Luego de haber corrido el proceso de auto calibración se encontró que el modelo SARIMA óptimo para representar el promedio de las estaciones M0003 y M0024 tiene una componente no estacional ARIMA con parámetros: ARIMA (1, 0, 1) y componente estacional nula, es

decir, el mejor es un modelo ARMA (1, 1). En la Figura 16 se puede observar los datos aproximados del pasado versus los datos reales.

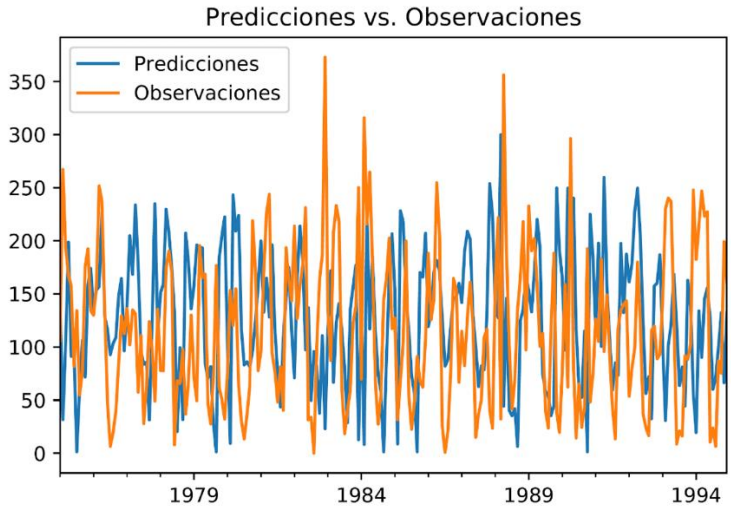


Figura 16. Reconstrucción de datos de la estación M0003 tomando como referencia a M0024 con un modelo SARIMA

Los parámetros de evaluación del modelo SARIMA son: correlación, 0.0071, error absoluto promedio, 76.18 mm y error cuadrático medio, 96.81 mm.

Luego de haber realizado el proceso de auto calibración del modelo SARIMA para las estaciones M0025 y M0026 se estimó que los parámetros óptimos son: ARIMA (1, 0, 1) y componentes de estacionalidad nulos, es decir, el mejor es un modelo ARMA (1, 1). En la Figura 17 se puede observar los datos aproximados del pasado versus los datos reales. Los parámetros de evaluación del modelo SARIMA son: correlación, 0.49, error absoluto promedio, 172.78 mm y error cuadrático medio, 245.62 mm. El error absoluto medio corresponde al 20% del valor máximo encontrado, un valor alto para tener en cuenta al momento de utilizar estos datos.

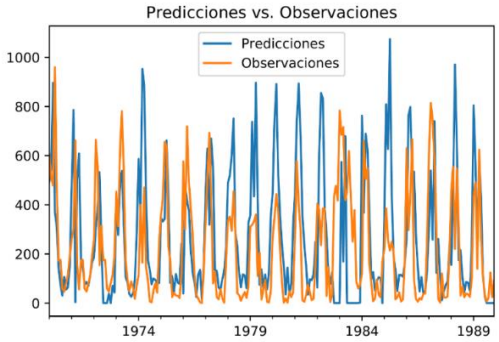


Figura 17. Reconstrucción de datos de la estación M0025 tomando como referencia a M0026 con un modelo SARIMA

Generación de datos del pasado utilizando una regresión lineal

Como se mencionó en párrafos anteriores el proceso de generación de datos aproximados también puede ser visto como una regresión lineal donde la estación de referencia constituye la entrada del modelo y los valores reconstruidos son la salida de este. En esta tesis se usó la librería *scikit learn*, que trabaja sobre el lenguaje de programación *Python*, para auto calibrar un modelo de regresión lineal.

En un primer intento la regresión lineal tuvo una correlación menor que la existente con los datos de referencia, se pudo observar, al descomponer la señal en sus partes, tendencia, estacionalidad y residuo, que la mayor parte del comportamiento lineal se podía observar en la tendencia y estacionalidad. Se decidió correr modelos individuales de regresión lineal para cada componente y al final sumar las reconstrucciones hechas.

Tras entrenar con 24 años de datos el modelo para las estaciones M0003 y M0024 se obtuvieron los siguientes resultados: correlación, 0.83, error absoluto medio, 30.08 mm y error medio cuadrático, 40.16 mm. Los resultados se pueden visualizar en la Figura 18.

De igual manera, tras entrenar con 28 años de datos el modelo para las estaciones M0025 y M0026 se obtuvieron los siguientes resultados: correlación, 0.87, error absoluto medio, 92.57 mm y error medio cuadrático, 133.53 mm. Los resultados se pueden visualizar en la Figura 19.

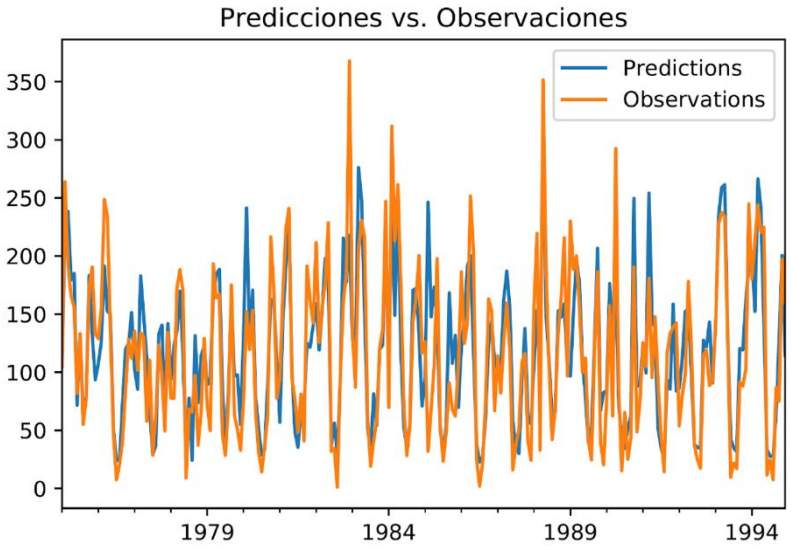


Figura 18. Reconstrucción de datos de la estación M0003 tomando como referencia a M0024 con un modelo de regresión lineal.

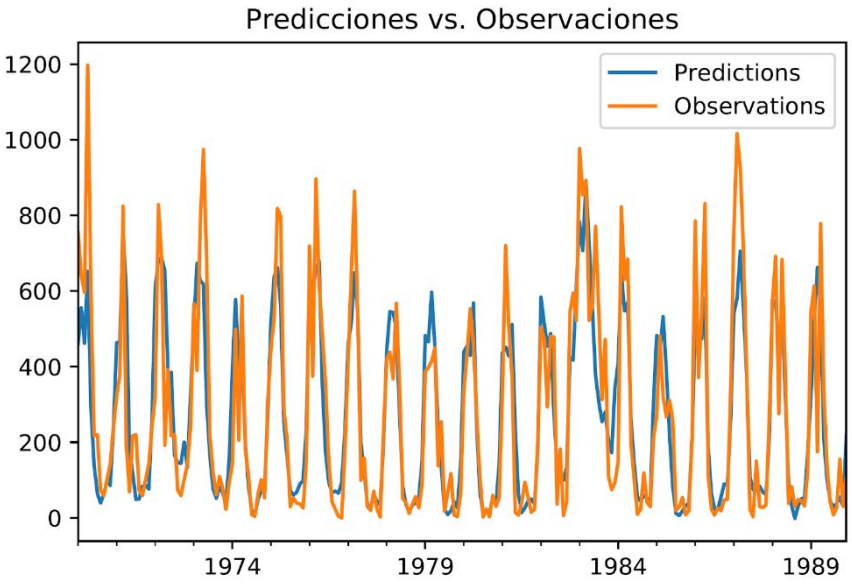


Figura 19. Reconstrucción de datos de la estación M0025 tomando como referencia a M0026 con un modelo de regresión lineal.

Generación de datos utilizando una red neuronal profunda (D.N.N)

La implementación de una red neuronal profunda (DNN) gana sentido bajo la hipótesis de que el problema de generación de datos aproximados del pasado se puede expresar como una regresión con posibles comportamientos no lineales. Los datos de lluvia pueden incorporar comportamientos no lineales, una red neuronal profunda puede ser capaz de modelar los comportamientos lineales y los comportamientos no lineales.

Previa a la creación de la red neuronal se dividieron los datos en sus componentes de tendencia, estacionalidad y componente residual. Esta descomposición se efectuó utilizando la librería *statsmodel* implementada en lenguaje *Python.* El objetivo de dividir la señal es poder crear diferentes redes neuronales para modelar distintos comportamientos de la señal.

Luego de un proceso de calibración de las tres redes neuronales, de tendencia, estacionalidad y residual, para la reconstrucción de datos de la estación M0003 con base a la información de la estación M0024, se obtuvo que las configuraciones óptimas fueron las mostradas en la

Tabla 4.

Tabla 4
Parámetros de las redes neuronales DNN para la reconstrucción de datos de las estaciones M0003 y M0024.

Componente	Parámetro	Descripción
Tendencial	Número de capas internas	3
	Distribución de neuronas	9 - 6 - 3
	Funciones de activación	relu
	Epochs	200
Estacional	Número de capas internas	1
	Distribución de neuronas	10
	Funciones de activación	relu
	Epochs	100
Residual	Número de capas internas	1
	Distribución de neuronas	10
	Funciones de activación	relu
	Epochs	100

Fuente y elaboración: propias

La red que caracteriza a la componente de tendencia mostró mejor desempeño con un diseño de tres capas intermedias, la primera capa con 9 neuronas, la segunda con 6 neuronas y la tercera con 3 tres neuronas. La función de activación, que mejor desempeño tuvo, en cada capa intermedia, fue una función lineal o *relu*. Finalmente, en la capa de salida se utilizó una función de activación *sigmoide,* de tipo no lineal, para modelar cualquier posible componente no lineal. La red neuronal descrita luce similar a la de la Figura 20.

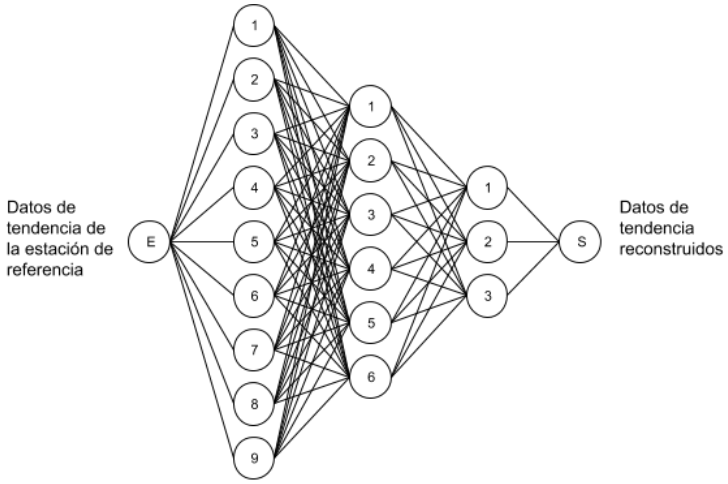


Figura 20. Red neuronal para la reconstrucción de la componente tangencial de M0003

La red que caracteriza a la componente estacional mostró mejor desempeño con un diseño de una sola capa intermedia, es una red neuronal sencilla, no una red neuronal profunda. La capa intermedia tiene diez neuronas. La función de activación, que mejor desempeño tuvo para la capa intermedia fue una función lineal *relu.* La capa de salida tiene una función de activación *sigmoide* para poder representar algún comportamiento no lineal existente. La red neuronal para representar la componente estacional luce como la de la Figura 21.

La red que caracteriza a la componente residual mostró un resultado óptimo con una configuración igual a la red utilizada para reconstruir la componente estacional, es decir, una neurona en la entrada, una sola interna con diez neuronas y una capa de salida con una neurona.

Al final, luego que cada componente fue reconstruida utilizando las redes neuronales descritas anteriormente, se sumó cada reconstrucción para obtener los datos aproximados de la estación M0003. En la Figura 22 se puede apreciar los datos generados. Los resultados de la evaluación de la generación de datos aproximados del pasado son: correlación 0.83, error absoluto medio, 40.15 mm y error medio cuadrático, 30.49 mm.

De igual manera, se siguió un proceso similar al descrito anteriormente para generar los datos de la estación M0025 tomando como referencia la estación M0026. La configuración de las redes neuronales es similar, únicamente, para la componente de tendencia se mostró mejor desempeño al eliminar la primera capa de nueve neuronas. Los resultados de la generación se pueden apreciar en la Figura 23. La evaluación de la red neuronal arrojó los siguientes resultados: correlación, 0.85, error absoluto medio, 100.24 mm, y error cuadrático medio, 141.15 mm.

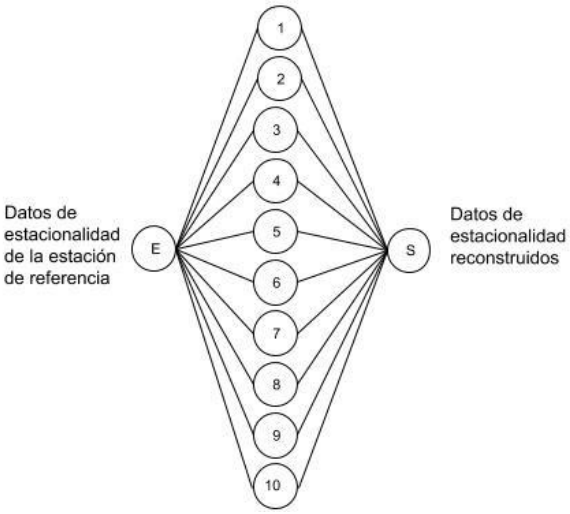


Figura 21. Red neuronal para la reconstrucción de la componente estacional de M0003

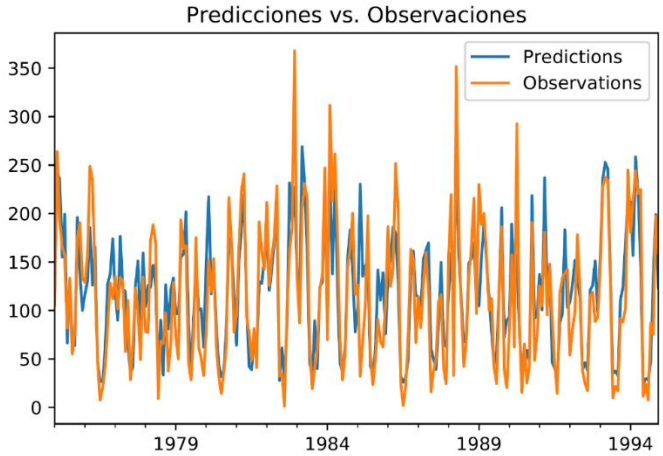


Figura 22. Reconstrucción de datos de la estación M0003 tomando como referencia a M0024 con una DNN

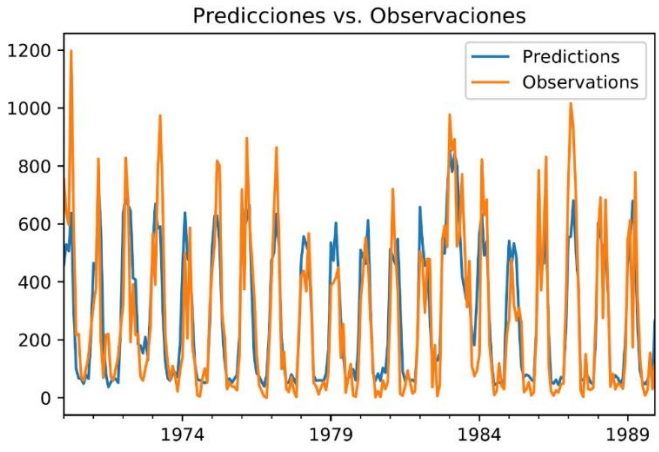


Figura 23. Reconstrucción de datos de la estación M0025 tomando como referencia a M0026 con una DNN

Reconstrucción de datos utilizando una red neuronal recurrente (L.S.T.M)

Los datos de lluvia, como se mencionó al inicio del capítulo, se pueden considerar como un proceso auto correlacionado en el que cada medición tiene relación y dependencia con las mediciones anteriores. Las redes neuronales que caracterizan de mejor manera este tipo de procesos son las redes neuronales recurrentes. Actualmente, la implementación de una red neuronal recurrente con mayores prestaciones es el tipo *Long Short-Term Memory* (*LSTM* por sus siglas en inglés).

Se implementó una red neuronal LSTM para generar datos del pasado de la estación M0003 tomando como base la información de la estación M0024 y otra red para generar los datos del pasado de la estación M0025 tomando como referencia la información de la estación M0026. El objetivo de la implementación de estas dos redes es determinar si una red neuronal recurrente es capaz de caracterizar de mejor manera que las otras redes implementadas el proceso de reconstrucción de datos.

Por la característica de las redes neuronales en las que hay una entrada y una salida no es necesario realizar un promedio de las señales como se efectuó en el modelo SARIMA. Por otro lado, fue necesario transformar los datos para que los conjuntos de entrada a la red neuronal contengan los *retardos* o información del pasado. Este procesamiento se lo ejecutó utilizando la librería especializada en redes neuronales *keras* implementada sobre el lenguaje de programación *Python.*

La configuración óptima de la red neuronal incluye una única capa de elementos de memoria LSTM con 200 elementos. La función de activación para cada elemento es lineal del tipo *relu*. La capa de salida incluye una neurona con una función de activación no lineal de tipo *sigmoide*.

Luego de haber entrenado a la red neuronal con un *epoch* de 200 se obtuvieron los siguientes resultados: correlación, 0.53, error absoluto medio, 48.94 mm y error absoluto medio cuadrático, 63.89 mm. Los resultados de la reconstrucción utilizando este tipo de red neuronal se pueden apreciar en la Figura 24.

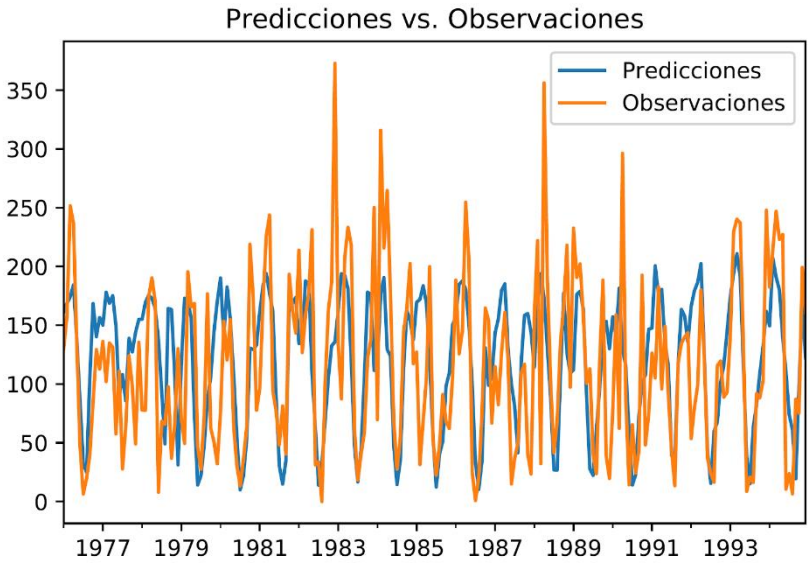


Figura 24. Reconstrucción de datos de la estación M0003 tomando como referencia a M0024 con una red LSTM

De igual manera, se utilizó la misma configuración de red neuronal para generar los datos de la estación M0025 tomando como referencia los datos de la estación M0026. Luego de haber entrenado a la red neuronal con un *epoch* de 200 se obtuvieron los siguientes resultados: correlación, 0.70, error absoluto medio, 119.54 mm, error cuadrático medio, 175.79 mm. Los valores de la reconstrucción se pueden apreciar en la Figura 25.

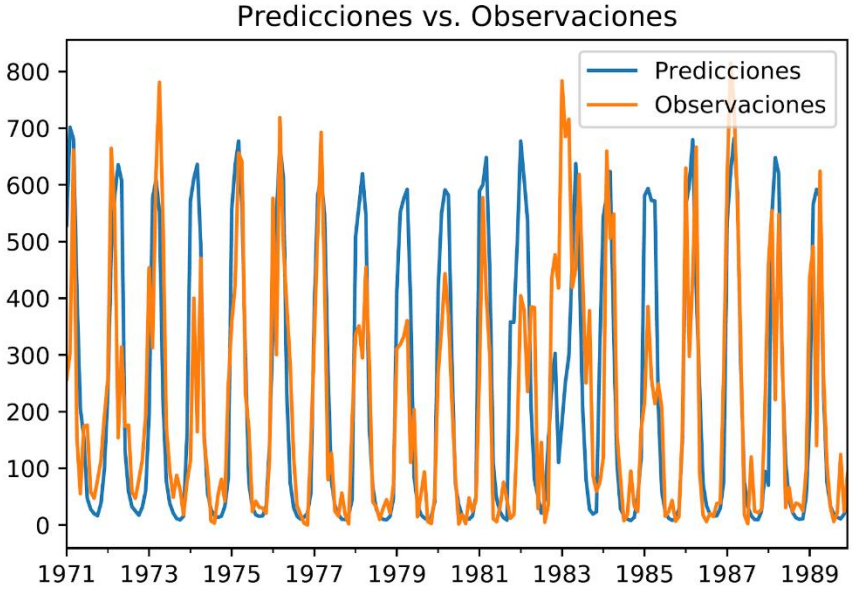


Figura 25. Reconstrucción de datos de la estación M0025 tomando como referencia a M006 con una red LSTM

Capítulo tercero
Análisis de resultados

1. Análisis del desempeño de los modelos utilizado para la generación de datos aproximados del pasado

Se implementaron dos redes neuronales y dos modelos estadísticos para realizar la reconstrucción de datos de las estaciones M0003 y M0025. De acuerdo con los estipulado en el artículo de Tran Anh Duong et. al. (2018), al menos dos medidas se deben analizar para cada reconstrucción realizada, se decidió utilizar tres, correlación, error absoluto medio y error cuadrático medio. También se realizó un análisis de bias y varianza, de los modelos con mejor desempeño, para determinar la capacidad de estos de reproducir los datos de lluvia.

El error absoluto medio y el error cuadrático medio son medidas de dispersión, es decir, miden que tan alejado se encuentran los errores del promedio. El error cuadrático medio *castiga* a los errores más grandes, y el error medio absoluto no lo hace. Por otro lado, la correlación es una medida de sesgo, indica que tan certera es la reconstrucción respecto del valor real.

En la Tabla 5 se muestran el resumen de los resultados de la aplicación de los modelos estadísticos y de dos diferentes tipos de redes neuronales para las estaciones en estudio.

Tabla 5
Evaluación de la generación de datos del pasado

Modelo / Estación	M0003	M0025
Regresión lineal	Correlación: 0.83 MAE: 30.08 mm RMSE: 40.16 mm RMSE Modelo: 0.04 Error Medio: 30.08 mm	Correlación: 0.87 MAE: 92.57 mm RMSE: 133.53 mm RMSE Modelo: 0.13 Error Medio: 92.57 mm
SARIMA	Correlación: 0.0071 MAE: 76.18 mm RMSE: 96.81 mm Error Medio: 76.18 mm	Correlación: 0.49 MAE: 172.78 mm RMSE: 245.62 mm Error Medio: 172.78 mm
Red neuronal profunda	Correlación: 0.83 MAE: 40.15 mm RMSE: 30.49 mm	Correlación: 0.85 MAE: 100.24 mm RMSE: 141.15 mm

Modelo / Estación	M0003	M0025
	RMSE Modelo: 0.03 Error Medio: 30.49 mm	RMSE Modelo: 0.14 Error Medio: 100.24 mm
LSTM	Correlación: 0.53 MAE: 48.94 mm RMSE: 63.89 mm Error Medio: 47.75 mm	Correlación: 0.70 MAE: 119.54 mm RMSE: 175.79 mm Error Medio: 114.81mm

Fuente y elaboración: propias

Como se puede observar, los modelos que presentan una mayor correlación son la regresión lineal simple y la red neuronal profunda. De acuerdo con (Armenta 2020, entrevista personal) los datos con una correlación mayor a 0.8 se podrían utilizar para alimentar a los modelos de impacto del cambio climático, es decir, se podría utilizar una regresión lineal simple o una red neuronal profunda. Para los dos modelos los niveles de MAE y RMSE tienen un promedio de 50 mm. Este valor se demuestra cuán dispersos están los errores del error promedio.

Por otro lado, el análisis de sesgo y desviación estándar demostró que, para la generación de datos aproximados del pasado de la estación M0025, el modelo de regresión lineal sobrestima en 151.5 mm. las mediciones de lluvia y subestima en 76.45 mm. la misma variable. La desviación estándar es de 8.17 mm. Para el modelo de regresión basado en redes neuronales artificiales se obtuvo una sobrestimación de 45.32 mm. y una subestimación de 39.04 mm. y una desviación estándar de 3.72 mm.

La generación de datos del pasado a través de los dos modelos, regresión lineal y red neuronal profunda, presenta un desempeño similar, sin embargo, la regresión lineal, desde el punto de vista logístico, es más fácil de llevar a cabo. Si un proyecto tiene la capacidad de contar con personal para el desarrollo de redes neuronales se recomienda utilizar la red neuronal profunda ya que esta metodología es capaz de representar comportamientos no lineales de la serie temporal.

2. Análisis de datos reconstruidos respecto a la base de datos WorldClim

Los datos provistos por WorldClim se encuentran contenidos en un archivo *raster* de formato *tif*, fue necesario transformar los datos para que puedan ser comparados con las mediciones reales y con los datos aproximados obtenidos de la red neuronal profunda. Se exportaron los datos con la ayuda del utilitario *QGis* a un formato *CSV*. En la Tabla 6

se puede observar un comparativo de los datos reales, los datos de WorldClim y los datos reconstruidos.

De la tabla se puede calcular el error promedio de los datos de WorldClim, 33% para M0003 y 28% para M0025, y de los datos aproximados del pasado, 11% para M0003 y 17% para M0025.

Tabla 6
Comparación de datos de WorldClim, Reales y Datos generados del pasado

	M0003					M0025				
Mes	**Real**	**WorldC.**	**% Error**	**DNN**	**% Error**	**Real**	**WorldC.**	**% Error**	**DNN**	**% Error**
Enero	131.0	97	26%	123.7	6%	463.6	429	7%	552.9	19%
Febrero	158.7	116	27%	157.1	1%	578.1	485	16%	560.4	3%
Marzo	182.3	141	23%	176.0	3%	591.5	522	12%	578.3	2%
Abril	195.8	157	20%	174.6	11%	639.1	466	27%	561.7	12%
Mayo	156.4	110	30%	140.9	10%	345.5	232	33%	311.1	10%
Junio	68.6	45	34%	50.5	26%	185.8	139	25%	131.7	29%
Julio	32.9	17	48%	46.7	42%	103.3	41	60%	84.6	18%
Agosto	41.9	26	38%	50.2	20%	64.0	44	31%	87.1	36%
Septiembre	93.0	58	38%	95.3	2%	90.2	55	39%	85.7	5%
Octubre	135.7	89	34%	147.7	9%	92.2	71	23%	104.0	13%
Noviembre	136.3	77	43%	137.1	1%	74.3	46	38%	95.3	28%
Diciembre	126.1	79	37%	125.1	1%	190.3	155	19%	254.1	33%

Fuente y elaboración: propias

Una de las críticas más fuertes a los datos de WorldClim es que no toma en cuenta los fenómenos locales para la generación de datos, este enunciado justifica que el error en los datos de WorldClim sea mayor para la estación ubicada en la sierra que para la estación ubicada en la costa. Por el contrario, los datos del pasado, generados en este estudio, presentan mayor exactitud para la estación de la sierra.

El error de los datos provistos por WorldClim excede en casi el doble al de los datos reconstruidos. Esta variación se justifica porque la metodología de WorldClim interpola datos de estaciones cercanas mientras que los datos aproximados generados por los modelos trabajan con estaciones con una mayor correlación.

3. Análisis de la generación de datos del pasado respecto a la distancia entre estaciones

Para generar datos aproximados del pasado es necesario contar con una estación de referencia que tenga datos reales en el periodo que se desea generar dichos datos, además, las dos estaciones deben tener una correlación aceptable entre ellas. En este apartado se realiza un análisis de los datos generados de la estación M0025 tomando como estaciones de referencia tres estaciones ubicadas a diferentes distancias con el fin de tener una referencia del desempeño del proceso de generación de datos aproximados respecto a la distancia entre estaciones. Las estaciones de referencia se encuentran a: 446 Km. (M0007), 101 Km (M0024) y 14.5 Km (M0026).

El análisis de desempeño del proceso de generación de datos respecto a la distancia entre estaciones se simplificó a un solo modelo y estación, el trabajo para las otras estaciones y otros modelos es similar. Se tomó como estación objetivo a M0025 y como modelo a una regresión basada en redes neuronales artificiales. Los resultados del ejercicio se pueden observar en la Figura 26, Figura 27 y Figura 28.

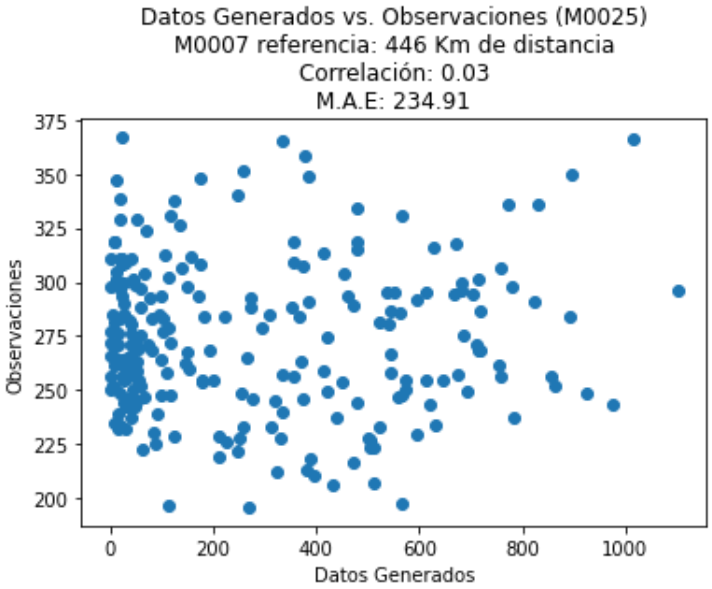


Figura 26. Generación de datos de M0025 con M0007 como estación de referencia

Se pudo determinar que la generación de datos aproximados tiene un mayor desempeño para estaciones de referencia y objetivo más cercanas. Entre los datos generados por el modelo entrenado tomando como referencia la estación M0007 y los

datos reales de M0025 existe una correlación baja y un error absoluto promedio muy alto comparado con las mediciones mayores, por el contrario, la correlación entre los datos generados por el modelo entrenado con la estación de referencia M0026 es alta y error absoluto promedio relativamente bajo.

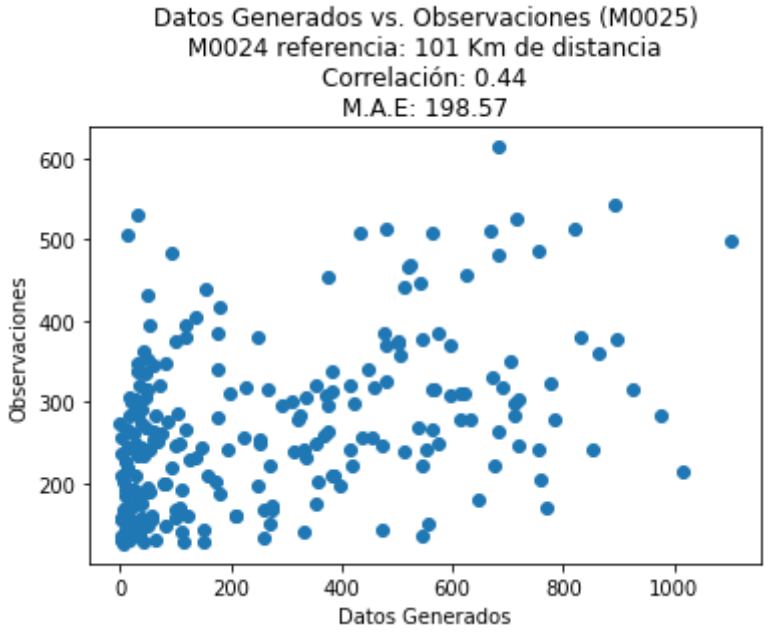


Figura 27. Generación de datos de M0025 con M0024 como estación de referencia

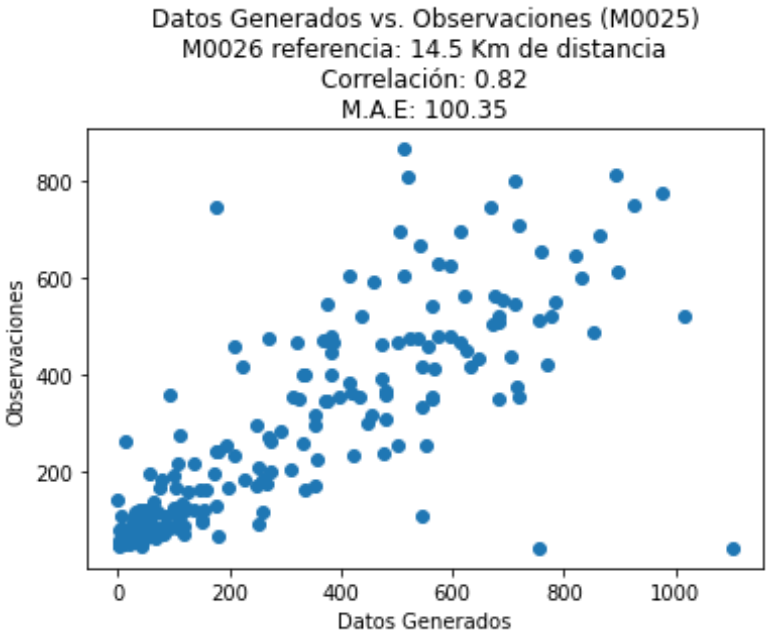


Figura 28. Generación de datos de M0025 con M0026 como estación de referencia

4. Análisis de otras metodologías para evaluación del impacto del cambio climático

Una de las alternativas, para la información requerida por los modelos de impacto al cambio climático, es la obtenida a través de sensores remotos. Según Guillermo Armenta (2020), el esfuerzo de procesamiento de los datos de sensores remotos es similar al de procesamiento de observaciones en sitio, la única diferencia es la exactitud y el proceso repetitivo de buscar fuentes con medidas adecuadas de correlación y sesgo. La exactitud de la información requerida es relativa a la escala del estudio, información de promedios mensuales es suficiente para escenarios de una región amplia, por el contrario, si se desea realizar análisis de eventos extremos o tendencias climáticas es necesario disponer de datos diarios (Armenta 2020). Por lo tanto, el tipo de estudio condiciona la calidad de información y por ende la fuente.

Los impactos del cambio climático se pueden evaluar alternativamente a través de una estimación cualitativa que toma como base las salidas de los modelos de circulación general o los datos de las bases que han pasado por un proceso de reanálisis. Por ejemplo, si los datos de WorldClim muestran una tendencia a la reducción de lluvia en un área en específico, se generan medidas de adaptación a esta tendencia sin tomar en cuenta el valor exacto esperado de esta variación. A pesar de tener la ventaja de llevar a la acción de una manera rápida, con este tipo de estimación, se corre el peligro de fomentar iniciativas que no estén alineadas con exactitud a las condiciones futuras (Núñez 2020, entrevista personal).

Conclusiones y recomendaciones

En este apartado se detallan las observaciones y aportes que se desprenden del trabajo de generación de datos aproximados de lluvia a través del uso de redes neuronales. Al final de esta sección se propondrá posibles trabajos, que, a partir de este documento, puedan incrementar el conocimiento sobre el uso de redes neuronales para la reconstrucción de datos y determinación de impactos del cambio climático.

Conclusiones sobre el desempeño de las redes neuronales en la generación de datos aproximados del pasado

El objetivo principal del trabajo de investigación es determinar el desempeño de las redes neuronales en la generación de datos aproximados del pasado de la lluvia. Dos redes neuronales se implementaron, una red neuronal profunda y una red LSTM. La red neuronal profunda tuvo un desempeño superior a la red LSTM.

La red neuronal profunda obtuvo un nivel de correlación (0.85) que permitiría su uso en modelos de impacto al cambio climático. La red LSTM tuvo un desempeño medio y alcanzó una correlación relativamente baja (0.70) por lo que no se podría utilizar para alimentar dichos modelos.

El desempeño alto de la red profunda se puede explicar desde la estadística y la inteligencia artificial. El proceso de generación de datos se asemeja más a un proceso de regresión que a un proceso auto correlacionado, la correlación entre la estación a reconstruir y la información de la estación de referencia es más fuerte que la correlación entre las mediciones actuales y las mediciones pasadas, por esta razón el modelo SARIMA y LSTM tienen un menor desempeño que la regresión lineal y la red neuronal profunda.

Conclusiones sobre la creación y entrenamiento de una red neuronal

Se elaboraron dos redes neuronales porque el proceso de generación de datos se comporta de dos maneras diferentes y era necesario evaluar cuál de ellas tenía más peso al momento de la generar los datos. De los resultados obtenidos por las redes neuronales y del análisis de correlación y autocorrelación se concluye que el proceso de reconstrucción de datos se asimila más a una regresión que a una serie temporal.

La calibración de los parámetros de una red neuronal no sigue un proceso definido. Existe una serie de consideraciones y aspectos particulares que se deben tomar en cuenta

al momento de calibrar la red, es un proceso que requiere de tiempo y conocimiento por parte del operador. Si se deseara utilizar una red neuronal para reconstruir los datos, es requerimiento que un proyecto cuente con los servicios de una persona con conocimiento de redes neuronales y aprendizaje supervisado.

El entrenamiento de la red, por otro lado, es un proceso sencillo que no requiere de un extenso poder computacional o tiempo. La red neuronal profunda del presente trabajo se entrenó presentando toda la serie de datos un promedio de 200 veces. El tiempo requerido para el entrenamiento no fue mayor a 5 minutos. Sin embargo, se debe tener cuidado en el proceso de entrenamiento para que la red neuronal no caiga en *overfitting* o baja capacidad de generalización, circunstancia en donde la red puede reproducir únicamente los datos de entrenamiento, pero no el comportamiento de un sistema.

Conclusiones sobre la generación de datos aproximados del pasado

El proceso de generación de datos se llevó a cabo como un pronóstico al invertir la línea de tiempo. Se obtuvieron correlaciones altas, 0.83 para la estación M0003 y 0.85 para la estación M0025 demostrando que esta propuesta metodológica constituye una alternativa más al momento de generar información del pasado con el objetivo de ser utilizada como entrada para la reducción de escala de los modelos de circulación general. Es posible que se alcancen niveles más altos de correlación y medidas de dispersión si se utilizan más datos de entrada como temperatura, viento, radiación solar, etc.

Conclusiones sobre la exactitud de la generación de datos aproximados y comparación con WorldClim

La medición de desempeño de un proceso de pronóstico presenta una serie de retos que no permiten expresar el resultado como un porcentaje fijo, por ejemplo, se pueden tener errores bajos en la mayoría de las muestras pronosticadas pero un error alto en algunas de las medidas desvirtuando el error promedio. Por esta razón a un proceso de pronóstico se califica al menos con dos medidas: una de error relativo o calidad del pronóstico y otra de error absoluto (Tran Anh, Bui, y Rutschmann 2018). En el presente trabajo también se realizó un análisis de sesgo utilizando el bias y la desviación estándar.

La medida de error relativo más común es el coeficiente de correlación. El coeficiente de correlación expresa el grado de relación de dos series de datos, puede variar entre -1 y 1, los extremos implican una relación fuerte y los valores cercanos a cero una

relación débil o nula. Los coeficientes obtenidos de 0.83 y 0.85 por lo tanto demuestran una relación fuerte entre la reconstrucción de datos y las observaciones.

La medida absoluta más utilizada para procesos de regresión es el error cuadrático medio. Esta medida representa la distancia promedio que existe entre los datos reconstruidos y las observaciones reales. Las mediciones de error cuadrático medio de 30.49 mm y 141.15 mm pueden ser explicadas como el ancho de la línea de regresión. Para el caso de estudio, donde los valores máximos de precipitación oscilan entre los 700 mm y 1000 mm, el ancho de las líneas sería del 3% del valor máximo para el primer caso y del 14% en el segundo.

Es importante comparar los datos aproximados generados con los datos de WorldClim para determinar si es relevante llevar a cabo el esfuerzo de generación de datos o no. Si los datos de WorldClim presentan una exactitud similar entonces la generación por redes neuronales sería un esfuerzo innecesario.

Los datos de WorldClim son promedios mensuales en el rango de años de 1970 al 2000. Esta característica hace que este conjunto de datos no sea óptimo para ser usado en modelos de impacto de cambio climático (Armenta 2020). Por otro lado, la exactitud de los datos reconstruidos es superior a los de WorldClim, por ende, la reconstrucción de datos es una mejor alternativa que la utilización de los productos de WorldClim.

Uso de redes neuronales en la predicción

Las redes neuronales no constituyen un mecanismo de reemplazo de las actuales técnicas de predicción del clima o del tiempo. La capacidad de representar fenómenos no lineales y de encontrar relaciones estadísticas entre diferentes variables hace que las redes neuronales, en lugar de ser un mecanismo de reemplazo, sean un apoyo para un pronóstico más exacto. Varios trabajos se han llevado a cabo con éxito en este sentido. Por ejemplo, se han utilizado redes neuronales para encontrar relaciones entre los datos de estaciones meteorológicas y las salidas de los modelos de circulación general para así reducir la escala de estos y poder representar fenómenos locales. Se han utilizado procesos de inteligencia artificial para incorporar datos de sensores remotos a las mediciones locales con el fin de incluir fenómenos más globales como insumo de la predicción. La aplicación de redes neuronales en al ámbito del pronóstico es un campo activo en el que una continua producción científica demuestra el interés y en la mayoría de los casos la efectividad del uso de estas.

Uso de redes neuronales en modelos de impacto

La capacidad de una red neuronal para generar datos aproximados del pasado de una estación tomando como base la información de otra está estrechamente relacionada a la correlación que exista entre las dos estaciones. Si una estación está muy lejana a la otra la correlación bajará y la reconstrucción no arrojará los resultados esperados.

El uso de los datos reconstruidos como información de entrada de los modelos de impacto del cambio climático está restringido por lo tanto a la correlación de la estación con una segunda estación que tenga datos desde 1970, De la restricción de correlación se desprende que no debe existir una distancia grande o una topografía variable entre ellas.

Por las razones mencionadas anteriormente, se puede observar una correlación mayor en la estación M0025 que en la estación M0003. La primera estación se encuentra en la costa, a una distancia de 33 Km de la estación de la cual se tomaron los datos base y la segunda estación se encuentra a 20Km de la estación de referencia, a pesar de que la segunda estación (M0003) está más cercana a la estación de referencia la correlación es menor.

Trabajo Futuro

La presente investigación constituye un punto de partida para el uso de redes neuronales en la generación de datos aproximados del pasado. Una serie de estudios se pueden generar tomando como referencia lo analizado. En los siguientes párrafos se mencionan algunos de los posibles trabajos que extiendan la investigación en este campo.

Se puede ejecutar un ejercicio similar al realizado en esta tesis agregando una mayor cantidad de parámetros. A pesar de que la correlación de los datos generados es alta, existen variables, como la topografía, la radiación solar y el viento, que podrían ser utilizadas como información de ingreso a las redes neuronales para ayudar a mejorar el proceso.

De igual manera el presente trabajo se podría extender tomando como referencia no sólo una sino varias estaciones relacionadas o una mezcla de estaciones relacionadas e información de sensores remotos o salidas de modelos de circulación general.

Obras citadas

Abhishek, Kumar, M. P. Singh, Saswata Ghosh, y Abhishek Anand. 2012. "Weather Forecasting Model Using Artificial Neural Network". *Procedia Technology*, 2nd International Conference on Computer, Communication, Control and Information Technology(C3IT-2012) on February 25 - 26, 2012, 4 (enero): 311–18. doi:10.1016/j.protcy.2012.05.047.

Tealab, Ahmed, Hesham Hefny y Amr Badr. 2018. "Forecasting of nonlinear time series using ANN". *Future Computing and Informatics Journal (2)* 39-47.

Aladag, Cagdas Hakan, y Erol Eğrioğlu. 2012. *Advances in Time Series Forecasting*. Oak Park, Ill.: Bentham eBooks. http://site.ebrary.com/id/10587894.

Alpaydin, Ethem. 2014. *Introduction to machine learning*. Third edition. Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press.

AlShehri, S.A., y S. Khatun. 2009. "Uwb Imaging For Breast Cancer Detection Using Neural Network". *Progress In Electromagnetics Research C* 7. https://pdfs.semanticscholar.org/1a74/a0a8392260792241ccf1ec0f00c465f6df69.pdf.

Ansari, Uzma, y Tanuja K. Sarode. 2017. "Skin Cancer Detection Using Image Processing". En *International Research Journal of Engineering and Technology (IRJET)*. Vol. 4.

Armenta, Guillermo. 2020. Alternativas a la información generada por estaciones meteorológicas en modelos de impacto al cambio climático.

Barr, Avron. 1982. "The Handbook of Artificial Intelligence".

Brady, Henry E. 2011. "Causation and Explanation in Social Science". *The Oxford Handbook of Political Science*, julio. doi:10.1093/oxfordhb/9780199604456.013.0049.

Campozano, Lenin, Rolando Célleri, Katja Trachte, Joerg Bendix, y Esteban Samaniego. 2016. "Rainfall and Cloud Dynamics in the Andes: A Southern Ecuador Case Study". Editado por Charles Jones. *Advances in Meteorology* 2016 (enero). Hindawi Publishing Corporation: 3192765. doi:10.1155/2016/3192765.

Ceccaroni, Luigi. 2007. "Inteligencia Artificial, Introducción a la inteligencia artificial". Universitat Politècnica de Catalunya. https://www.cs.upc.edu/~luigi/II/IA-2007-fall/1-introduccion-a-la-inteligencia-artificial-%28es%29.pdf.

Ch.Jyosthna, Devi, y B Syam Prasad Reddy. 2012. "ANN Approach for Weather Prediction using Back Propagation". http://www.ijettjournal.org/volume-3/issue-1/IJETT-V3I1P204.pdf.

De la Fuente, Santiago. s. f. "Series Temporales: Modelo ARIMA". http://www.estadistica.net/ECONOMETRIA/SERIES-TEMPORALES/modelo-arima.pdf.

Flato, G, J Marotzke, y B Abiodun. 2013. "Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change". En *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge.

Gomez, Emiliio. 2001. "El rompecabezas del cerebro: La conciencia. Capítulo 2".

"Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies". 2009. En *A Field Guide to Dynamical Recurrent Networks*, de John F. Kolen y Stefan C. Kremer. IEEE. doi:10.1109/9780470544037.ch14.

Hecht-Nielsen, Robert. 2014. “Theory of the Backpropagation Neural Network”. En *Neural Networks for Perception Computation, Learning, and Architectures*, 65–93.

Hijmans, Robert J, Susan E Cameron, Juan L Parra, Peter G Jones, y Andy Jarvis. 2005. “Very High Resolution Interpolated Climate Surfaces for Global Land Areas”. *Int. J. Climatol. International Journal of Climatology* 25 (15): 1965–78.

Hinton, Geoffrey E., y Terrence J. Sejnowski, eds. 1999. *Unsupervised learning: foundations of neural computation*. Computational neuroscience. Cambridge, Mass: MIT Press.

Hirani, Dawal, y Nitin Mishra. 2016. “A Survey On Rainfall Prediction Techniques”. *International Journal of Computer Application* 6 (2): 2250–1797.

Information Resources Management Association, ed. 2019. *Deep learning and neural networks: concepts, methodologies, tools, and applications*. Hershey: Engineering Science Reference.

IPCC. 2017. “General Guidelines On The Use Of Scenario Data For Climate Impact And Adaptation Assessment”.

———. 2020. “What is a GCM?” Accedido enero 23. http://apps.ipcc-data.org/guidelines/pages/gcm_guide.html.

Kenny, David A. 1979. *Correlation and causality*. New York: Wiley.

Luk, Kin C., J. E. Ball, y A. Sharma. 2001. “An Application of Artificial Neural Networks for Rainfall Forecasting”. *Mathematical and Computer Modelling* 33 (6): 683–93. doi:10.1016/S0895-7177(00)00272-7.

MAE. 2020. “Informe de sistematización”. En *Taller de presentación de avances en la investigación sobre Cambio Climático en el Ecuador y articulación interinstitucional*. http://suia.ambiente.gob.ec/documents/907429/972259/20+Relatoria+del+Taller+MAE-RECC.pdf/47d878e4-cd59-47a7-8958-5672932b10c1.

Maisincho, Luis. 2019. Las estaciones meteorológicas y los datos que generanEscrito.

Matich, Damián. 2001. “Redes Neuronales: Conceptos Básicos y Aplicaciones”. Argentina: Universidad Tecnológica Nacional – Facultad Regional Rosario. https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf.

McCullagh, Peter. 2002. “What Is a Statistical Model?” *The Annals of Statistics* 30 (5): 1225–1310.

N.I.S.T. 2020. “Autocorrelation”. Accedido marzo 29. https://www.itl.nist.gov/div898/handbook/eda/section3/eda35c.htm.

NIST. 2020. “Definitions, Applications and Techniques”. Accedido enero 23. https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc41.htm.

NOAA. 2020. “ESRL : PSD : NCEP/NCAR Reanalysis 1”. Accedido enero 23. https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html.

NOAA. 2020. “Precipitation Forecasting Using a Neural Network”. Accedido julio 13. https://www.nssl.noaa.gov/users/brooks/public_html/hall/neural.html.

NOAA. 2020. “Improve CFS Week 3-4 Precipitation and 2 Meter Air Temperature Forecasts with Neural Network Techniques”. Accedido julio 13. https://www.nws.noaa.gov/ost/climate/STIP/43CDPW/43cdpw-YFan.pdf.

Nuñez, Jorge. 2020. Alternativas cualitativas para la evaluación del impacto del cambio climático.

Olah, Christopher. 2015. “Understanding LSTM Networks”. agosto 27. https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Parkhi, Omkar, Andrea Vedaldi, Andrew Zisserman, y C.V. Jawahar. 2020. “Cats and Dogs”. University of Oxford. Accedido enero 23. https://www.robots.ox.ac.uk/~vgg/publications/2012/parkhi12a/parkhi12a.pdf.

S. Lawrence, y C. L. Giles. 2000. “Overfitting and neural networks: conjugate gradient and backpropagation”. En *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, 1:114–19 vol.1. doi:10.1109/IJCNN.2000.857823.

Snell, Seth E., Sucharita Gopal, y Robert K. Kaufmann. 2000. “Spatial Interpolation of Surface Air Temperatures Using Artificial Neural Networks: Evaluating Their Use for Downscaling GCMs”. *Journal of Climate* 13 (5): 886–95. doi:10.1175/1520-0442(2000)013<0886:SIOSAT>2.0.CO;2.

Roux, Stephane G., Venugopal V., Fienberg Kurt, Arneodo Alain y Fofula-Geoirgiou Efi. 2009. “Evidence for inherent nonlinearity in temporal rainfall”. *Advances in Water Resources* 32 (1): 41-48.

Tapoglou, Evdokia, Ioannis C. Trichakis, Zoi Dokou, Ioannis K. Nikolos, y George P. Karatzas. 2014. “Groundwater-level forecasting under climate change scenarios using an artificial neural network trained with particle swarm optimization”. *Hydrological Sciences Journal* 59 (6): 1225–39. doi:10.1080/02626667.2013.838005.

Tran Anh, Duong, Minh Bui, y P. Rutschmann. 2018. “Long Short Term Memory For Monthly Rainfall Prediction In Camau, Vietnam”. Institute of Hydraulic and Water Resources Engineering, Technische Universität München.

Trigo, Ricardo, y Jean Palutikof. 1999. “Simulation of daily temperatures for climate change scenarios over Portugal: a neural network model approach”. *Climate Research* 13: 45–49.

UNFCCC. 2020. “Statistical Downscaling”. Accedido enero 23. https://unfccc.int/files/adaptation/methodologies_for/vulnerability_and_adaptation/application/pdf/statistical_downscaling.pdf.

University of Edinburgh. 2020. “Impact Modelling: Data for Decision Makers”. *The University of Edinburgh*. Accedido enero 23. https://www.ed.ac.uk/sustainability/what-we-do/climate-change/case-studies/climate-research/impact-modelling.

Weart, Spencer R. 2009. *Discovery of Global Warming.* Cambridge, USA: Harvard University Press. http://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=4642429.

Yale. 2020. “Linear Regression”. Accedido enero 23. http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm.