Universidad Tecnológica de la Mixteca

División de Estudios de Posgrado

RECONSTRUCCIÓN DE SERIES DE PRECIPITACIÓN DEL ESTADO DE OAXACA USANDO REDES NEURONALES ARTIFICIALES

TESIS

Para obtener el grado de:

MAESTRA EN TECNOLOGÍAS DE COMPUTO APLICADO

Presenta:

LIC. ROSA MARÍA GUTIÉRREZ APOLONIO

DIRECTORA DE TESIS:

DRA. GABRIELA ÁLVAREZ OLGUÍN

Huajuapan de León, Oaxaca, agosto de 2019.

A mis padres
Enrique J. Gutiérrez Morales
Rosaelia Apolonio Martinez
por su amor y apoyo incondicional.

Agradecimientos

Agradezco principalmente a mi directora de tesis, la Dra. Gabriela Álvarez Olguín por su disponibilidad para dirigir este trabajo, por su paciencia dedicada así como su aporte de experiencia a las revisiones. También agradezco a la Dra. Lluvia Carolina Morales Reynaga por el tiempo dedicado a la realización del proyecto.

Mis más sinceros agradecimientos a mis sinodales al Dr. Arturo Téllez Velázquez y al Dr. Alejandro Ivan Aguirre por sus siempre oportunas e inteligentes observaciones que han contribuido de forma significativa a mejorar este trabajo.

Agradezco a la Dra. Marisol López Cerino no solo por la disponibilidad de revisar el documento y sus conocimientos aportados, también por sus consejos y su apoyo incondicional.

También agradezco al Dr. Felipe de Jesús Trujillo Romero por su motivación y apoyo durante la maestría.

A la Universidad Tecnológica de la Mixteca por haberme dado la oportunidad de realizar mis estudios en sus aulas y a mis profesores, en especial aquellos que hacían su trabajo con verdadero entusiasmo.

A mis padres Rosaelia y Enrique, que sin su apoyo incondicional y cariño hubiese sido difícil terminar mis estudios, porque siempre han estado y me han impulsado en los momentos más difíciles de mi carrera y de mi vida. Ellos que han dado su vida entera, su amor y que han procurado la felicidad de sus hijos antes que la de ellos.

Índice general

Índice de figuras

Índice de tablas

Resumen

Debido a fallas en los instrumentos de medición o por complicaciones causadas por la ocurrencia de eventos meteorológicos extremos, en Oaxaca, así como en la mayor parte del país, es común que las series de tiempo de variables hidrometeorológicas presenten datos faltantes. El análisis de las series de precipitación, es la base de diversos estudios hidrológicos, necesarios para el diseño de obras hidráulicas y la administración de los recursos hídricos. Sin embargo, en México las bases de datos de las Instituciones Gubernamentales a cargo de las estaciones climatológicas, presentan errores o están incompletos y esto puede llevar a la toma de decisiones equivocadas, que podría provocar mayor riesgo de falla en las obras o una mala planeación en la prevención de inundaciones o mitigación de sequías.

Debido a esto, se desea reconstruir las series de precipitación del estado de Oaxaca usando los enfoques de redes neuronales de retropropagación y contrapropagación, utilizando la información de estaciones climatológicas vecinas y analizando las variables climatológicas del estado de Oaxaca. Al tener reconstruida la base de datos de precipitación de Oaxaca, podrá ser utilizada para realizar predicciones de precipitación confiables.

Para emprender este estudio se hizo un análisis de la cantidad y calidad del conjunto de datos de precipitación a cargo del Servicio Meteorológico Nacional. Se detectaron los datos atípicos y se eligió el rango adecuado para realizar la reconstrucción de las series de precipitación. Después se aplicó un Análisis de Componentes Principales a las variables climatológicas disponibles incluyendo las distancias de las estaciones climatológicas al Golfo de México y al Océano Pacifico, con el objetivo de encontrar alguna correlación entre estas variables con la variable de precipitación.

Una vez que se obtuvo la correlación entre las variables, se procedió al agrupamiento de las estaciones climatológicas que correspondan a una misma región hidrológicamente homogénea, con ayuda de la red de Kohonen. Por último, se procedió a la reconstrucción de las series de precipitación aplicando la red de retropropagación y contrapropagación.

Para los resultados y comparativas, se calculó el error cuadrático medio, para evaluar los algoritmos de redes neuronales; y el coeficiente de correlación, para evaluar la confiabilidad de las series de precipitación reconstruidas. Con base a los resultados obtenidos, el tiempo de ejecución de la red de retropropagación es menor al de la red de contrapropagación. Sin embargo, se obtienen mejores resultados de la reconstrucción de las series utilizando la red de contrapropagación.

Capítulo 1

Introducción

La importancia de la predicción de eventos de lluvia recae en que con ello se puede diseñar estrategias para la mitigación de desastres naturales como las sequías e inundaciones, o realizar diseños de obras hidráulicas para el control de escurrimientos (Álvarez-Olguín y Escalante-Sandoval, 2016). Sin embargo, este tipo de predicción es sumamente complicado, porque requiere un registro temporal completo, el cual en la mayoría de casos no lo está, siendo muy común que por una falla durante el período de monitoreo, los datos no se almacenan completamente. (Kim y Pachepsky, 2010). Por lo tanto, reconstruir las series de precipitación donde se presentan datos faltantes es una cuestión clave para la funcionalidad de tales predicciones.

La reconstrucción de las series de precipitación se realiza utilizando información disponible de la misma estación o datos del mismo día de estaciones vecinas. Cuando se utilizan datos de la misma estación, se realiza la reconstrucción de las series mediante una simple interpolación entre los datos disponibles (Lowry, 1972), o el valor medio de la serie de datos (Linacre, 1992), entre otros métodos estadísticos. Por otro lado, cuando se realiza reconstrucción de las series con información de estaciones climatológicas vecinas, algunos de los métodos estadísticos propuestos son:

- Distancias geométricas con base en los datos de las estación más cercana (Xia et al., 1999).
- El promedio aritmético de los datos de varias estaciones vecinas (Willmott et al., 1994).
- Interpolación espacial utilizando correlaciones entre las series de las estaciones vecinas (Filippini et al., 1970).

Muchos de los modelos estadísticos empleados para la reconstrucción de las series de precipitación, han demostrado no ser tan efectivos, esto se debe a que los datos son no estacionarios por las alteraciones climáticas. En consecuencia, se han propuesto diversos estudios basados en redes neuronales artificiales, que son capaces de modelar problemas complejos. Entre algunos estudios está el de Coulibaly y Evora (2007) donde se comparan seis modelos de redes neuronales, o el de Faucher et al. (1999) donde utilizaron árboles de regresión para estimar vientos marinos. En conclusión, para elegir el modelo de red

neuronal que permita obtener una buena estimación, se debe considerar la zona geográfica, la distribución espacial de las observaciones adyacentes, el día, el mes y la temporada de la estación a estimar (Lucio et al., 2007).

El presente trabajo de tesis tiene como objetivo realizar un análisis y la reconstrucción de las series de precipitación, de las estaciones climatológicas del estado de Oaxaca, usando información de estaciones climatológicas vecinas. El análisis de datos consiste en comparar la variable precipitación con las variables climatológicas disponibles (temperatura, granizo, evaporación, tormenta eléctrica, niebla y nublados), a través del modelo de análisis de componentes principales. Para obtener las estaciones climatológicas vecinas, y utilizando el resultado del método anterior, se utiliza la red neuronal de Kohonen, la cual ayuda a clasificar a las estaciones climatológicas en clases con características similares. Dado el problema, también se toman en cuenta aquellas estaciones climatológicas de los estados vecinos de Oaxaca. Para la reconstrucción de las series de precipitación se implementan los aprendizajes de retropropagación y contrapropagación.

1.1. Estado del arte

La reconstrucción de datos faltantes en las series de datos climatológicos, también es conocida como deducción, rellenado o estimación y a ha sido investigada desde el enfoque estadístico e inteligencia artificial (Kim y Pachepsky, 2010). Las series de datos se pueden reconstruir utilizando información de:

1. La misma estación climatológica pero de días anteriores y posteriores a la fecha de interés.
2. De estaciones climatológicas vecinas.

Algunos de los primeros enfoques para resolver este problema son una simple interpolación entre los datos disponibles (Jeffrey et al., 2001, Lowry, 1972) o el promedio de la serie de datos (Linacre, 1992). Otros métodos estadísticos para la reconstrucción de los datos diarios son el pronostico con mínimos cuadrados o regresión no lineal (Acock y Pachepsky, 2000). Estas investigaciones utilizaron información proporcionada de la misma estación, de días anteriores y posteriores al día de interés.

Existen otros trabajos que resuelven el problema con métodos estadísticos, pero a diferencia de los mencionados en el párrafo anterior, en estos se utiliza información de estaciones vecinas. Ramos-Calzado et al. (2008) reconstruyen las series de datos mensuales de precipitación usando regresión lineal con los datos de las estaciones vecinas y estimando el valor faltante como la mediana de las salidas de la regresión lineal. Mientras que Young (1992) utiliza tres modelos de interpolación: (1) el método de regresión lineal múltiple, (2) análisis discriminante múltiple y (3) el método del cociente normal. Por otro lado, Filippini et al. (1970) utilizan un método basado en matrices de correlaciones para elegir a las estaciones vecinas y calculan el valor faltante a partir de aproximaciones autorregresivas.

En los últimos 15 años, las Redes Neuronales Artificiales (RNAs) han mostrado un potencial prometedor en varias aplicaciones hidrológicas, en particular, para reconstruir

datos climatológicos faltantes, por ejemplo, Coulibaly y Evora (2007) comparan seis tipos de RNAs con tres capas (capa de entrada, oculta y de salida): perceptrón multicapa (MLP), feed-forward con retraso de tiempo (TLFN), base radial generalizada (RBF), red recurrente (RNN), red difusa de contrapropagación (CFNN) y red recurrente con retardo de tiempo (TDRNN). Ellos utilizan registros diarios de precipitación y temperatura (máxima y mínima) de 15 estaciones climatológicas ubicadas en la provincia de Quebec, Canadá. La entrada de las RNAs es información de tres a cuatro estaciones climatológicas vecinas las cuales se eligen de la matriz de correlación entre la estación de interés y el resto de las estaciones, la proximidad entre las estaciones y la longitud de registro de los datos disponibles.

Para evaluar el rendimiento de las RNAs Coulibaly y Evora (2007) utilizan el error absoluto medio (Mean Absolute Error, MAE) y el coeficiente de correlación (r) entre la serie diaria observada y reconstruida. Como conclusión, demuestran que el MLP y la TLFN son los más eficaces en la reconstrucción de series de precipitación diaria y temperatura, caso contrario con el RNN y el TDRNN que resultan ser los menos apropiados para este caso de estudio.

Por otro lado, Devi et al. (2016) utilizan: retropropagación (BPN), retropropagación en cascada (CBPN), red de tiempo distribuido (DTDNN) y red exógena autorregresiva no lineal (NARX), para la reconstrucción de series de precipitación, temperatura y humedad diarios del estado de Tamil Nadu en India, de más de 14 estaciones climatológicas del 2001 al 2012. Primero aplican una normalización (mín-máx) para escalar los datos en el intervalo (0,1), después utilizan autocorrelación para analizar la relación entre las lluvias del día de interés con días anteriores y el análisis de sensibilidad para elegir a los predictores claves. La función de activación de las RNAs es una sigmoidal en la capa oculta y una función lineal en la capa de salida. Cerca del 70 % de los datos son utilizados como conjunto de entrenamiento para la red y el resto como el conjunto de prueba. Para medir el desempeño de la RNA utilizan el error cuadrático medio (mean squared error, MSE) y el coeficiente de correlación para comparar las predicciones. Con base en el análisis de rendimiento, el modelo NARX superó al resto aunque BPN obtuvo resultados muy buenos.

Lucio et al. (2007) utilizan una red con tres capas con los entrenamientos de retropropagación y propagación rápida para la reconstrucción de series de precipitación en el estado de Río Grande en Brasil. Toman en cuenta las series pluviales mensuales totales registradas en el período de 1961 a 2005, con un total de 14 estaciones climatológicas. Dado que cuatro estaciones climatológicas presentan menor número de valores perdidos, estas se utilizan como estaciones de entrenamiento. Para los datos de entrada también utilizan estaciones climatológicas vecinas elegidas con el coeficiente de correlación y la distancia entre ellas. Con las pruebas realizadas, lograron demostrar que la red presenta una respuesta óptima en la fase de entrenamiento, sin embargo, durante la fase de validación y prueba, la red sobre estimó los valores observados. Además, entre los dos métodos de entrenamiento, propagación rápida es inestable e inclinado a estancarse en mínimos locales.

Luk et al. (2000) utilizan una RNA multicapa con retraso (Feed-forward modificada con procesos Markovianos) para identificar un conjunto óptimo de entradas para el pronóstico

de precipitaciones a corto plazo, aplicado a los 16 pluviómetros de la cuenca urbana Parramatta situada en Sydney, Australia, en el período de 1991 a 1996. Además, utilizan información de pluviómetros vecinos desde 2 hasta 15 tomando en cuenta la distancia más corta. Con las pruebas realizadas concluyen que la red con mejor rendimiento es la de menor orden de retraso.

Con el objetivo de mantener el agrupamiento estacional, Khalil et al. (2001) desarrollan un modelo de segmentación a los datos denominado modelo "grupo-valor", que permite agrupar datos con atributos similares y así conocer estaciones vecinas. Este modelo lo combinan con RNAs lo cual da lugar a dos modelos: el modelo auto variable de múltiples capas feed-forward con retardo de orden uno (MASM) y el modelo bivariado de múltiples capas feed-forward (BSM). Estos modelos son aplicados a 10 caudales localizados en Canadá en un periodo de 1963 a 1995. Para medir la eficiencia de los modelos utilizan el error medio relativo (MRE) con lo cual concluyen que el modelo MBSM obtiene mejores resultados, y así indican que las técnicas bivariadas son más adecuadas para los propósitos de reconstrucción de datos de caudales.

Kim y Pachepsky (2010) obtienen mejores resultados si combinaban RNAs con árboles de regresión (AR), para este caso, los AR se utilizan para determinar los datos de entrada para la red. El área de estudio está localizada en la parte central de la Bahía de Chesapeake, que consta de 39 estaciones climatológicas en el periodo de 2001 a 2007. Para evaluar la eficiencia del modelo y de las series reconstruidas se utiliza el coeficiente de correlación y el MSE.

Por otro lado, Faucher et al. (1999) utilizan el método CANFIS para reconstruir los datos de vientos marinos, dicho método es una combinación de métodos de clasificación, árboles de regresión y RNAs difusas. Los AR son utilizados para seleccionar a los predictores relevantes. El objetivo de este estudio es reconstruir vientos marinos superficiales de 13 sitios de boyas canadienses a lo largo de la costa de la Columbia Británica (BC) durante el período de 1958 a 1997.

Además de utilizar la matriz de correlación para la elección de estaciones climatológicas vecinas, también se han propuesto utilizar distancias geométricas entre las estaciones y emplear datos de la estación más cercana (Xia et al., 1999), así como el promedio aritmético de los datos de varias estaciones vecinas (Willmott et al., 1994). Estudios más actuales como Teegavarapu and Chandramouli (2005) utilizan el método de proximidad del polígono, el coeficiente de correlación, el vecino más cercano y el método del exponencial inverso, todos estos métodos calculan la distancia inversa entre las estaciones.

Se puede observar que en todos estos estudios utilizan diferentes enfoques para la reconstrucción de series de precipitación, en particular se obtienen resultados óptimos combinando RNAs con información de estaciones climatológicas vecinas. Además, el entrenamiento que destaca en muchos de los estudios es retropropagación ya que es fácil de programar y proporciona buenos resultados. Sin embargo, la elección de la metodología depende de varios factores: la variable climatológica considerada, la zona geográfica, la distribución espacial de las observaciones vecinas además del día, mes y temporada (Lucio et al., 2007).

1.2. Planteamiento del Problema

El registro de datos observados sobre precipitaciones es de suma importancia, ya que con ellos es posible desarrollar una adecuada gestión del agua, por ejemplo, a partir del tratamiento e interpretación de los datos es posible conocer la disponibilidad de este recurso en un determinado lugar, para minimizar los problemas derivados de su uso (Solís and Rivera, 2004). Con estos datos, se puede modelar y predecir la cantidad de precipitación (Coulibaly y Evora, 2007), la simulación del impacto del cambio climático en los sistemas de agricultura (Serrano Altamirano et al., 2005), el diseño de obras hidráulicas para el control de escurrimientos (Álvarez-Olguín y Escalante-Sandoval, 2017), el diseño de presas o puentes, entre otros. Una subestimación de la disponibilidad de agua implica fallos en las estrategias de asignación de este recurso a las personas, lo cual, en tiempos prolongados de sequías es un gran problema (Álvarez-Olguín y Escalante-Sandoval, 2016).

Sin embargo, para realizar dichas tareas se debe contar con datos que puedan proporcionar registros periódicos que tengan continuidad en las observaciones. Además, la Organización Meteorológica Mundial (WMO, por sus siglas en inglés, World Meteorological Organization) recomienda que se cuente con datos registrados durante 50 años continuos, como mínimo para realizar predicciones (Kundzewicz et al., 2000). Sin embargo, es común encontrarse con el problema de la falta de continuidad de las bases de datos disponibles que pueden afectar al resultado obtenido, como es el caso de las estaciones climatológicas que por diferentes motivos provocan la ausencia de datos, entre ellos: variabilidad de la precipitación, la falla del equipo por falta de mantenimiento o algún desastre natural (huracanes, temblores, etc.), la falla en el proceso de medición o su destrucción por parte de animales silvestres (Antelo y Long, 2014, Luk et al., 2000).

Para el caso de Oaxaca, el Servicio Meteorológico Nacional cuenta con información de diferentes estaciones climatológicas en distintas regiones del estado. Pero dicha información varía en el inicio y fin de captura de los datos. Además, en el estado de Oaxaca como en la mayor parte del país, los registros de precipitación comúnmente presentan discontinuidades, o son de corta duración.

Por lo tanto, el objetivo de este trabajo de tesis es ayudar a resolver el problema de reconstrucción de las series de precipitación en el estado de Oaxaca, México. Dados los resultados obtenidos en diferentes investigaciones, se utiliza el enfoque de redes neuronales artificiales, en particular, el aprendizaje de retropropagación y contrapropagación. La base de datos se obtiene del CLImate COMputing project (CLICOM, `http://clicom-mex.cicese.mx/`) a cargo del Servicio Meteorológico Nacional (SMN). Esta base de datos tiene información nacional de todos los estados en México, en particular del estado de Oaxaca existe información sobre 351 estaciones climatológicas en distintas regiones del estado, almacenados en hojas de cálculo Excel. La información almacenada es diaria sobre: temperatura observada, temperatura máxima, temperatura mínima, precipitación diaria, evaporación, tormenta eléctrica, granizo, niebla y nublados. Los datos capturados no inician ni terminan en la misma fecha para cada estación climatológica, lo cual ocasiona un problema. El inicio de captura en general del estado de Oaxaca va desde 1922 hasta 2004 y el fin de captura va desde 1969 hasta 2009, por lo que se debe elegir un rango de años en el cual se hará la reconstrucción de los datos. Además, hay datos en

los que su valor es dudoso, inapreciable o es estimado.

1.3. Justificación

Es importante contar con un registro continuo de datos de precipitación para realizar investigaciones y manejo de la información para la toma de decisiones en materia de recursos, costos de producción, entre otros. En particular, tener este registro ayuda a conocer la disponibilidad y calidad del agua, la simulación del impacto del cambio climático en los sistemas de agricultura, industrial y energético (Coulibaly y Evora, 2007). Además permiten analizar las corrientes de lluvia lo cual ayuda a las zonas que requieran de alcantarillados para evitar inundaciones. También, ayudan a tomar precauciones en carreteras o poblaciones en zonas montañosas propensas a derrumbes por causa de intensas lluvias (Almeida Román, 2010).

Para realizar una buena gestión del agua y obtener resultados que se ajusten a la realidad del fenómeno se debe contar con datos que puedan proporcionar registros periódicos y además, tengan continuidad en las observaciones. Sin embargo, por diferentes factores es difícil contar dichos registros, debido a esto se opta por técnicas que ayuden a la reconstrucción de series de precipitación.

Para resolver la problemática de la reconstrucción de datos se han realizado diversos estudios, principalmente estadísticos. Sin embargo, debido al comportamiento no estacionario que presentan los datos de precipitaciones, algunos métodos estadísticos como regresión lineal o interpolación no han dado buenos resultados (Creutin et al., 1997). También, los datos son no estacionarios debido a los diferentes cambios climáticos que existen, como los huracanes que se han vuelto más intensos y frecuentes y son una de las principales causas de lluvias y escorrentías[1] de mayor magnitud (Álvarez-Olguín y Escalante-Sandoval, 2017).

Algunas investigaciones sobre reconstrucción de series de precipitación en México basados en aproximación por métodos estadísticos son: Alvarado Medellin (2007) que analiza los datos de la cuenca del río Amajac en Hidalgo, Ruíz Corral et al. (2006) y Ruíz Corral et al. (2007) analizan los datos obtenidos por las estaciones en los estados de Baja California Sur y Guanajuato, respectivamente. Pero para el estado de Oaxaca hay pocos estudios sobre la reconstrucción de series de precipitación, en Serrano Altamirano et al. (2005) tienen como objetivo generar información actualizada climatológica del estado de Oaxaca y para ello realizan la reconstrucción de datos con el programa CLIMGEN, el cual se basa en estadísticas generadas a partir de los datos diarios. Otros estudios solo tratan sobre predicción (Bustamante, 2013) o análisis de hidrografía.

Últimamente, se ha mostrado que los métodos de RNAs han dado mejores resultados en la reconstrucción de series de precipitación (Coulibaly y Evora, 2007, Devi et al., 2016, Faucher et al., 1999, Lucio et al., 2007), debido a que son herramientas de modelado de datos estadísticos no lineales. Muchos estudios han desarrollado modelos complejos para poder resolver el problema de reconstrucción de series (Khalil et al., 2001, Kim

[1] Agua de lluvia que circula libremente sobre la superficie de un terreno

y Pachepsky, 2010) pero otros han optado por los modelos clásicos: retropropagación o feed-forward (Coulibaly y Evora, 2007, Luk et al., 2000), que además de su facilidad para programarlos también obtienen aproximaciones aceptables, pero la elección del modelo depende de diferentes factores (Lucio et al., 2007).

En la actualidad no existe en el estado de Oaxaca un algoritmo que ayude en la reconstrucción de series de precipitación, se pretende seleccionar un modelo de red neuronal que permita estimar los datos faltantes en la base de datos de precipitación, con el fin de contar con series completas de datos y con esto poder ayudar al desarrollo de otras investigaciones. Se pretende que las series reconstruidas de precipitación ayuden al trabajo futuro del instituto de Hidrología de la Universidad Tecnológica de la Mixteca y profesionales que puedan utilizar los resultados de la presente tesis.

Además, varias investigaciones en reconstrucción de series de precipitación utilizando el enfoque de redes neuronales han obtenido buenos resultados utilizando como entrada información de estaciones meteorológicas vecinas, por lo que se pretende utilizar esta estrategia y para ello se buscan dichas estaciones a partir de la categorización proporcionada por la red neuronal de Kohonen.

1.4. Hipótesis

Es posible que los enfoques de redes neuronales de retropropagación y contrapropagación, junto con la información de estaciones climatológicas vecinas, puedan ayudar a reconstruir las series de precipitación en el estado de Oaxaca.

1.5. Objetivos

1.5.1. Objetivo general

Reconstruir las series de precipitación del estado de Oaxaca usando los enfoques de redes neuronales de retropropagación y contrapropagación, utilizando la información de estaciones climatológicas vecinas y analizando las variables climatológicas del estado de Oaxaca.

1.5.2. Objetivos específicos

1. Realizar un análisis de las variables climatológicas mediante el análisis de componentes principales para encontrar correlaciones con la variable precipitación.
2. Encontrar las estaciones climatológicas vecinas de las estaciones objetivos que ayuden a la reconstrucción de las series de precipitación por medio de mapas autoorganizados SOM.

3. Implementar los modelos de retropropagación y contrapropagación para la reconstrucción de series de precipitación en el estado de Oaxaca.

1.6. Metas

Para poder alcanzar los objetivos fijados se presentan las siguientes metas:

1. Análisis de los datos para encontrar datos atípicos.
2. Análisis estadístico para encontrar el rango de años para el cual se hará la reconstrucción de las series de precipitación.
3. Localización de las series de datos que ayudarán a entrenar a los modelos de redes neuronales artificiales y la serie de datos que serán parte del conjunto de prueba.
4. Reporte de como influyen el resto de las variables climatológicas con la variable de precipitación a partir del análisis PCA.
5. Implementación de la red SOM en el lenguaje Python para elegir a las estaciones climatológicas vecinas que brinden mayor información en la reconstrucción de series de precipitación.
6. Implementación de los modelos propuestos para la reconstrucción de series de tiempo, retropropagación y contrapropagación en el lenguaje Python utilizando la información obtenida en los puntos anteriores.
7. Comparativa entre los algoritmos propuestos para la reconstrucción de series de precipitación a partir del error cuadrático medio.
8. Comparativa entre la series de precipitación obtenida por los algoritmos implementados y la observada, para analizar la confiabilidad de los datos mediante el coeficiente de correlación.
9. Elaboración del documento de tesis.
10. Elaboración de al menos un artículo, divulgación o indexado.

Capítulo 2

Marco Teórico

En este capítulo se describen los métodos utilizados para la reconstrucción de las series de precipitación. En la sección 2.1 se define lo que es un dato atípico así como la prueba de Grubbs que nos permite detectar a los datos atípicos que se presentan en nuestra base de datos. En la sección 2.2 se presenta la teoría del análisis de componentes principales que ayuda a encontrar correlaciones entre las variables climatológicas de nuestra base de datos. Por último, en la sección 2.3 describimos la teoría de redes neuronales artificiales. Además, esta sección se divide en tres subsecciones: retropropagación, mapas auto-organizados y red de contrapropagación.

2.1. Datos atípicos

Cuando se trabaja con un conjunto de datos reales con frecuencia ocurren en ellos hechos puntuales que desconocemos. Por ejemplo, los datos pueden haber estado sometidos a intervenciones desconocidas como errores de medición. Las observaciones afectadas por estas intervenciones pueden presentar una estructura distinta de las demás, es decir, no mantienen la estructura de dependencia del resto de los datos. A estos datos se les denomina datos atípicos y varios autores han propuesto una definición:

- Un dato atípico es una observación o conjunto de observaciones que parecen ser inconsistentes con el resto del conjunto de datos (Barnett y Lewis, 1984).
- De forma similar Hawkins (1980) define un dato atípico como una observación que se desvía mucho de otras observaciones y despierta sospechas de ser generada por un mecanismo diferente.
- Beckman y Cook (1983) se refieren a los datos atípicos ya sea como observaciones discordantes o como contaminantes. Una observación discordante es cualquier observación sorpresiva o discrepante para el investigador. Un contaminante es cualquier observación que no es parte de la distribución objetivo.

A los datos atípicos también se les conocen como outliers, anomalías, observaciones discordantes, excepciones, fallas, defectos, ruido, errores, contaminantes, valores extremos

entre otros. La detección de datos atípicos ha sido ampliamente investigada y ha sido aplicada en diferentes áreas tales como detección de fraudes con tarjetas de crédito, seguros o asistencia en salud, detección de intrusos en sistemas de cómputo, aprobación de préstamos para detectar clientes potencialmente problemáticos, limpieza de datos, predicción meteorológica, etc. (Chandola et al., 2007). Los datos atípicos no pueden ser caracterizados como benéficos o problemáticos sino que deben ser contemplados en el contexto del análisis y debe evaluarse el tipo de información que pueden proporcionar.

Es importante identificar datos atípicos debido a que:

- Si sus efectos son grandes pueden sesgar la estimación de los parámetros, lo que conduce a malas predicciones futuras.
- Si el suceso ha ocurrido en los últimos datos y alguna observación afectada se utiliza para generar predicciones estas no serán buenas, aunque los parámetros estén bien estimados.
- Si se identifican correctamente los datos atípicos y estimamos sus efectos, en el momento que estos aparezcan en el futuro, se puede incorporar esta información en las predicciones y obtener intervalos de predicción más realistas.
- El no identificar a los datos atípicos puede conducir a decisiones erróneas con consecuencias como pérdida de dinero, tiempo y credibilidad.
- Aunque los datos atípicos pueden aparentar ser inválidos pueden ser correctos y viceversa.
- Su principal problema radica en que son elementos que pueden no ser representativos de la población pudiendo distorsionar seriamente el comportamiento de los contrastes estadísticos.
- Aunque los datos atípicos son diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de un segmento válido de la población y, por consiguiente, una señal de la falta de representación de la muestra.

Peat and Barton (2005) establecen la existencia de dos tipos de datos atípicos: Univariantes y Multivariantes:

Dato atípico univariante: son aquellos datos muy diferentes al resto para una sola variable. Por ejemplo, un partido de fútbol con una diferencia de goles de 50, en este caso solo se considera la variable "gol".

Dato atípico multivariante: es aquel dato extremo para una combinación de variables. Por ejemplo, un caso inusual es un niño de 8 años cuya estatura sea de 1.55 m y su peso 45 kg, en este ejemplo se tiene un dato atípico considerando tres variables (edad, estatura y peso).

Los datos atípicos multivariantes son los más difíciles de identificar y se usan estadísticos como la distancia de Cook o la de Mahalanobis. Por otro lado, Chandola et al. (2007) clasifican los datos atípicos en tres tipos de a cuerdo a su composición y relación con el resto de los datos:

Tipo I: Son aquellos que corresponden a instancias individuales de los datos. Por ejemplo, retiros realizados por usuarios en cajeros de las 6 am a 12 pm, si la mayoría de los movimientos se encuentran entre \$100 y \$500, y solo hay dos retiros mayores a \$500 (como \$5000) serán datos extremos.

Tipo II: Son datos individuales pero se definen con respecto a un contexto específico. Por ejemplo, en diciembre en Huajuapan se ha registrado temperaturas de $9°C$ pero esta temperatura en abril sería una anomalía.

Tipo III: Constituyen un subconjunto de datos que se encuentra por fuera del conjunto total de datos. Estos son significativos solamente cuando se trata de datos espaciales o de naturaleza secuencial. Por ejemplo, el electrocardiograma humano en el cual se presenta una línea extendida durante un tiempo prolongado corresponde a un valor atípico, pues a pesar de que existen valores igualmente bajos, no lo son por un período de tiempo tan prolongado.

Los datos extremos del tipo I pueden ser detectados en cualquier tipo de conjuntos de datos, mientras que los de tipo II y III necesitan la presencia de una estructura secuencial o espacial en los datos. Se debe tener en cuenta que, en los datos se pueden encontrar datos atípicos mínimos así como máximos y, tienen diferentes efectos en el análisis. Los valores extremadamente altos desplazan la media hacia la derecha, mientras que los valores extremadamente bajos atraen la media hacia a la izquierda. Esto ocasiona que la media pierda su cualidad de representar el punto medio.

Desde tiempo atrás se vienen desarrollando técnicas con diferentes enfoques para identificar observaciones candidatas para su eliminación o sustitución, como diagramas de caja, prueba de Grubbs, métodos basados en proximidad como: el vecino más cercano, regresión, análisis de componentes principales, máquina de soporte vectorial, redes neuronales, entre otros (Uribe, 2010). Además se han realizado algunos trabajos comparativos que pueden servir para determinar las técnicas adecuadas para una determinada situación. Sin embargo, no son una guía general para elegir la técnica adecuada. La elección de la técnica de detección y corrección de valores atípicos es de suma importancia pero no es una tarea sencilla ya que a veces el dato erróneo es muy similar al resto de los datos; pero la técnica elegida debe cumplir las siguientes características (Chandola et al., 2007):

- Debe identificar a los valores extremos de acuerdo con las características, requisitos, limitaciones y la naturaleza de los datos.
- Debe ser capaz de detectar el tipo de valores atípicos que se presentan en el problema.
- Debe obtener óptimos resultados en términos de precisión así como de eficiencia computacional.

Al elegir la técnica de detección de datos atípicos, es importante que dicha técnica no esté afectada por el efecto de enmascaramiento (masking effect) y el efecto inundación (swamping effect). El primer efecto se refiere a que valores bastante alejados del centro de la nube de datos, no aparezcan como atípicos. Por el contrario, el efecto inundación se refiere a que algunos datos pueden parecer atípicos cuando no lo son, simplemente porque están alejados de la media (Uribe, 2010).

Prueba de Grubbs

La prueba de Grubbs o el método Extreme Studentized Deviate (ESD) fue planteado por Frank E. Grubbs desde el año 1969 (Grubbs, 1969). La prueba de Grubbs se utiliza para detectar datos atípicos en un conjunto de datos univariante además los datos deben aproximarse a una distribución normal antes de aplicar la prueba. Este método sirve para detectar un valor atípico a la vez y funciona bien en tamaños de muestras muy grandes (Uribe, 2010).

La prueba de Grubbs es muy utilizada en el área de hidrología (Pérez, 2012) y permiten identificar el limite inferior y superior de los datos, X_L y X_H respectivamente, definidos por la ecuación (2.1) y (2.2).

$$X_L = \exp\left(\overline{X} - K_N \cdot S\right) \tag{2.1}$$

$$X_H = \exp\left(\overline{X} + K_N \cdot S\right) \tag{2.2}$$

Donde $\overline{X}$ y S son la media y desviación estándar de los logaritmos naturales de la muestra, K_N es la estadística de prueba que depende del tamaño de la muestra N y se puede consultar en (Grubbs and Beck, 1972). Los valores que estén fuera del intervalo $[X_L, X_H]$ se consideran datos atípicos.

2.2. Análisis de componentes principales

El análisis de componentes principales (PCA, Principal Components Analysis) es una técnica estadística de reducción de dimensión, es decir, dado un conjunto de datos con n observaciones y p variables correlacionadas, $X = (X_1, X_2, \ldots, X_p)$, PCA permite resumir al conjunto X en un conjunto $Z = (Z_1, Z_2, \ldots, Z_q)$ con el menor número de variables $(q < p)$ que contiene la mayor cantidad posible de la variación (James et al., 2014).

Así, PCA es una técnica no supervisada que permite calcular las componentes principales y una herramienta que permite visualizar las observaciones o las variables. Ya que es muy difícil visualizar las p variables de X en $\mathbb{R}^p$, por lo que en algunos casos se opta por hacer gráficos de dispersión entre cada dos variables. Sin embargo, en total se tienen $\binom{p}{2}$ gráficos de dispersión, que para p grande es tedioso la visualización de los datos como se puede observar en la Figura 2.1. Claramente, es mejor encontrar una representación de los datos de menor dimensión que capture la mayor información como sea posible, de preferencia en dos o tres dimensiones que es fácil de visualizar.

La idea de reducir la dimensión se basa en que cada una de las n observaciones viven en el espacio p-dimensional, pero no todas estas dimensiones son igualmente interesantes. Por lo que cada una de las variables encontradas por PCA es una combinación lineal de las características de p. Esto es, que existe un vector de escalares $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})$ tal

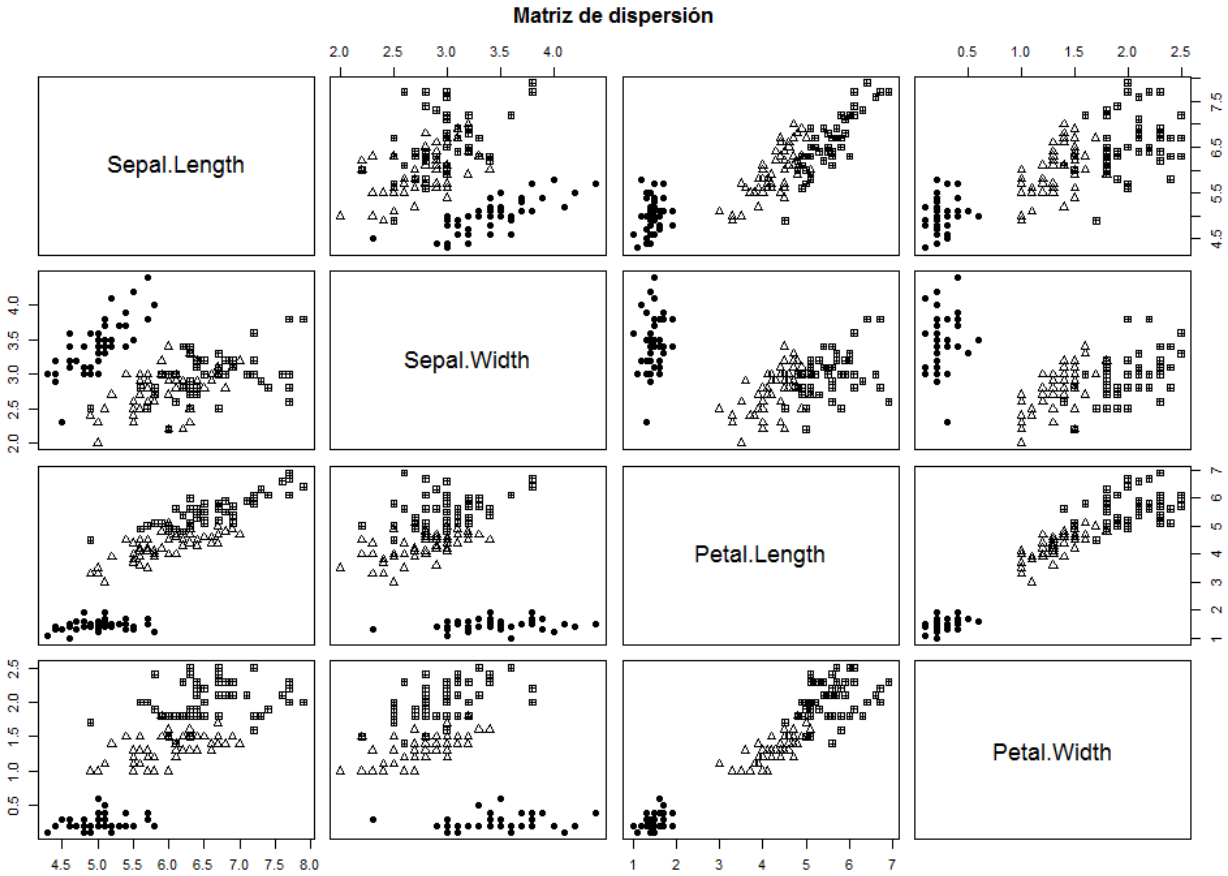


Figura 2.1: Matriz de dispersión Sepal y Petal de la base de datos Iris, obtenida con el proyecto estadístico R-studio.

que la primer componente principal Z_1 de un conjunto de características $X_1, X_2, \ldots, X_p$ es la combinación lineal normalizada de estas características:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p, \tag{2.3}$$

que tiene la mayor varianza y normalizando el vector de escalares se debe cumplir que $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. Entonces, para encontrar a la primera componente principal se debe resolver el problema dado por 2.4:

$$\begin{aligned} &\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximizar}} \quad \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\}, \\ &\text{sujeto a} \qquad \sum_{j=1}^{p} \phi_{j1}^2 = 1. \end{aligned} \tag{2.4}$$

De forma similar se encuentra la segunda componente principal Z_2, que es la combinación lineal de las variables $X_1, \ldots, X_p$ con la varianza máxima de todas las combinaciones

lineales que no están correlacionadas con Z_1:

$$Z_2 = \phi_{12}X_1 + \phi_{22}X_2 + \ldots + \phi_{p2}X_p. \tag{2.5}$$

El decir que Z_2 no está correlacionada con Z_1 implica que ϕ_2 es ortogonal a ϕ_1, donde $\phi_2 = (\phi_{12}, \phi_{22}, \ldots, \phi_{p2})$. Los vectores ϕ_i, para $i = 1, \ldots, q$, son las direcciones en el espacio de características a lo largo de las cuales los datos varían más, y las componentes principales se proyectan a lo largo de estas direcciones. También, las componentes principales proporcionan superficies lineales de baja dimensión que están más cerca de las observaciones. Esto quiere decir, que la primera componente principal es la línea en el espacio p-dimensional más cercana a las n observaciones como se puede observar en la Figura 2.2. Las líneas punteadas indican la distancia entre cada observación y el primer vector de la componente principal. Con esto, se busca una dimensión única de los datos que se encuentre lo más cerca posible de todos los datos y que proporcione un buen resumen de ellos.

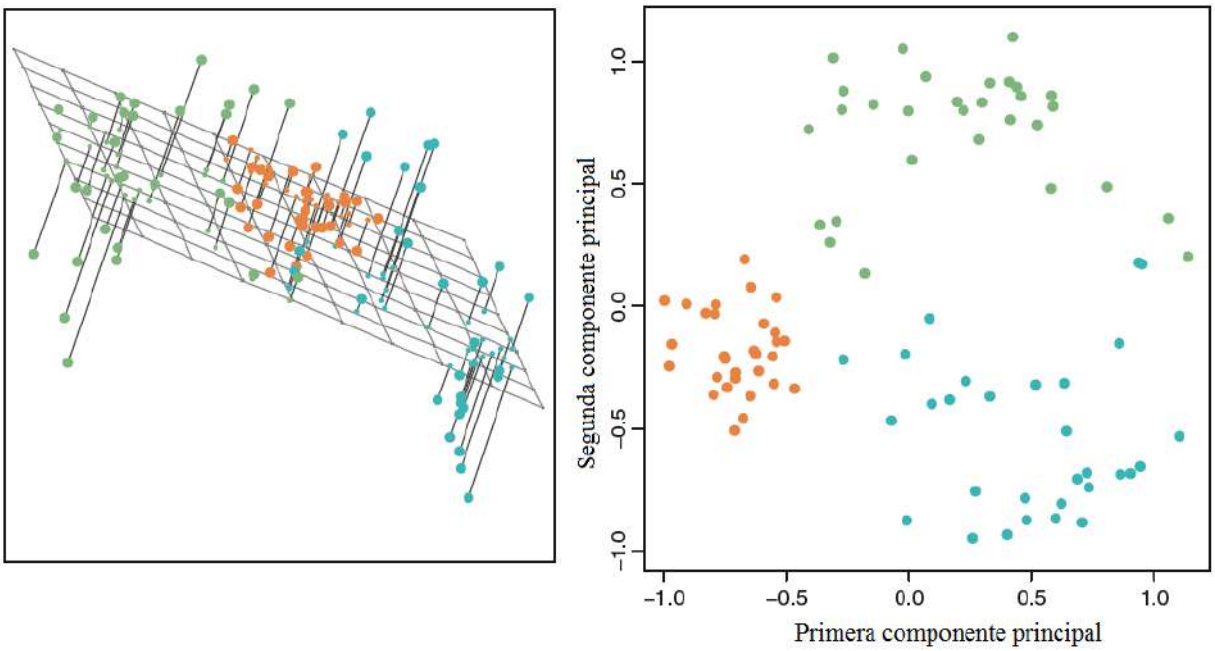


Figura 2.2: La imagen de la izquierda son datos en tres dimensiones y muestra las dos primeras direcciones de las componentes principales que abarcan el plano que mejor se ajusta a los datos, minimizando la distancia de cada punto al plano. En la imagen de la derecha se observan los datos proyectados al plano con la varianza máxima. Imagen obtenida de James et al. (2014).

Para encontrar la tercera componente principal se sigue un procedimiento similar al anterior, y el vector de escalares ϕ_3 resulta ser ortogonal a ϕ_1 y ϕ_2, de forma similar para las siguientes componentes principales. El número q de componentes principales depende de la dimensión a la que se quieren proyectar los datos.

2.3. Redes neuronales artificiales

La idea de las Redes Neuronales Artificiales (RNAs) fue formulada por Warren McCulloch y Walter Poitts en 1943 y tienen el objetivo de modelar la forma de procesamiento de la información, en sistemas nerviosas biológicas, es decir, tratan de modelar el proceso de una red neuronal biológica (Perez Aguila, 2012). En la Figura 2.3 (a) se puede observa la estructura de una neurona biológica mientras que en la Figura 2.3 (b) es la representación de una neurona artificial.

El cerebro humano es un sistema altamente complejo, no lineal y paralelo, es decir, realiza diferentes tareas simultáneamente a diferencia de una computadora común que son de tipo secuencial. Una RNA es un procesador de información, de distribución altamente paralela, constituido por muchas unidades sencillas de procesamiento llamadas neuronas. Además, estas dos clases de redes neuronales comparten las características: adaptabilidad y auto-organización, es decir, si en el proceso se pierden neuronas especializadas en una determinada tarea, la red puede obligar a otras neuronas a especializarse en dicha tarea y continuar teniendo un buen comportamiento. Con esto, las neuronas son capaces de cambiar de forma dinámica con el medio y se preparan para recibir y procesar nuevas entradas y dar las salidas correspondientes.

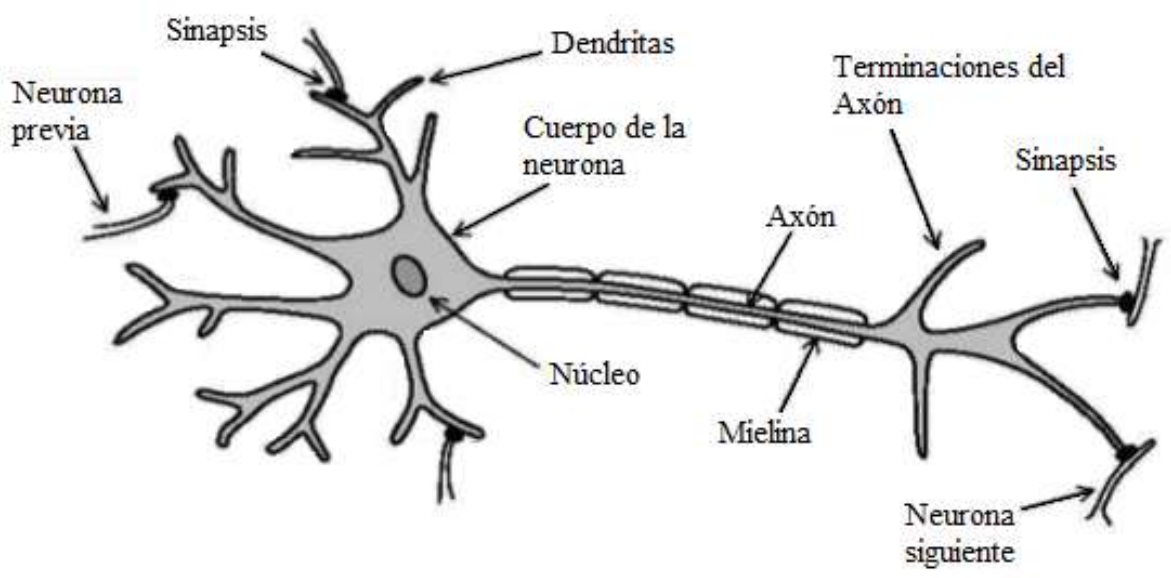
Cada una de las neuronas que componen a la RNA utilizan únicamente operaciones computacionales elementales: suma, multiplicación y operadores lógicos. Esta característica hace que las RNAs sean sencillas de implementar.

La estructura de una red neuronal biológica está dada por (ver Figura 2.3 (a)):

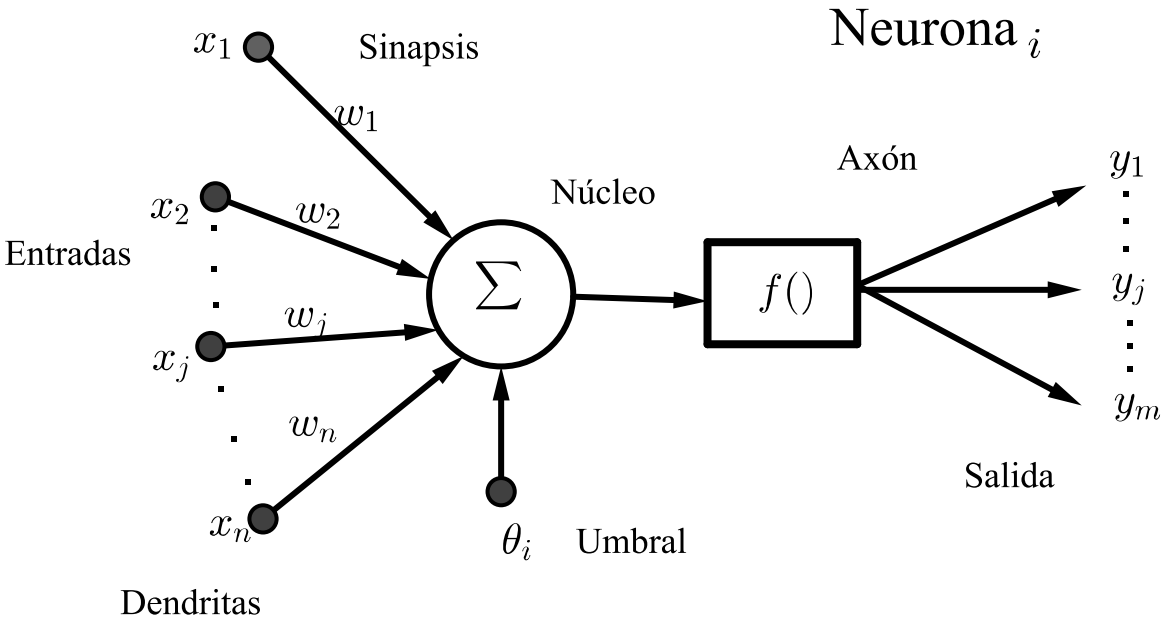
- Dendritas: ramificaciones a través de las cuales una neurona recibe las señales de entrada.
- Núcleo: donde se realizan la mayor parte de los procesos neuronales.
- Axón: prolongación que sale del núcleo de la neurona y termina en una ramificación que tiene como objetivo propagar la señal.
- Región pre-sináptica: es el conjunto de ramificaciones de un axón.
- Sinapsis: es la interconexión entre dos neuronas.

Los mensajes que transmiten y que reciben las neuronas son impulsos nerviosos que se mueven primero a las dendritas, luego al núcleo, después al axón y por último a la región pre-sináptica, es decir, a la sinapsis. Una vez que el impulso nervioso alcanza a una sinapsis se libera una sustancia química denominada neurotransmisor, esta sustancia provoca los cambios de permeabilidad en la membrana de la neurona receptora y como consecuencia el impulso nervioso pasa a la siguiente neurona. Finalmente, el neurotransmisor es destruido y se libera el espacio sináptico.

Las neuronas pueden tener miles de sinapsis por lo que puede transmitir su impulso a otras miles de neuronas, más aún puede conectarse consigo misma, lo que se conoce como retroalimentación. Es importante destacar que no todas las conexiones entre neuronas tienen los mismos pesos, dicho peso determina si la membrana de la neurona que recibe el impulso lo transmitirá o no.



(a) Neurona biológica



(b) Neurona artificial

Figura 2.3: Representación de una neurona biológica y una red neuronal artificial.

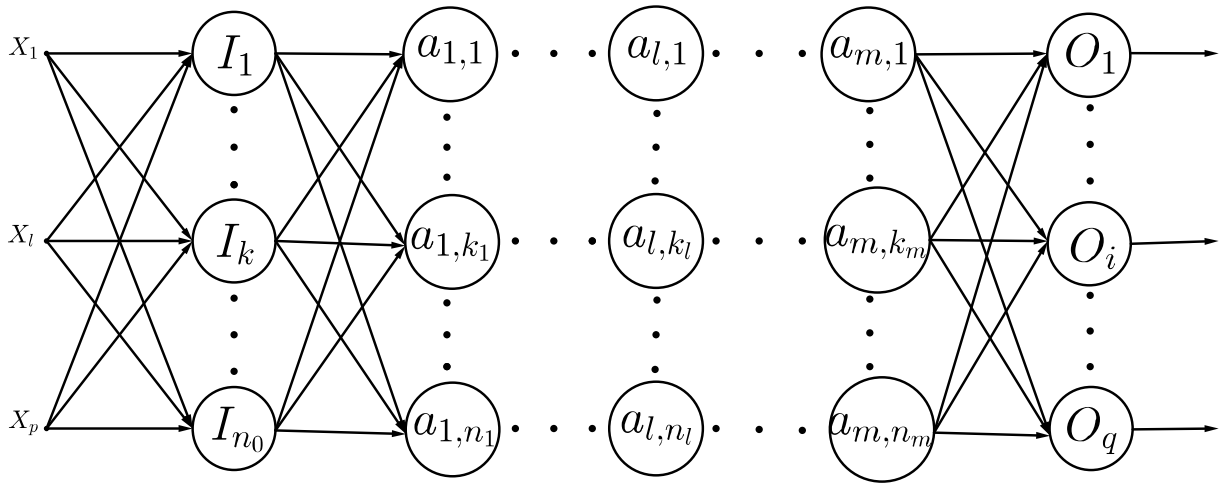


Figura 2.4: Representación de una red neuronal con múltiples capas.

Con estas características se puede visualizar a una neurona (ya sea biológica o artificial) como una entidad con n entradas (dendritas) representado por un vector con n componentes $[x_1, \ldots, x_n]$. Cada entrada x_j, para $j = 1, \ldots, n$, recibe un valor que proviene de otra neurona o del mundo exterior y está asociado a un peso w_j. Cada par (x_j, w_j), para $j = 1, \ldots, n$, es llevado al núcleo donde se realiza el cómputo neuronal y se obtiene como resultado un único valor que se distribuye en m salidas (sinapsis). Dichas salidas se pueden conectar con otra neurona o con el mundo exterior (ver Figura 2.3 (b)).

Una red neuronal (biológica o artificial) está formada por varias neuronas interconectadas agrupadas en capas. En la Figura 2.4, se puede observar una representación gráfica de una red neuronal la cual recibe un conjunto de entradas formado por p vectores: $\{X_1, X_2, \ldots, X_p\}$ donde,

$$X_l = \begin{bmatrix} x_{l,1} \\ \vdots \\ x_{l,k} \\ \vdots \\ x_{l,p} \end{bmatrix},$$

con $x_{l,k} \in \mathbb{R}$, para $l = 1, \ldots, p$ y $k = 1, \ldots, p$. La estructura de una red neuronal está formada por tres tipos de capas:

Capa de entrada: es el conjunto de neuronas que recibe información proveniente del conjunto de entrada. Cada neurona en esta capa se denota por I_k, para $k = 1, \ldots, n_0$ y n_0 es el número de neuronas.

Capas ocultas: conjunto de neuronas internas a la red y no tienen contacto directo con el entorno exterior. El número de capas ocultas m, puede variar dependiendo del problema a resolver y las neuronas en estas capas pueden estar interconectadas de distintas formas. Esta última característica junto con su número determina las

distintas topologías de redes neuronales. Cada capa oculta de la Figura 2.4 consta de n_l neuronas y cada neurona se denota por a_{l,k_l}, $1 \leq l \leq m$ y $1 \leq k_l \leq n_m$.

Capa de salida: conjunto de neuronas que proporciona la salida final obtenida por la red. Esta capa esta constituida por q neuronas y se denotan por O_i, para $i = 1, \ldots, q$.

Dado que una neurona biológica tiene un "estado de activación", es decir, puede estar activa o inactiva, para las neuronas artificiales es similar, estas neuronas no solo tienen dos estados, otras pueden tomar cualquier valor dentro de un conjunto determinado. Para modelar el estado de activación se utiliza una función g diferenciable en los $\mathbb{R}$, es decir, g hace el papel de un neurotransmisor. La función de activación g recibe la suma de los pesos multiplicados por sus respectivas entradas y el resultado es enviado a la siguiente capa. Algunos ejemplos de funciones de activación son:

- Función identidad:
$$g(z) = z.$$
- Función sigmoide:
$$g(z) = \frac{1}{1 + \exp(-z)}.$$
- Función tangente hiperbólica:
$$g(z) = \tanh(z) = \frac{1 - \exp^{-2z}}{1 + \exp^{-2z}}.$$

Pero si solo se quiere indicar que la neurona está activa o inactiva se pueden utilizar las siguientes funciones:

- Función escalón:
$$g(z) = \begin{cases} 1, & \text{si } z \geq 0, \\ 0, & \text{en otro caso.} \end{cases}$$
- Función escalón simétrica:
$$g(z) = \begin{cases} 1, & \text{si } z \geq 0, \\ -1, & \text{en otro caso.} \end{cases}$$

Los datos de entrada se procesan a través de la red neuronal con el objetivo de obtener una salida, para esto, una red neuronal debe aprender a calcular la salida correcta para el conjunto de entrada. A este proceso se le denomina proceso de entrenamiento o de aprendizaje, y el conjunto de datos sobre el cual este proceso se basa se denomina conjunto de datos de entrenamiento. El proceso de aprendizaje es donde la red neuronal modifica el valor de sus pesos, si el peso es distinto de cero entonces se crea una nueva conexión, por otro lado, si el peso es cero entonces la conexión se destruye. Además, si los pesos son suficientemente pequeños entonces los valores de entrada muy grandes de la neurona correspondiente tendrán poca influencia en la red.

El algoritmo básico para entrenar a una red neuronal artificial se le conoce como propagación. Este proceso consiste en que primero el conjunto de entrada X, llega a la capa de

entrada y estos valores se multiplican por los pesos previamente ingresados a la neurona. Se aplica la función de activación a la suma de estos productos como se observa en la ecuación (2.6).

$$I_k = g\left(\sum_{l=1}^{p} w_{l,k} \cdot X_l\right), \quad 1 \leq l \leq p,\ 1 \leq k \leq n_0, \tag{2.6}$$

donde, $w_{l,k}$ es el l-ésimo peso de la k-ésima neurona. Cada uno de los valores de I_k, para $1 \leq k \leq n_0$, es la k-ésima entrada de las neuronas de la primera capa oculta. Se aplica el mismo proceso de la capa de entrada, las entradas que llegan a la primera capa oculta se multiplican por sus respectivos pesos correspondientes la primera capa oculta. A la suma de este producto se le aplica la función de activación por lo que la salida de cada neurona de la primera capa oculta está dada por la ecuación (2.7).

$$a_{1,k_1} = g\left(\sum_{k=1}^{n_0} w_{k,k_1,1} \cdot I_k\right), \quad 1 \leq k_1 \leq n_1. \tag{2.7}$$

El valor de a_{1,k_1} es la salida de la k_1-ésima neurona en la primera capa oculta y $w_{k,k_1,1}$ es el k-ésimo peso de la k_1-ésima neurona en la primera capa oculta.

Este proceso se repite en cada una de las capas ocultas y la salida de la k_l-ésima neurona en la l-ésima capa oculta está dada por la ecuación (2.8).

$$a_{l,k_l} = g\left(\sum_{k_{l-1}=1}^{n_{l-1}} w_{k_{l-1},k_l,l} \cdot a_{l-1,k_{l-1}}\right), \quad 1 \leq k_l \leq n_l \text{ y } 1 < l \leq m. \tag{2.8}$$

Una vez que se obtiene este valor de la última capa oculta, estas serán la entrada de la capa de salida. Se define el valor de salida de la i-ésima neurona en la capa O_i, por la ecuación (2.9).

$$O_i = g\left(\sum_{k_m=1}^{n_m} w_{k_m,i},\ \cdot a_{m,k_m}\right), \quad 1 \leq i \leq q. \tag{2.9}$$

Estos pasos se resumen en el algoritmo 1 que es obtenido de Perez Aguila (2012). En este caso, los pesos se definen por el conjunto W.

Existen dos métodos de aprendizaje importantes:

Aprendizaje supervisado: Este aprendizaje determina la respuesta que debería generar la red a partir de una entrada determinada, en caso de que la salida de la red no coincida con la deseada, se procede a modificar los pesos de las conexiones hasta obtener una salida lo mejor aproximada a la deseada. Quien decide esta aproximación es un valor conocido como error.

Aprendizaje no supervisado: Este aprendizaje no recibe ninguna información que le indique si la salida generada por la red es o no correcta a una determinada entrada.

Algoritmo 1 Propagación(X, W, n_0, n_l, q)

Entrada: Entrada X con p elementos, conjunto de pesos W, número de neuronas en cada capa n_0, q y n_l, para $1 \leq l \leq m$.

Salida: Resultado de cada capa oculta a_l, $1 \leq l \leq m$, y de la capa de salida O.

1: **para** cada elemento en la capa de entrada, $1 \leq k \leq n_0$ **hacer**
2: Calcular I_k dada por la ecuación (2.6).
3: **fin para**
4: **para** cada neurona en la primera capa oculta $1 \leq k_1 \leq n_1$ **hacer**
5: Calcular la salida a_{1,k_1} dada por la ecuación (2.7).
6: **fin para**
7: **para** la segunda hasta la última capa oculta, $1 < l \leq m$ **hacer**
8: **para** cada neurona en la l-ésima capa oculta, $1 \leq k_l \leq n_l$ **hacer**
9: Calcular la salida a_{l,k_l} dada por la ecuación (2.8).
10: **fin para**
11: **fin para**
12: **para** la i-ésima neurona en la capa de salida, $1 \leq i \leq q$ **hacer**
13: Calcular el valor de O_i dada por la ecuación (2.9).
14: **fin para**

2.3.1. Retropropagación

El algoritmo de retropropagación (Back-Propagation) fue propuesto por Rumelhart Hinton y Williams en 1986. Este algoritmo utiliza el aprendizaje supervisado como estrategia de aprendizaje. Una vez que se aplica el algoritmo de propagación a la RNA se identifica el error entre la salida obtenida de la red con la esperada por la ecuación (2.10).

$$E = \sum_{i=1}^{q} (T_i - O_i)^2 . \tag{2.10}$$

Si el algoritmo de propagación obtiene una salida que coincide con la esperada entonces no se hace nada. Sin embargo, si existen diferencias entre la salida obtenida por la red y la esperada entonces, el algoritmo de retropropagación ajusta todos los pesos de cada capa de tal forma que el error se reduzca. Con el ajuste de los pesos se busca reducir de manera gradual el error pero no eliminarlo por completo en ese momento.

Dado que el error se calcula a nivel de la capa de salida, los pesos de esta capa son los primeros en ser modificados por la ecuación (2.11).

$$w_{k_m,i} = w_{k_m,i} + \alpha \cdot \Delta_i \cdot a_{m,k_m}, \tag{2.11}$$

donde $w_{k_m,i}$ indica el k_m-ésimo peso de la i-ésima neurona en la capa de salida, $\alpha \in \mathbb{R}^+$ denominada coeficiente de aprendizaje y

$$\Delta_i = \left[T_i - g\left(\sum_{k_m=1}^{n_m} w_{k_m,i} \cdot a_{m,k_m} \right) \right] \cdot g'\left(\sum_{k_m=1}^{n_m} w_{k_m,i} \cdot a_{m,k_m} \right), \quad \text{para } 1 \leq i \leq q.$$

Una vez que se actualizan los pesos de la capa de salida se prosigue a la actualización de los pesos de la última capa oculta. La ecuación (2.12) permite actualizar el k_{m-1}-ésimo peso de la k_m-ésima neurona de la última capa oculta, m.

$$w_{k_{m-1},k_m,m} = w_{k_{m-1},k_m,m} + \alpha \cdot \Delta_{k_m,m} \cdot a_{m-1,k_{m-1}}. \tag{2.12}$$

Donde,

$$\Delta_{k_m,m} = \sum_{i=1}^{q} \Delta_i \cdot w_{k_m,i} \cdot g' \left(\sum_{k_{m-1}=1}^{n_{m-1}} w_{k_{m-1},k_m,m} \cdot a_{m-1,k_{m-1}} \right) \cdot a_{m-1,k_{m-1}}.$$

Después de actualizar los pesos de la capa oculta m, se prosigue a actualizar los pesos de la penúltima capa oculta, así sucesivamente hasta llegar a la primera capa oculta. La ecuación (2.13) permite actualizar al k_{l-1}-ésimo peso de la k_l-ésima neurona en la l-ésima capa oculta, para $l = m-1, m-2, \ldots, 1$.

$$w_{k_{l-1},k_l,l} = w_{k_{l-1},k_l,l} + \alpha \cdot \Delta_{k_l,l} \cdot a_{l-1,k_{l-1}}, \tag{2.13}$$

con,

$$\Delta_{k_l,l} = \sum_{k_{l+1}=1}^{n_{l+1}} \Delta_{l+1} \cdot w_{k_l,l+1} \cdot g' \left(\sum_{k_{l-1}=1}^{n_{l-1}} w_{k_{l-1},k_l,l} \cdot a_{l-1,k_{l-1}} \right) \cdot a_{l-1,k_{l-1}}.$$

Una vez que se termina de actualizar los pesos de las capas ocultas, lo que resta es actualizar el l-ésimo peso de la k-ésima neurona en la capa de entrada utilizando la ecuación (2.14).

$$w_{l,k} = w_{l,k} + \alpha \cdot \Delta_k \cdot X_l. \tag{2.14}$$

Donde,

$$\Delta_k = \sum_{k_1=1}^{n_1} \Delta_{k_1,1} \cdot w_{k,k_1,1}\ g' \left(\sum_{l=1}^{p} w_{l,k} \cdot X_l \right).$$

El proceso de actualización de los pesos se repite hasta encontrar un error aceptable. Una vez que se ha terminado el proceso de ajuste de los pesos, la red recibe el conjunto de prueba que ayuda a predecir la salida con ayuda de dichos pesos.

La red recibe como entrada el conjunto de entrenamiento X, la salida esperada T, el número de capas ocultas y de neuronas en cada capa, el coeficiente de aprendizaje α y el error con el cual la salida de la red será aceptada. Los pasos se resumen en el Algoritmo 2.

Algoritmo 2 Retropropagación(X, W, n_0, n_l, q)

Entrada: Entrada X con $N \times p$ elementos, pesos $T_{N\times q}$, el número de capas ocultas m, número de neuronas en cada capa n_0, q y n_l, para $1 \leq l \leq m$.

Salida: Pesos W de cada capa oculta y de la salida.

1: Inicializar aleatoriamente los pesos, W_l para $1 \leq l \leq m$ y el peso de la capa de salida W_q, entre (-1,1).
2: **para** $t = 1$ hasta L iteraciones **hacer**
3: Inicializar el coeficiente de aprendizaje $\alpha = \frac{1}{1+t}$.
4: **para** $n = 1$ hasta N **hacer**
5: $[a_l, O]$=Propagación(X, W, n_0, n_l, q).
6: Calcular el error E dado por la ecuación 2.10).
7: Actualizar el peso de la capa de salida dada por la escuación 2.11
8: Actualizar los pesos de las capas ocultas:
9: **para** la m-ésima capa oculta hasta la primera capa oculta **hacer**
10: Actualizar los pesos de estas capas con las ecuaciones (2.12) y (2.13).
11: **fin para**
12: Actualizar los pesos de la capa de entrada utilizando la ecuación (2.14).
13: **fin para**
14: **fin para**

2.3.2. Mapas auto-organizados de Kohonen

Los mapas auto-organizados (SOM por sus siglas en inglés) es un modelo de red neuronal artificial que es entrenada por aprendizaje no supervisado y obtiene una representación discreta del espacio de entrenamiento. Uno de los modelos más populares de SOM fue propuesto por el profesor Teuvo Kohonen, por lo que es conocida como Red de Kohonen. En la Figura 2.5 se puede observar una representación de la arquitectura de la red de Kohonen.

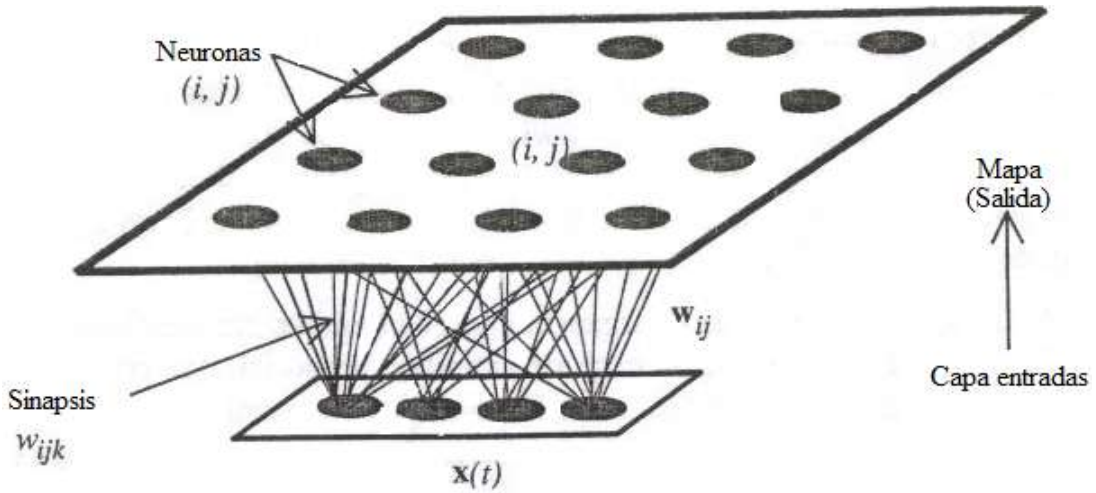


Figura 2.5: Arquitectura de la red de Kohonen.

Una de las ventajas de la red de Kohonen, y en general de las RNAs no supervisadas es el hecho de que no necesitan las salidas correctas a priori. Algunos problemas solo cuentan

con el conjunto de vectores de entrada y no con el conjunto de salidas correctas, por lo que no es posible medir el error de aprendizaje.

Dado un conjunto A de vectores de entrada, también conocido como espacio de la red, la red de Kohonen tiene como objetivo generar una partición del conjunto A en m regiones disjuntas, $a_1, a_2, \ldots, a_m$. Esto quiere decir:

1. $a_i \cap a_j \neq \emptyset$, para $i \neq j$.
2. $\bigcup_{i=1}^{m} a_i = A$.

La idea es que la red de Kohonen "cubra" al conjunto A, de tal forma que para cada vector de entrada se active una y solo una neurona. Es decir, si el conjunto A se divide en m regiones, entonces la red de Kohonen debe contar con al menos m neuronas y cada neurona se especializa en una y solo una región.

Con esto, la red de Kohonen se puede ver como una función cuyo dominio es el espacio de la red, A, y codominio las regiones en que se divide al conjunto A. La relación consiste en que cada elemento del espacio de la red es mapeado a la región que le corresponde. En otras palabras, la red de Kohonen realiza una clasificación de los vectores de entrada y estas clases definen una representación de la estructura del espacio de la red. La malla que resulta de la red de Kohonen sobre el conjunto A se denomina mapa del espacio.

Las redes de Kohonen son arreglos de neuronas con topologías de diferentes dimensiones, en este caso, se mencionará el problema del mapeo de un espacio n-dimensional utilizando una red Kohonen unidimensional.

La estructura de la red está formada por m neuronas y cada neurona recibe un vector de entrada X con n componentes:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Cada neurona j, para $j = 1, \ldots, m$, tendrá asociado un vector de pesos $W_j \in \mathbb{R}^n$ denominado centro de gravedad y servirá para obtener la salida correspondiente de la red. El objetivo de la red de Kohonen es encontrar, al presentar el vector de entrada a la red, a la neurona que proporcione la salida con valor mínimo. Dicha neurona está asociada a la clase que pertenece el vector de entrada y se le denomina neurona ganadora.

Para encontrar la región a la que pertenece el vector de entrada X, se calcula la distancia euclidiana entre X y cada uno de los vectores de pesos W_j, para $j = 1, \ldots, m$. La neurona ganadora es aquella cuya distancia sea la mínima. Si la distancia entre X y W_j es cercana a cero es porque entre estos vectores hay poca diferencia. Entre más alejada esté la distancia de cero significa que X y W_j son muy diferentes.

Para la actualización de los pesos, la red cuenta una vecindad de radio r asociada a la j-ésima neurona, $j = 1, \ldots, m$, y se define como el conjunto de neuronas localizadas

hasta r posiciones de la neurona j. Para el caso de las neuronas en los extremos de la red se utiliza una vecindad asimétrica. Además, en el proceso de entrenamiento se define la fuerza de enlace entre dos neuronas j y k, para $j, k = 1, \ldots, m$, como una función $\varphi(j, k, r)$ con codominio en [0,1] definida por la función (2.15).

$$\varphi(j,k,r) = \begin{cases} 1, & \text{si } j = k; \\ a, & a \in (0,1) \text{ si } j \text{ está en la vecindad de } k; \\ 0, & \text{en otro caso.} \end{cases} \tag{2.15}$$

La actualización del peso W_j, para $j = 1, \ldots, m$, está dado por la ecuación (2.16).

$$W_j = W_j + \alpha \cdot \varphi(j,k,r) \cdot (X - W_j). \tag{2.16}$$

Donde α es el coeficiente de aprendizaje y por lo general $\alpha \in (0,1)$. El coeficiente de aprendizaje, al igual que en el algoritmo de retropropagación, tiene como objetivo controlar la magnitud de la actualización de los pesos. Lo interesante del valor de α es que puede tomar un valor fijo en todo el entrenamiento o puede ir disminuyendo su valor conforme al número de iteraciones que se estén efectuando. El objetivo de esto, es suponer que entre cada iteración se va acercando a un estado óptimo, por lo que las actualizaciones requeridas deben ser cada vez más pequeñas, y así asegurarse de que no se está alejando del punto óptimo. La idea de ir reduciendo gradualmente a α también se puede aplicar al radio de vecindad, r. Con esto, al principio la neurona ganadora tiene fuerte influencia sobre su vecindad pero conforme pasan las iteraciones, se reduce gradualmente su influencia con el objetivo de que un peso asociado a una neurona vecina no sea desplazado de su posición óptima debido a la actualización de la neurona ganadora.

El objetivo de repetir la actualización de los pesos en diferentes iteraciones es hacer que los pesos se distribuyan uniformemente en el espacio de la red. Aunque, en la práctica esto depende del conjunto de entrenamiento y de los vectores de entrada.

Con el conjunto de entrenamiento, el SOM construye el mapa y con el conjunto de prueba se realiza la clasificación dada la entrada. El algoritmo recibe el conjunto de vectores de entradas, el coeficiente de aprendizaje y el número de neuronas en la red. El proceso se describe en el Algoritmo 3.

Puede ocurrir que una vez que termina el entrenamiento de la red, existan clases vacías, es decir, que ninguno de los vectores de entrada pertenece a dicha clase. En este caso, la neurona, correspondiente a esa clase, puede ser omitida de la red sin afectar la partición del espacio que se obtuvo durante el entrenamiento.

Si se desea convertir una red Kohonen unidimensional a bidimensional o multidimensional, se necesita que la la fuerza de enlace considere una vecindad bidimensional o multidimensional, respectivamente.

2.3.3. Red de contrapropagación

El red de contrapropagación (Counter-propagation neural networks, CPN) no es descubrimiento nuevo ya que es una combinación novedosa de tipos de redes previamente

Algoritmo 3 Kohonen(X, m, r, L)

Entrada: Conjunto de entrenamiento $X_{N\times n}$, número de neuronas en la red m, el valor del radio r, número de iteraciones L.

Salida: Pesos W.

1: Inicializar aleatoriamente los pesos, W de la red.
2: **para** $t = 1$ hasta L iteraciones **hacer**
3: Inicializar el coeficiente de aprendizaje $\alpha = \frac{1}{1+t}$.
4: **para** $n = 1$ hasta N **hacer**
5: $minD = \infty$ que indica la distancia mínima.
6: $minN = -1$ que se utiliza para identificar a la neurona ganadora.
7: **para** $j = 1$ hasta m **hacer**
8: Calcular la distancia euclidiana entre el k-ésimo vector de entrada y el vector de pesos,

$$d = \sqrt{\sum (E[k] - W[j])^2}.$$

9: **si** d<minD **entonces**
10: $minD = d$,
11: $minN = j$.
12: **fin si**
13: **fin para**
14: Neurona ganadora = $minN$.
15: Actualizar los pesos con la ecuación (2.16).
16: **fin para**
17: **fin para**

existentes. La red de contrapropagación fue propuesta por el doctor en matemáticas Robert Hecht-Nielsen en 1987. La estructura de esta red esta constituida por dos capas, una de ellas es la red de Kohonen y la otra es la capa de salida. En la Figura 2.6, se puede observar una representación de la red de contrapropagación.

La red de contrapropagación asocia un conjunto de entradas X con sus respectivas salidas Y, es decir, si existe una función f tal que $X = f(Y)$, entonces la red encontrará esa función. Además, la red aprenderá la función f^{-1} si existe la inversa de f.

El proceso del algoritmo de contrapropagación se realiza de manera similar al proceso de formación del algoritmo de Kohonen. La capa de Kohonen tiene el objetivo de asignar los datos de entrada multidimensionales en una matriz de dimensión inferior, por lo general en una bidimensional, ya que no es fácil el estudio y visualización humana de más de dos dimensiones de la red de Kohonen (Kuzmanovski and Novič, 2008). El aprendizaje de contrapropagación combina el aprendizaje supervisado y el no supervisado. El aprendizaje no supervisado se encuentra en la capa de Kohonen y el aprendizaje supervisado se presenta porque la red debe conocer la salida que se desea para cada información de entrada (Freeman y Skapura, 1991).

Dado que vectores de entrada similares son clasificados por una única neurona, es necesario que estas entradas lleven asociados vectores de salida parecidos para que la

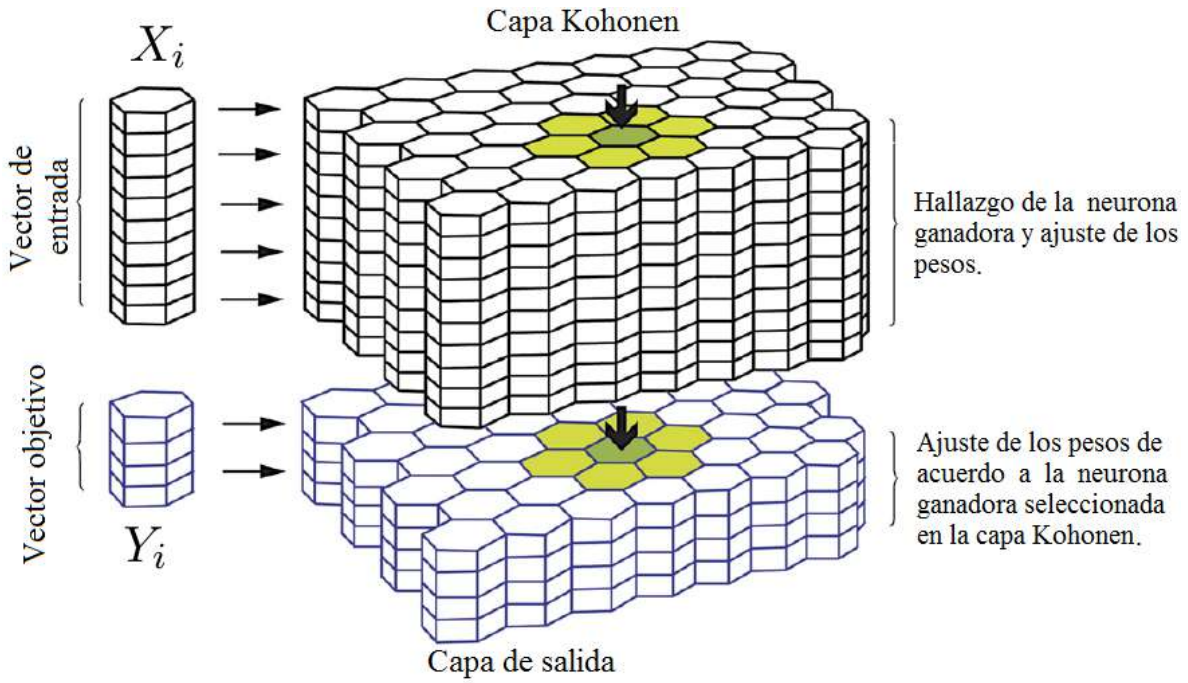


Figura 2.6: Representación gráfica de la red neuronal de contrapropagación. Imagen obtenida de Kuzmanovski and Novič (2008).

red de contrapropagación funcione correctamente. En caso contrario, se debe aumentar el número de neuronas en la capa de Kohonen, logrando que cada neurona represente un número menor de entradas.

La técnica de entrenamiento de la red de contrapropagación es similar al de la red de Kohonen. Cada vector de entrada X_i, para $i = 1, \ldots, N$:

$$X_i = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,n} \end{bmatrix},$$

es comparado con los pesos, W_j, de las neuronas, para $j = 1, \ldots, m$, para obtener la neurona ganadora. Una vez que se obtiene la neurona ganadora c en la capa de Kohonen, los pesos de ambas capas son ajustados a partir de los vectores de entrada X_i, vectores objetivo Y_i, la taza de aprendizaje α y la función de vecindad $\phi(d_j - d_g)$.

$$W_j = W_j + \alpha(t) \cdot \phi(d_j - d_c) \cdot (X_i - W_j). \tag{2.17}$$

$$U_j = U_j + \alpha(t) \cdot \phi(d_j - d_c) \cdot (Y_i - U_j). \tag{2.18}$$

Los vectores W_j y U_j en las ecuaciones (2.17) y (2.18) son los pesos de la capa de Kohonen y de la capa de salida, respectivamente. La diferencia $dj - dc$ es la distancia topológica entre la neurona ganadora c y la neurona j. La tasa de aprendizaje $\alpha(t)$ sigue

teniendo el mismo papel, controlar la intensidad de la actualización de los pesos durante el entrenamiento. En este caso, α es una función decreciente, y decrece conforme al número de iteraciones t. Al actualizar el peso W_j de la neurona j también se actualizan los pesos de las neuronas vecinas, y el número de neuronas vecinas depende del radio de vecindad r.

El entrenamiento de la red de contrapropagación se divide en dos fases:

1. Entrenamiento de la capa de Kohonen:

   a) Inicializar de manera aleatoria los pesos de la red.

   b) Ingresar el conjunto de entrada X. En esta capa se encontrará a la neurona ganadora c empleando la distancia Euclidiana, definida en el Algoritmo 3.

   c) Modificar los pesos utilizando la ecuación (2.17).

   d) Se repiten los últimos dos pasos hasta estabilizar los pesos.

   Una vez que concluya esta fase del entrenamiento, los pesos de cada neurona de esta capa serán, aproximadamente, la media de todas las entradas que pertenecen a la clase correspondiente a dicha neurona.

2. Entrenamiento de la capa de salida:

   En esta fase se pueden dar dos casos. El primero consiste en que cada clase consta de sólo una salida. En este caso, el vector de la neurona será la salida deseada para cada clase. En el segundo caso, cada vector de entrada de una clase tiene asociada una salida deseada diferente, por lo que la neurona proporciona diferentes salidas deseadas. Para resolver este problema se utiliza un algoritmo de aprendizaje que permita corregir el error, y así hacer que los pesos se aproximen a las salidas deseadas:

   a) Ingresar a la red un vector de entrada X_i con su respectiva salida deseada Y_i.

   b) Obtener la neurona ganadora c de la capa de Kohonen.

   c) Actualizar los pesos de la capa de salida por la ecuación (2.18).

   d) Repetir los pasos anteriores hasta obtener una buena aproximación de las salidas deseadas para cada vector de entrada X_i.

En la Figura 2.7 se resumen los pasos principales del algoritmo de contrapropagación.

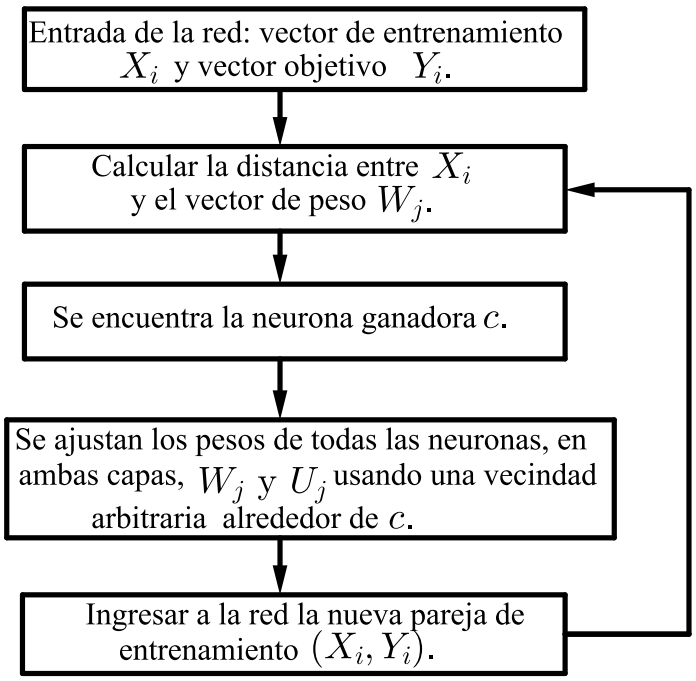


Figura 2.7: Representación del algoritmo de contrapropagación.

Capítulo 3

Desarrollo

Este capítulo se divide en siete secciones y describe la implementación de los métodos utilizados. En la primera sección se muestra un diagrama que representa de forma general el desarrollo de la tesis paso por paso. La segunda sección describe la base de datos que se utiliza en el proyecto de tesis. La tercera sección se describe el análisis que se le aplico a la base de datos: elección del rango de años para el cual se hace la reconstrucción de precipitación, detección de datos atípicos, así como la distancia de cada estación climatológica al Golfo y al Océano Pacífico. En la sección 3.6 se describe el análisis de componentes principales realizado con las variables climatológicas con las que se cuentan y las distancias calculadas en la sección 3.5.

Las últimas tres secciones se describe la implementación de los algoritmos de redes neuronales: SOM, retropropagación y contrapropagación, respectivamente.

3.1. Caso de estudio y base de datos

Dentro de la República Mexicana, el estado de Oaxaca está ubicado en la región suroeste del país, limitando al norte con los estados de Puebla y Veracruz, al este con el estado de Chiapas, al sur con el Océano Pacífico y al oeste con el estado de Guerrero. Geográficamente, se encuentra entre las coordenadas de $16^\circ\,53'\,53''$ de latitud norte y $96^\circ\,24'\,51''$ de longitud oeste (ver Figura 3.1). El estado está dividido en 570 municipios, con 93 757 km^2 de territorio y su capital es la Ciudad de Oaxaca de Juárez. Alberga una rica composición multicultural donde conviven más de 16 grupos étnicos y es el estado con más diversidad de México.

De acuerdo con información recopilada por el Instituto Nacional de Estadística y Geografía (INEGI) existen cinco tipos de climas diferentes en la entidad: cálido subhúmedo, seco y semiseco, cálido húmedo, templado subhúmedo y templado húmedo (ver Figura 3.6). En las regiones altas de la sierra, tiene un clima templado con inviernos fríos. En las regiones de Valles Centrales y Mixteca Alta, se tiene un clima templado subhúmedo y seco extremoso. En La Cañada y la Costa, el clima es cálido húmedo; mientras que en el Istmo el clima es cálido subhúmedo con vientos potentes.

Figura 3.1: Localización del estado de Oaxaca y regiones. Imagen obtenida de `http://iohio.org.mx/esp/mapas.htm`.

Los datos climáticos utilizados en esta tesis, se extrajeron del CLImate COMputing project (CLICOM) a cargo del Servicio Meteorológico Nacional (SMN). Esta base de datos tiene información nacional de todos los estados en México. Del estado de Oaxaca existe información sobre 351 estaciones climatológicas en distintas regiones del estado, almacenados en hojas de cálculo Excel. La información almacenada corresponde a los registros diarios de diferentes variables climatológicas, a las cuales se asignaron los siguientes códigos:

1. Temperatura observada (1),
2. Temperatura máxima (2),
3. Temperatura mínima (3),
4. Lluvia (5),
5. Evaporación (18),
6. Tormenta eléctrica (30),
7. Granizo (31),
8. Niebla (32),
9. Nublados (43).

Los registros de datos para cada estación, no inician ni terminan en la misma fecha. En general, el inicio de registro va desde 1922 hasta 2004 y el fin de registro va desde 1969 hasta 2009. Además, hay datos de valor dudoso, inapreciable o estimado. Para indicar esto, el SMN utiliza las etiquetas descritas en la Tabla 3.1.

Tabla 3.1: Descripción de las banderas que se presentan en la base de datos.

Bandera	**Significado**
M	Missing, valor no existente; siempre irá acompañado de un valor -99999.
D	Dudoso, no confirmado en registro en papel.
J	Dudoso, confirmado en registro en papel.
T	Lluvia inapreciable, siempre irá acompañado de un valor 0.
E	Valor estimado.
C	Temperatura corregida ó intercambio entre temperatura mínima y máxima no revisado en expediente.
H	Temperatura verificada en expediente.
O	Intercambio entre temperatura ambiente y mínima no revisado en expediente.
P	Intercambio entre temperatura ambiente y máxima no revisado en expediente.
Q	Intercambio entre las tres temperaturas no revisado en expediente.
R	Intercambio entre temperatura ambiente y mínima revisado en expediente.
S	Intercambio entre temperatura ambiente y máxima revisado en expediente.
V	Intercambio entre las tres temperaturas revisado en expediente.
A	Precipitación revisada y corregida en expediente.
B	Precipitación comparada con estaciones vecinas no revisado en expediente.
G	Valor generado.
I	Temperaturas corregidas mediante otros métodos diferentes a la revisión en expediente (diarios). Resultado estimado a partir de un período incompleto (solo para mensuales).
*	Valor que se repite más de una vez (solo para mensuales).

Tabla 3.2: Archivo con información de todas las estaciones climatológicas en el país, clasificada por estado.

Estación	**Nombre**	**Municipio**	**Estado**	**Estado**	**Long**	**Lat**	**Xutm**	**Yutm**	**Huso**	**Altitud**	**Inicio**	**Fin**	**Años**	**datos**	**Rev.**
1003	CALVILLO	CALVILLO	1	Aguascal	-102.7	21.8	115623.7	2424567.5	14	1640	1932-01	1988-12	57	88.9	ok
1004	CAÑADA	AGUASCALI	1	Aguasca	-102.1	22.0	169787.8	2437143.7	14	1910	1970-03	2010-10	40.7	99.8	ok
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
20162	TEQUISISTL	MAGDALENA	20	Oaxaca	-95.5	16.4	863305.5	1818168.1	14	182	1940-11	2008-12	68.2	86.1	ok
20163	TEZOATLAN	TEZOATLAN	20	Oaxaca	-97.8	17.6	626140.3	1952075.6	14	1527	1963-09	1996-02	32.5	91.9	ok
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30113	NANCHITAL	NANCHITAL	30	Veracruz	-94.4	18.1	985442.1	2003591.9	14	20	1939-03	1983-12	44.8	67.7	ok
30114	NAOLINCO	NAOLINCO	30	Veracruz	-96.8	19.6	723014.3	2174358.7	14	1542	1955-07	2008-12	53.5	98.5	ok

La información de todas las estaciones climatológicas del país se describe como se indica en la Tabla 3.2. Además se cuenta con un archivo para cada estado, con información diaria de cada estación y de las diferentes variables climatológicas clasificada por mes. Estos archivos contienen la información descrita en la Tabla 3.2.

Se consultaron las coordenadas de las estaciones climatológicas para calcular las distancias de las estaciones a la frontera del país. La hoja de cálculo que tiene información del estado de Oaxaca ("Oax_DLY"), junto con la información de sus estados vecinos: Chiapas, Guerrero, Puebla y Veracruz, fue la base de los análisis realizados. El número de estaciones climatológicas por estados utilizados en las pruebas se muestran en la Figura 3.2. Se contó con 1437 estaciones climatológicas. El mayor número corresponde al estado de Oaxaca, con 355 estaciones, seguido por Veracruz con 347.

Tabla 3.3: Archivo que contiene información de todas las estaciones climatológicas en el país, clasificada por estado. Cuando el DATASET-ID es igual a 20 corresponde al estado de Oaxaca.

DATASET-ID[a]	**Station-ID** [b]	**ELEMENT-CODE** [c]	**YEAR-MONTH**[d]	**VALUE-1** [e]	**FLAG1-1** [f]	**VALUE-2**	**FLAG1-2**	...	**VALUE-31**	**FLAG1-31**
20	20001	1	1950-04	17		17		...	-99999	M
20	20001	2	1950-04	22		24		...	-99999	M
20	20001	3	1950-04	12		15		...	-99999	M
20	20001	5	1950-04	0		0		...	-99999	M
20	20001	30	1950-04	0		0		...	-99999	M
20	20001	31	1950-04	0		0		...	-99999	M
20	20001	32	1950-04	1		1		...	-99999	M
20	20001	43	1950-04	0		0		...	-99999	M

[a] DATASET-ID: valor asignado al "Estado".
[b] Station-ID: valor asignado a la estación climatológica.
[c] ELEMENT-CODE: Código asignado a la variable climatológica.
[d] YEAR-MONTH: información de la variable climatológica en un cierto mes de un cierto año.
[e] VALUE-i: Valor de la variable climatológica en el día i, para $i = 1, 2, \ldots, 31$.
[f] FLAG-i: Bandera de la variable climatológica en el día i, para $i = 1, 2, \ldots, 31$.

3.2. Diseño experimental

El objetivo de este trabajo es reconstruir las series de precipitación mensual del estado de Oaxaca, se eligió trabajar con información mensual debido a que en las series de precipitación diaria se pueden ver varios picos, caso contrario si se toman las series mensuales como se puede observar en la Figura 3.3.

Para lograr dicho objetivo se realizaron tareas que permiten garantizar el funcionamiento correcto de los algoritmos utilizados para la reconstrucción de series. En la Figura 3.4 se observa el esquema desarrollado para lograr el objetivo de la tesis.

Como primer paso se obtuvo la base de datos con información de las estaciones climatológicas del estado de Oaxaca y de sus estados vecinos (Guerrero, Puebla, Veracruz y Chiapas), así como información diaria de las variables climatológicas de cada estación

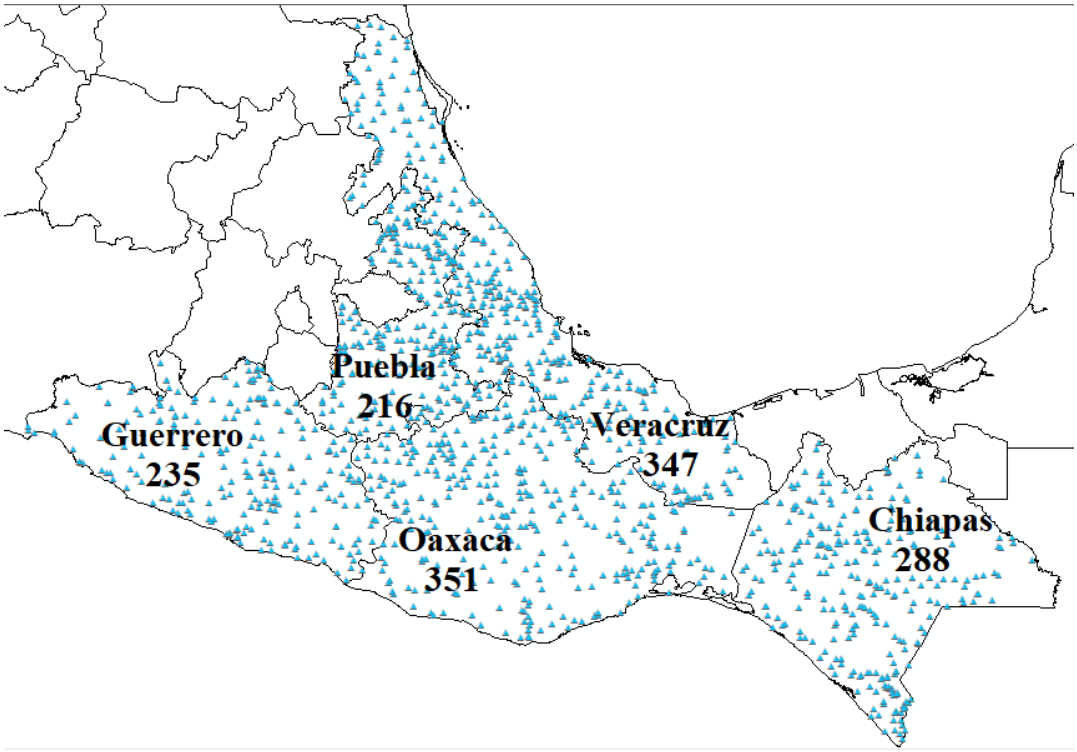


Figura 3.2: Ubicación de las 1437 estaciones climatológicas utilizadas en la base de datos. Oaxaca con 355 estaciones, Guerrero con 235, Puebla con 216, Veracruz con 347, y Chiapas con 288 estaciones climatológicas.

(esta información se detalla en la sección 3.1), con lo cual hasta el momento se contaba con una base de datos inicial.

El siguiente paso fue delimitar el rango de años de reconstrucción de las series de precipitación, para esto se analizó el porcentaje de datos por año de todas las estaciones climatológicas del estado de Oaxaca. Además, fue importante realizar un análisis de calidad de los datos obtenidos debido a que los datos atípicos inválidos, influyen negativamente en la aplicación de las técnicas utilizadas y en la obtención de buenos resultados. Como tercera tarea en esta sección se hizo el cálculo de las distancias de las estaciones climatológicas al Golfo y Pacífico de México. Esto con el fin de considerar en los análisis posteriores, la influencia de la cercanía (o lejanía) de los océanos sobre el comportamiento de la precipitación. Como resultado, se obtuvo un filtro de la base de datos, que fue ingresada al PCA.

Posteriormente se analizaron las variables climatológicas, mediante PCA, con el cual, se encontraron las variables correlacionadas con la precipitación, con el objetivo de obtener una mejor reconstrucción de la series de precipitación. En este caso, solo se utilizo información de las estaciones climatológicas del estado de Oaxaca. Los resultados obtenidos se emplearon para generar la base de datos que fue ingresada al SOM.

Como resultado se obtuvieron grupos de estaciones climatológicas, donde en cada grupo corresponde a una región hidrológicamente homogénea. Los datos de salida obtenidos, fueron ingresados a los algoritmos de reconstrucción.

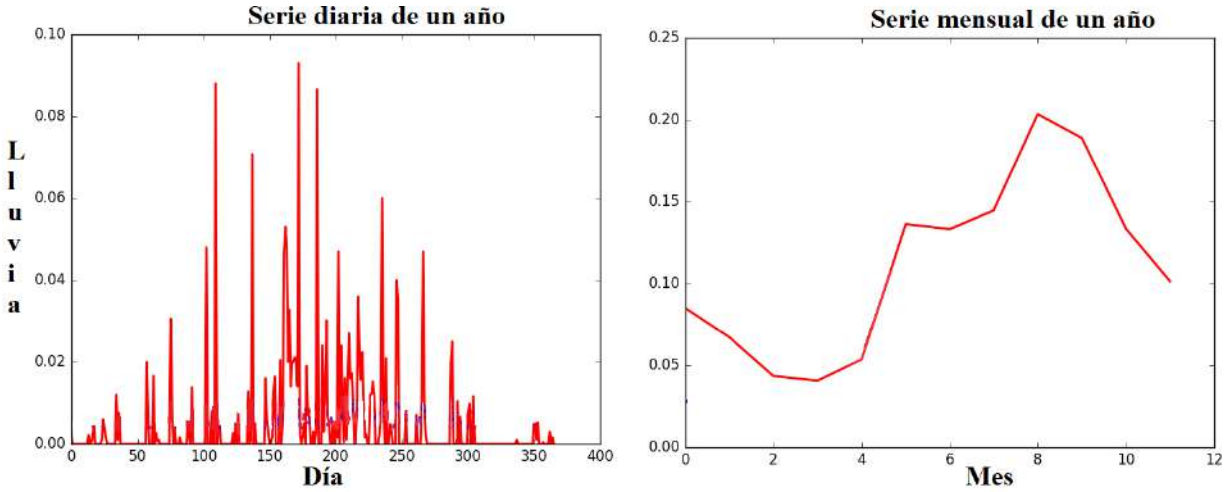


Figura 3.3: La gráfica de la izquierda representa la serie diaria de la precipitación que se presenta durante un año y, en la gráfica de la derecha es la información mensual que se presenta durante un año.

Por último, se implementaron los algoritmos de reconstrucción: retropropagación y contrapropagación.

3.3. Elección del rango de años

Como se mencionó, los datos capturados no inician ni terminan en la misma fecha para cada estación climatológica, por lo cual se debe de elegir el rango de años en el que será factible realizar la reconstrucción de datos de precipitación.

Para calcular el porcentaje de datos de todas las estaciones en el año x, p_x, se utilizó la ecuación (3.1).

$$p_x = \frac{\sum_i IE_{i,x} * 100}{TIA_x}, \text{ para } x = 1922, \ldots, 2009, \tag{3.1}$$

donde $IE_{i,x}$ indica la cantidad de información con la que se cuenta en el año x de la estación i, para $x = 1922, \ldots, 2009$ e $i = 1, \ldots, 351$. TIA_x es la cantidad total de información en el año x por todas la estaciones climatológicas y está dado por la ecuación (3.2).

$$TIA_x = nd * ne, \tag{3.2}$$

donde, nd es el número de días en el año x, para $x = 1922, \ldots, 2009$, el cual varia si el año es o no bisiesto (365 o 366), y ne es el total de estaciones climatológicas, es decir, ne=351.

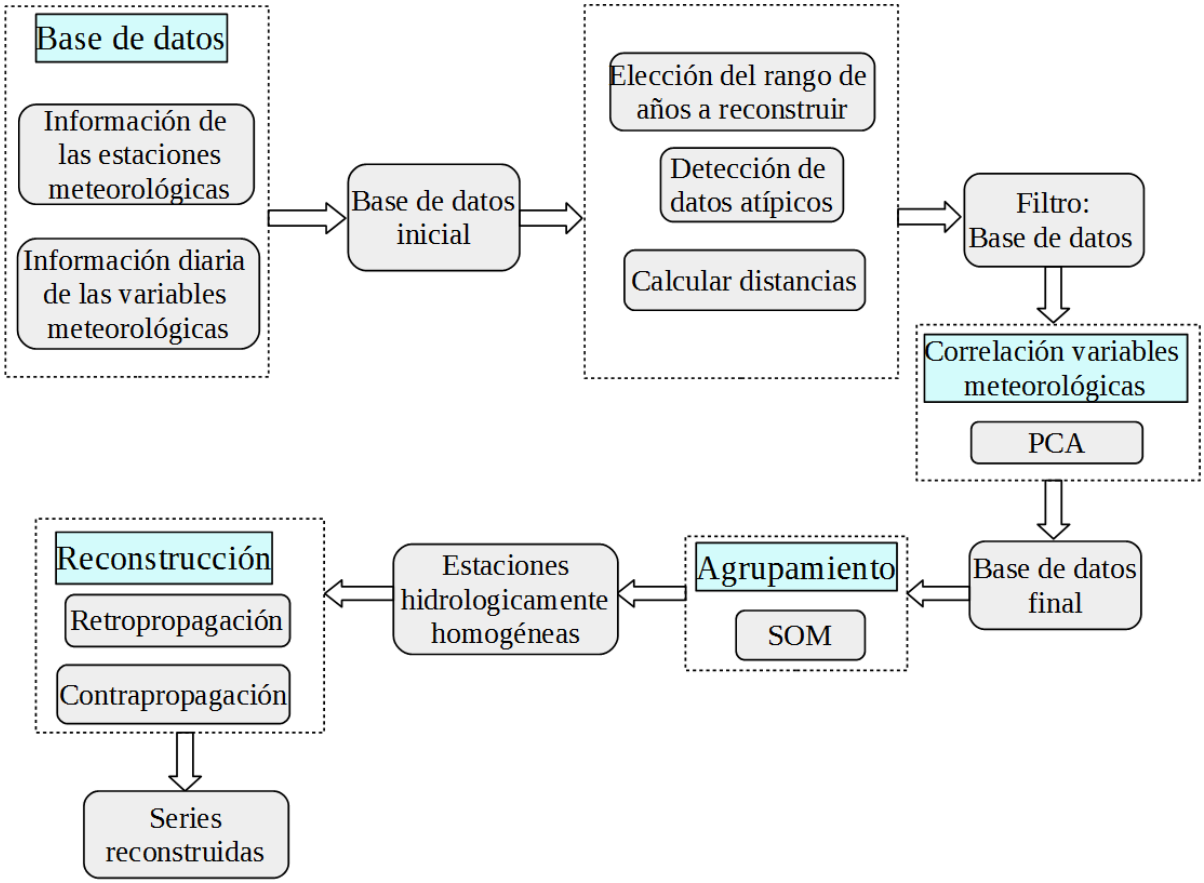


Figura 3.4: Esquema general del desarrollo para la reconstrucción de series de precipitación del estado de Oaxaca.

Para obtener los datos de precipitación mensual se obtuvo la suma de la precipitación que se presentó en un determinado mes mediante la ecuación (3.3).

$$PreM_{m,x,i} = \sum_{k=0}^{nd} PreD_{i,x,m,k}, \tag{3.3}$$

donde $PreD_{i,x,m,d}$ es la precipitación que se presenta el día d en el mes m, en el año x de la estación i, para $d = 1, 2, \ldots, nd$, $m = 1, 2, \ldots, 12$, $x = 1957, \ldots, 2007$, $i = 1, 2, \ldots, 1437$ y nd el número de días en dicho mes.

3.4. Detección de datos atípicos

En las bases datos por lo general se presentan datos alejados de la media por lo que se realizó la detección de datos atípicos. Es este caso, se utilizó la prueba de Grubbs que se describe en la sección 2.1. Para aplicar esta técnica se llevaron a cabo los siguientes pasos:

1. Se calculó el logaritmo natural de los datos diferentes de cero.

2. Se eligió el porcentaje de confianza a utilizar, en nuestro caso utilizamos el 1 %.
3. Se obtuvo el valor de K_N, el cual depende del total de datos N que se tengan y se puede obtener consultando la Tabla 3.4.
4. Se estimó el valor de X_L y X_H dados por las ecuaciones (2.1) y (2.2), respectivamente.

Tabla 3.4: Valor de K_N para la prueba de Grubbs.

N	**1 %**	**N**	**1 %**	**N**	**1 %**	**N**	**1 %**
3	1.155	16	3.052	28	3.464	40	3.673
4	1.499	17	3.103	29	3.486	41	3.687
5	1.78	18	3.149	30	3.507	42	3.7
6	2.011	19	3.191	31	3.528	43	3.712
7	2.201	20	3.23	32	3.546	44	3.724
8	2.358	21	3.266	33	3.565	45	3.736
9	2.492	22	3.3	34	3.582	46	3.747
10	2.606	23	3.332	35	3.599	47	3.757
11	2.705	24	3.362	36	3.616	48	3.768
12	2.791	25	3.389	37	3.631	49	3.779
13	2.867	26	3.415	38	3.646	50	3.789
14	2.935	27	3.44	39	3.66	51	3.798
15	2.997						

Dado que hubo meses en los que no se registro ningún dato o que en todo el mes no se presentó precipitación, entonces la cantidad de precipitación de esos meses es de cero. Al calcular los logaritmos naturales de dichos meses ocasiona problemas, por lo que es importante solo calcular el logaritmo natural de los valores de precipitación mensual distintos de cero.

El en análisis de datos atípicos para cada estación climatológica se calculó el valor de X_L y X_H para el mes de enero de todos los años registrados de esa estación, por lo que se definieron, respectivamente, las variables descritas en las ecuaciones (3.4) y (3.5) como los límites inferior y superior de la estación l del mes m del año x, donde $l = 1, 2, \ldots, 1437$, $m = 1, \ldots, 12$ e $N = 1, 2, \ldots, 51$. Esto se debe a que se cuentan con 1437 estaciones climatológicas, son 12 meses en un año (Enero, Febrero, ..., Noviembre, Diciembre) y el rango de años para realizar la reconstrucción es de 1957 a 2007, con un total de 51 años, por lo que el valor de K_N a lo más es de 3.798 cuando $N = 51$ (ver Tabla 3.4).

$$X_L^{l,m,N} = \exp\left(\overline{X} - K_N S\right), \tag{3.4}$$

$$X_H^{l,m,N} = \exp\left(\overline{X} + K_N S\right). \tag{3.5}$$

Con esto, para la estación l, se tienen 12 límites inferiores y 12 límites superiores:

$$\left[\left[X_L^{l,1,N}, X_H^{l,1,N}\right], \left[X_L^{l,2,N}, X_H^{l,2,N}\right], \ldots, \left[X_L^{l,12,N}, X_H^{l,12,N}\right]\right],$$

donde K_N, depende del total de años en los que la precipitación mensual sea distinta de cero. En particular, $X_L^{l,1,N}$ es el límite inferior de la estación l, del mes de enero con un total de N años con lluvia mensual distinta de cero.

Una vez calculados los límites inferiores y superiores se procede a encontrar los datos que están fuera de su respectivo intervalo. El resultado que arrojó la prueba de Grubbs son datos extremos que bien pueden ser datos atípicos válidos o inválidos. La forma correcta de comprobar si estos datos corresponden a eventos que realmente ocurrieron, o son el resultado de errores de medición o captura, sería necesario revisar cada observación para verificar la veracidad del dato acudiendo a la información histórica, es decir, verificar si ese día hubo algún desastre natural para presentar valores tan altos de precipitación o si fue época de sequía para que por días no se presentará lluvia. Sin embargo, acudir a esta información llevaría más tiempo por lo que se decidió descartar a todos los datos extremos que arrojo la prueba de Grubbs. Esto quiere decir, los meses que se encuentran fuera de su rango se consideraron como información faltante en la base de datos.

3.5. Cálculo de las distancias

Uno de los primeros pasos en el proyecto de tesis es realizar un análisis de las variables climatológicas mediante PCA para encontrar correlaciones entre las variables climáticas y la precipitación. Pero también se cree que la distancia de las estaciones climatológicas al mar influye en las lluvias, por tal motivo se calculó la distancia de cada estación al Océano Pacífico y al Golfo de México.

La determinación de las distancias se realizó con ArcGIS que es un software utilizado en los sistemas de información geográfica. Los pasos que se siguieron para calcular las distancias fueron:

1. Se crearon distintos puntos a lo largo de la línea costera de México y se obtuvieron sus coordenadas latitud y longitud (ver Figura 3.5). Con esto, se generaron dos archivos diferentes: uno que contiene información de los puntos que se seleccionaron en la frontera con el Océano Pacífico y otro del lado de la frontera del Golfo de México.

2. Se calcularon las distancias de las estaciones climatológicas del estado de Oaxaca a cada uno de los puntos creados, para esto se utilizó la distancia del círculo grande (great circle) dada por la ecuación (3.6). El cálculo de la distancia se realizó con Python 2.7, como salida se obtiene un archivo .txt que contiene ambas distancias para cada una de las estaciones climatológicas.

$$d(x,y) = 2 \cdot R \cdot \arctan\left(\frac{\sqrt{\cos^2(\phi_2)\cdot\sin^2(\Delta\lambda) + \cos(\phi_1)\cdot\cos(\phi_2)\cdot\sin^2\left(\frac{\Delta\lambda}{2}\right)}}{\sin(\phi_1)\cdot\sin(\phi_2) + \cos(\phi_1)\cdot\cos(\phi_2)\cdot\cos(\Delta\lambda)}\right), \tag{3.6}$$

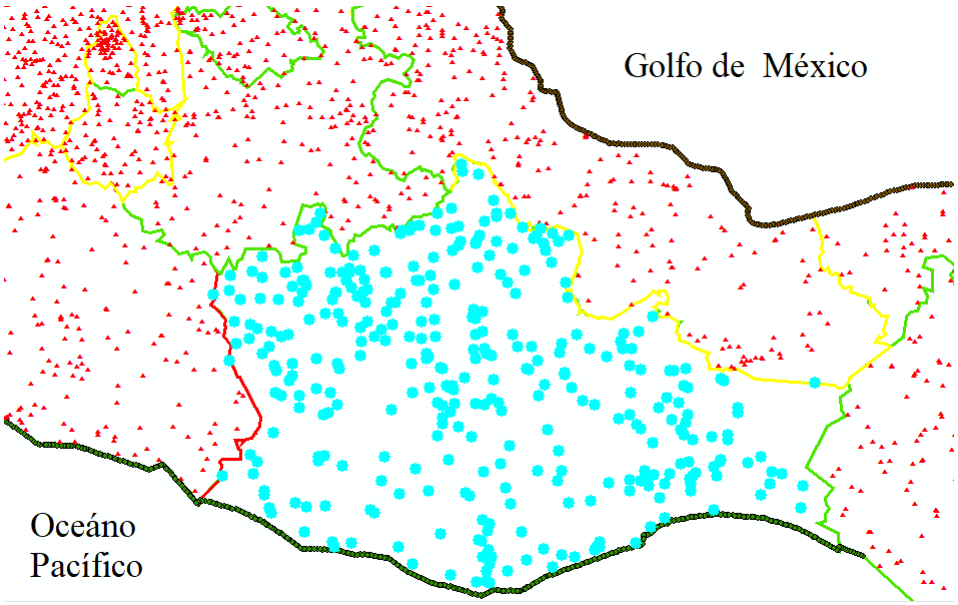


Figura 3.5: Mapa de la República Mexicana realizado con ArcGIS donde se observa: las estaciones climatológicas del estado de Oaxaca (azul), estaciones climatológicas del resto del país (rojo), puntos con el Golfo (café) y con el Pacífico (verde) de México.

donde, $x = (\phi_1, \lambda_1)$ y $y = (\phi_2, \lambda_2)$ son dos coordenadas en radianes, R es el radio de la tierra y,

$$\Delta\lambda = |\lambda_2 - \lambda_1|.$$

3. Se seleccionó la distancia mínima entre la línea costera y cada estación climatológica, es decir, para cada estación se obtuvieron dos distancias mínimas una al Océano Pacífico y otro al Golfo de México.

3.6. Análisis de componentes principales

La variable climatológica de interés de este trabajo es la precipitación, sin embargo se cree que sobre está variable puede estar asociada a otras variables climatológicas. Para conocer la influencia de las variables climatológicas sobre la precipitación, se realizó un PCA.

El estado de Oaxaca tiene una gran variedad en su clima, el 47 % de la superficie del estado presenta clima cálido subhúmedo que se localiza en toda la Zona Costera y hacia el este, el 22 % presenta clima cálido húmedo localizado principalmente en la región Norte, el 16 % presenta clima templado húmedo en las partes altas orientales de los cerros Volcán Prieto y Humo Grande, el 11 % presenta clima seco y semiseco en la región centro Sur y Noroeste, el restante 4 % presenta clima templado subhúmedo hacia el Sur y Noroeste del estado en zonas con altitudes entre 2 000 y 3 000 metros. En la Figura 3.6, se puede observar la variedad en el clima del estado de Oaxaca y se puede apreciar que la parte seca se encuentra en el centro del estado, mientras que en las fronteras con Veracruz y

el Pacífico es más húmedo, por lo que se puede creer que la distancia a los mares puede afectar a las precipitaciones.

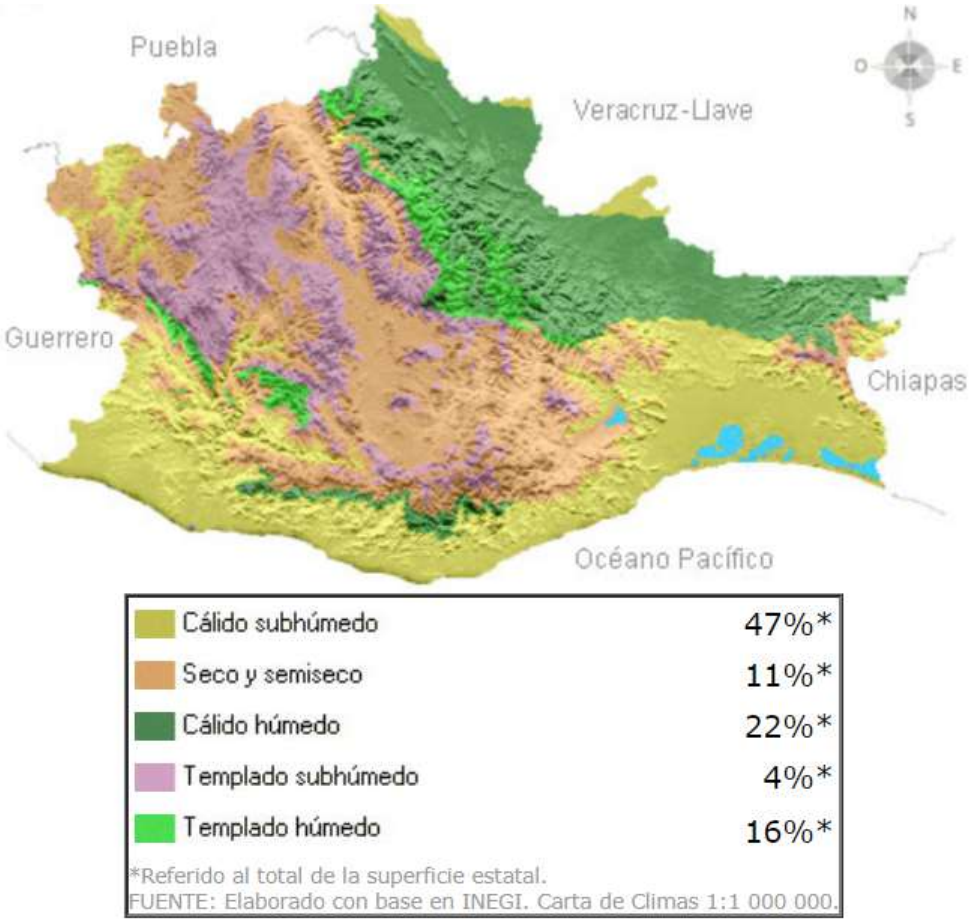


Figura 3.6: Mapa del estado de Oaxaca, elaborado por el INEGI (2014), donde se describen los tipos de clima por zonas.

Al analizar la base de datos se observa que se tienen variables cualitativas y cuantitativas. Las variables cuantitativas son aquellas que son expresadas por un valor numérico las cuales son: temperatura observada, temperatura máxima, temperatura mínima, lluvia diaria, evaporación, distancia al Golfo de México y al Océano Pacífico. Las variables: tormenta eléctrica, granizo, niebla y nublados; se consideran cualitativas porque por ejemplo, si durante el día no hubo granizo entonces a esta variable en ese día se le asigna el valor de 0, si hubo un poco de granizo se le da el valor de 1, pero si cae mucho granizo en ese día entonces se le asigna el valor de 2, y así para el resto de las variables cualitativas. Sin embargo, calificar si es poco, más o menos, o mucho granizo depende de la percepción de la persona que estuvo ese día en la estación climatológica.

Debido a lo anterior se decide descartar a las variables cualitativas y solo realizar el PCA con las variables cuantitativas. Para realizar el análisis se normalizaron cada una de las variables, esto es, por cada variable climatológica se obtuvo el valor máximo y cada valor se dividió por ese valor máximo. La base de datos ingresada al análisis se puede observar en la Tabla 3.5, donde cada columna es el valor de cada una de las variables

climatológicas que se presentan en el estado de Oaxaca y cada fila es información diaria de cada estación climatológica.

Tabla 3.5: Base de datos ingresado al PCA aplicado a las variables climatológicas cuantitativas del estado de Oaxaca.

Temp. Obs ($°C$)	**Temp. Max** ($°C$)	**Temp. Min** ($°C$)	**Lluvia** mm	**Evaporación** mm	**Distancia Golfo** Km	**Distancia Pacífico** Km	**Estación**
0.341	0.305	0.235	0.0	0.050	0.495	0.409	20001
0.390	0.481	0.294	0.0	0.089	0.495	0.409	20001
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.658	0.592	0.735	0.021	0.516	0.322	0.813	20509

3.7. Algoritmo de Kohonen

El modelo de Kohonen en redes neuronales nos ayudó a encontrar las estaciones climatológicas que son hidrológicamente homogéneas, es decir, identificar grupos de estaciones que pertenecen a una misma región hidrológicamente homogénea.

Los pasos que se siguen para realizar el agrupamiento de las estaciones climatológicas fueron los siguientes:

1. Se hizo un filtro de la información, en este filtro solo se quedaron las estaciones climatológicas del estado de Oaxaca y sus estados vecinos (Guerrero, Puebla, Veracruz y Chiapas) con información de precipitación de 1957 a 2007.

2. Se generó el conjunto de entrenamiento E. Para esto, cada vector del conjunto, $e \in E$ tiene longitud de 12×1, cada vector contiene información mensual por año de la precipitación, son 12 elementos por variable debido a que estos son los meses durante un año.

$$E = [e_1, e_2, \ldots, e_j, \ldots, e_n],$$

donde,

$$e_j = [mes_1, \ldots, mes_{12}],$$

e_j, para $1 \leq j \leq n$, es una serie de una determinada estación de un determinado año.

En el conjunto de entrenamiento se tomaron todas aquellas series mensuales con información faltante menor a 5 días, esto para no afectar demasiado el análisis de dichos datos. Esto quiere decir que, si la estación x tiene información de 1957 a 1984, cada serie mensual se ordenó de forma ascendente tomando como criterio el número de días que no se registro la precipitación durante todo el año. Una vez ordenados se eligieron todos aquellas series anuales que tienen menos de 5 días no registrados de precipitación.

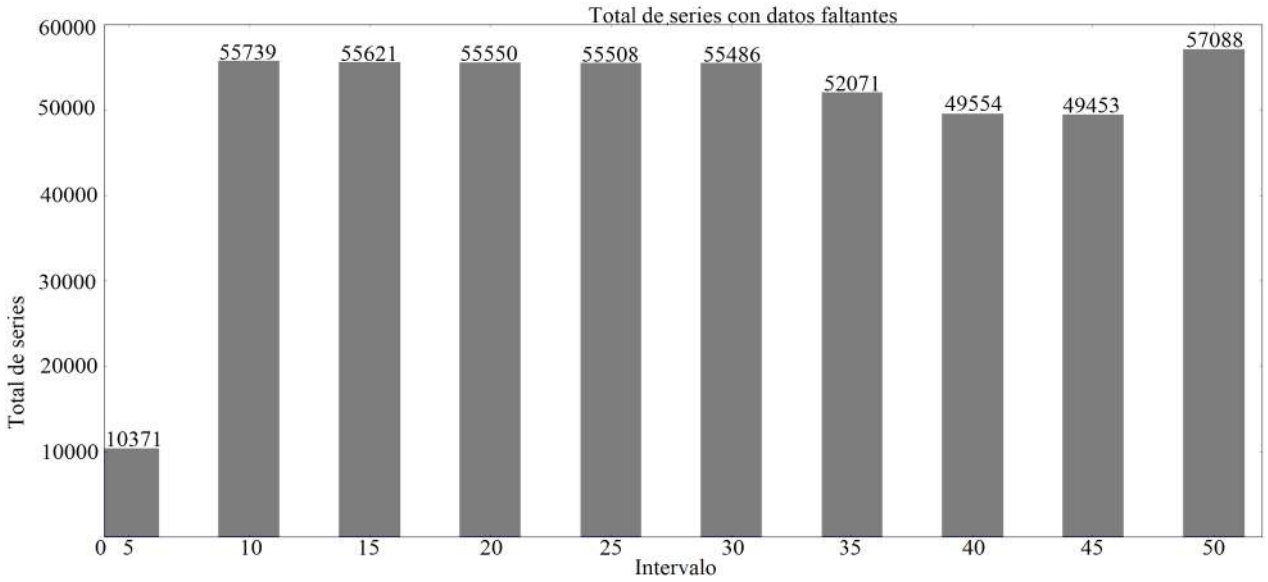


Figura 3.7: Cantidad de series anuales de precipitación de la base de datos que tienen cierta cantidad de información faltante. En el último intervalo representa a todas las series que tienen más de 50 datos faltantes diarios en la serie, es decir, un total de 57088 series.

3. El conjunto E es la entrada de la red de Kohonen. La salida de la red depende de la cantidad de grupos que se desean.

Los parámetros utilizados en la red de Kohonen se describen a continuación:

- Debido a que el INEGI divide al estado de Oaxaca en cinco grupos de clima diferentes o de forma más detallada en 15 grupos (INEGI, 2012; INEGI 2018), entonces se decidió hacer un agrupamiento de 5, 10 y 15 grupos. Esto quiere decir que, el número de neuronas en la capa de salida m, de la red de Kohonen está formada por 5, 10 o 15 neuronas dependiendo de la prueba.
- Se hicieron diferentes pruebas con el valor del radio r, para poder encontrar el valor que obtenga un buen agrupamiento, ya sea que en todas las iteraciones se tenga un valor fijo o que cambie conforme al número de iteración t, dado por la ecuación (3.7).

$$r = \frac{1}{1+t}. \tag{3.7}$$

- Se utilizó la distancia euclidiana, $d(P,Q)$, definida por la ecuación (3.8).

$$d(U,V) = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2}, \tag{3.8}$$

donde $U = (u_1, u_2, \ldots, u_n)$ y $V = (v_1, v_2, \ldots, v_n)$.

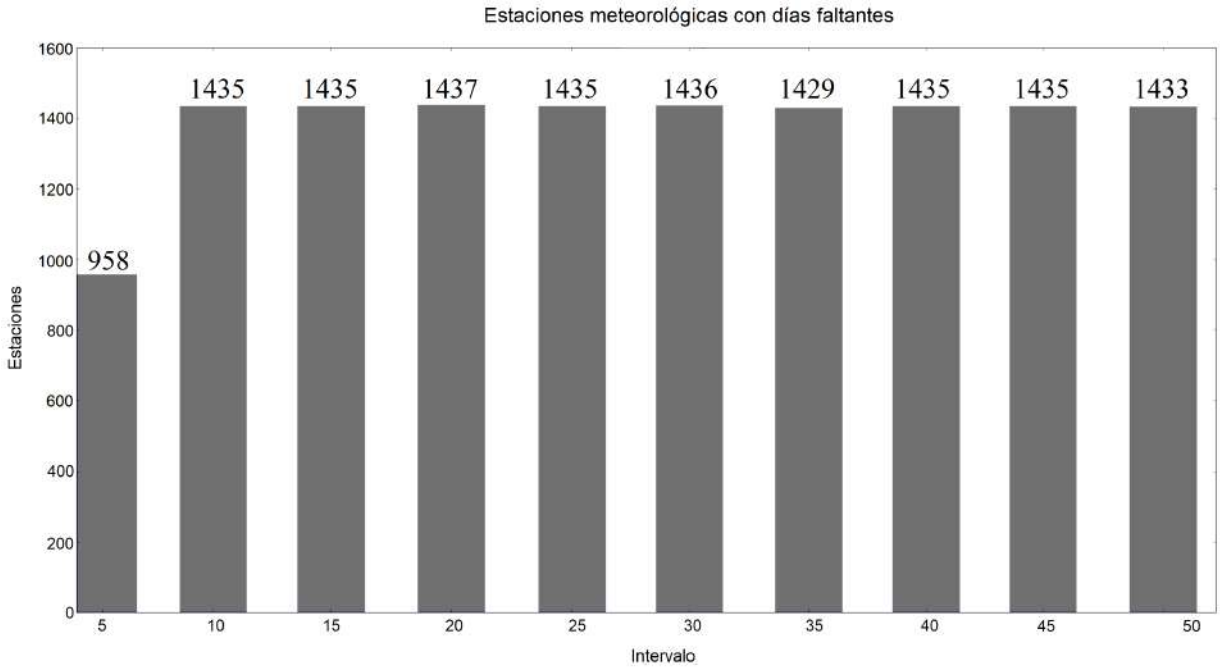


Figura 3.8: Cantidad de estaciones climatológicas que tienen cierta cantidad de información faltante. Por ejemplo, se tienen 958 estaciones meteorológicas con series con 5 o menos datos faltantes

- Cada elemento del conjunto de pesos W son inicializados de dos formas, aleatoria con una distribución $U(0,1)$ y de tal forma que todos tenga el mismo valor. El conjunto W está formado por m vectores donde cada vector está asociado a una neurona de la capa de salida, y está dado por:

$$W = \{w_1, w_2, w_3, \ldots, w_m\},$$

  donde $w_i = [w_{i,1}, w_{i,2}, \ldots, w_{i,12}]$ y $w_{i,j} \in \mathbb{R}$, con $i = 1, 2, \ldots, m$ y $j = 1, 2, \ldots, 12$.

- El coeficiente de aprendizaje α al igual que el radio, se probaron con valores fijos en todas las iteraciones o de manera que cambie en cada iteración utilizando la ecuación (3.7).

- El número de iteraciones L utilizados en las pruebas es de 1000, 3000 y 5000.

Como salida en la red de Kohonen se obtiene el conjunto de pesos W donde cada w_i, para $i = 1, 2, \ldots, m$, se interpreta como la serie anual representante de cada grupo i. Las pruebas de reconstrucción fueron aplicadas a cada uno de los agrupamientos obtenidos por la red de Kohonen, y la información utilizada para la red de retropropagación es el conjunto formado por los representantes de cada grupo, es decir, W.

3.8. Retropropagación

El algoritmo de Retropropagación ayudó a la reconstrucción de los datos de precipitación del estado de Oaxaca utilizando información de las estaciones climatológicas vecinas. Las siguientes pruebas que se muestran en esta sección se aplicaron a datos de precipitación mensual.

Para entrenar a la red de retropropagación se necesitaron dos conjuntos, el de entrada X y el que se espera obtener T, estos dos conjuntos se obtuvieron de la siguiente forma:

- Los representantes de cada grupo de la red de Kohonen conformaron al conjunto T, que la red aproxima y con el que se evalúa el error de aprendizaje MSE.
- El conjunto X es de tamaño $NV \times N$ donde NV es el número de vecinos y $N = 12*n$ con n el número de grupos formados en la red de Kohonen tal que cada grupo tenga tres o más elementos, esto se debe a que el número de vecinos $NV \in \{3, 5\}$. En forma vectorial, X está dado por:

$$X^T = [\ldots, [d_{i,j}^1, d_{i,j}^2, \ldots, d_{i,j}^k, \ldots, d_{i,j}^{NV}] \ldots],$$

  $d_{i,j}^k$ representa mes i del año j del vecino k, con $1 \leq i \leq 12$, $1 \leq j \leq 51$, $1 \leq k \leq NV$.

En la Figura 3.9 se puede observar la red utilizada para la reconstrucción de precipitación para un determinado mes y que es entrenada por retropropagación, con las características descritas en los puntos anteriores con $NV = 3$. En este caso, ingresa el elemento x_j, del vector de entrenamiento X, y se obtiene la salida o_j, para $1 \leq j \leq N$. El vector x_j está formado por tres elementos que son los valores de la precipitación de las tres estaciones climatológicas vecinas de la estación objetivo. Para el caso NV=5, x_j está compuesto por cinco elementos y la capa de entrada lo constituyen cinco neuronas $(I_1, I_2, \ldots, I_5)$. La salida o_j representa la precipitación de un determinado mes, que fue aproximada por la red de retropropagación.

En las primeras pruebas realizadas se utilizaron los siguientes parámetros:

- Se utilizaron tres y cinco vecinos por lo que el número de neuronas en la capa de entrada es $n_1 \in \{3, 5\}$, y los vecinos fueron elegidos de tal manera que sean los más similares al elemento objetivo. Para calcular la similaridad se calculó la distancia euclidiana entre el vector objetivo con todas las series del grupo y se escogieron las 3 o 5 series que hayan tenido la menor distancia, respectivamente.
- La capa de salida solo consta de una neurona, $n_3 = 1$ ya que se espera que la red devuelva la cantidad de lluvia para un determinado día.
- Solo se utilizó una capa oculta con diferente número de neuronas n_2.
- El coeficiente de aprendizaje α fue de 0.001, 0.01, 0.05, 0.1, 0.25 y $\dfrac{1}{1+t}$.
- Los pesos W para cada capa se generaron de manera aleatoria con una distribución uniforme (-1,1) o inicializados por un valor en particular.

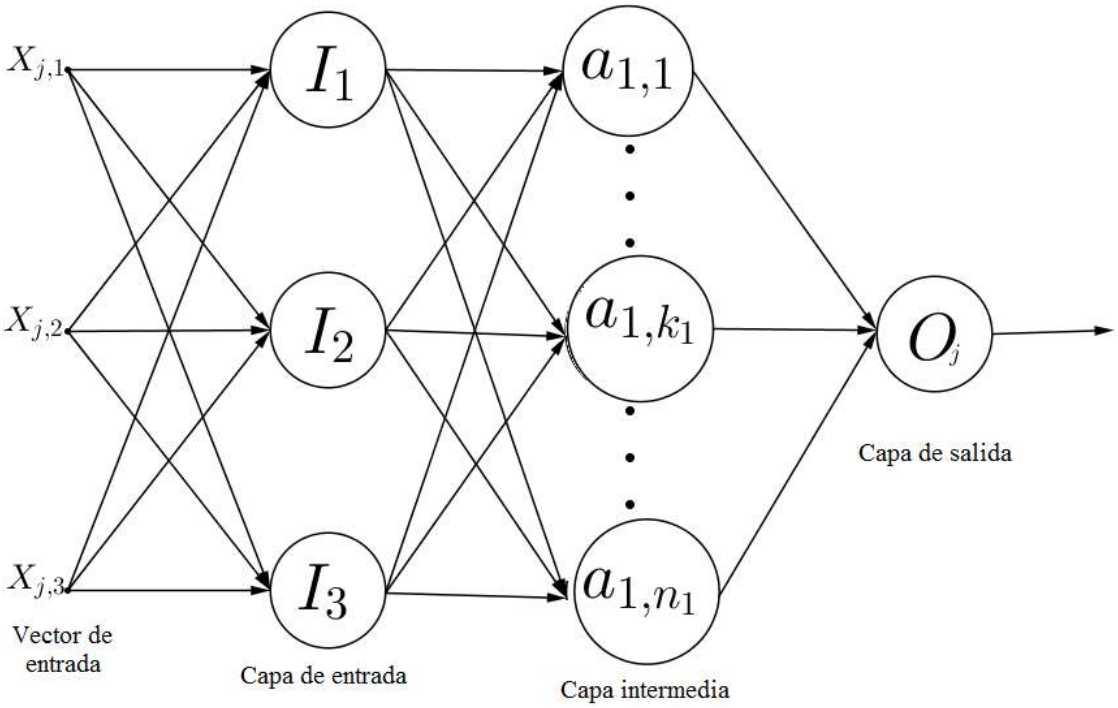


Figura 3.9: Red utilizada para la reconstrucción de precipitación entrenada por retropropagación con NV=3.

- El número de iteraciones fue L: 500000.
- La función de activación en cada capa es la sigmoidal, tangente hiperbólica modificada o Gaussiana dadas por las funciones $f, g, h : \mathbb{R} \to (0, 1)$, respectivamente:

$$f(x) = \frac{1}{1 + \exp(-x)}, \tag{3.9}$$

$$g(x) = \frac{1}{1 + \exp(-2x)}, \tag{3.10}$$

$$h(x) = \exp(-x^2). \tag{3.11}$$

  Se eligen estas funciones debido a que en la capa de salida se desea obtener un valor en el intervalo (0,1) porque la cantidad de lluvia está normaliza en el intervalo [0,1] (ver Figura 3.10).

Una vez que se realizaron las pruebas utilizando el conjunto de entrenamiento, se eligió la prueba que arrojó el menor MSE para cada agrupamiento y se guardo el conjunto final W que consta de los pesos de la capa de entrada, oculta y de salida. Estos pesos fueron utilizados para obtener la reconstrucción de las series anuales de precipitación del estado de Oaxaca. Se construyeron dos conjuntos P y T_p, donde T_p contiene 17901 series anuales de precipitación del estado de Oaxaca que resultaron de los 51 años totales del intervalo [1957-2007] por las 351 estaciones climatológicas. El conjunto P de prueba es de tamaño

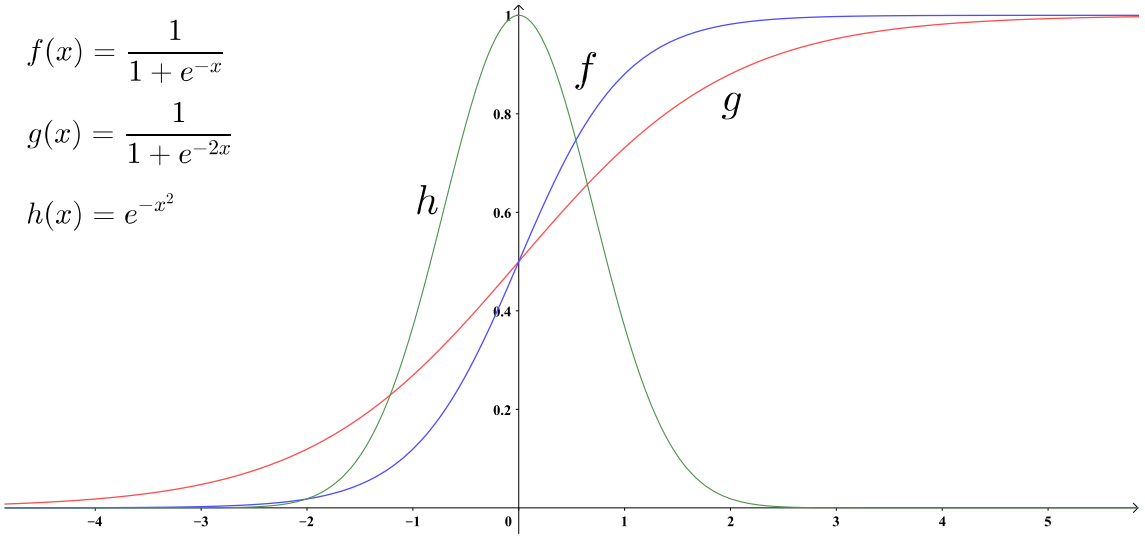


Figura 3.10: Gráficas de las funciones de activación utilizadas en la red de retropropagación.

$NV \times N$ donde NV es el número de vecinos y $N = 12 * n$ con n =17901 que son el número de series que constituyen a T_p:

$$P^T = [\ldots, [d^1_{i,j}, d^2_{i,j}, \ldots, d^k_{i,j}, \ldots, d^{NV}_{i,j}] \ldots].$$

El conjunto P^T fue ingresado a la red mediante el algoritmo de propagación con los pesos obtenidos con la red de retropropagación y se compararon con el conjunto T_p que corresponde a los datos medidos de la precipitación del estado de Oaxaca.

Una vez que se reconstruyeron las series, se calculó el coeficiente de correlación de Pearson y la diferencia absoluta entre la serie original con la reconstruida, con el fin de determinar que tan eficiente fue el modelo propuesto. El coeficiente de correlación de Pearson de dos series X e Y denotado por ρ_{XY}, está dado por la ecuación (3.12).

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \, \sigma_Y}, \tag{3.12}$$

donde σ_{XY} es la covarianza de (X, Y), σ_X y σ_Y son la desviación estándar de la serie X e Y, respectivamente. Por otro lado, la diferencia absoluta de dos series X e Y denotado por DA_{XY} se define por la ecuación (3.13).

$$DA_{XY} = \sum_{i=1}^{n} \mid X_i - Y_i \mid, \tag{3.13}$$

donde n es el total de elementos de las series. Es importante mencionar que la comparación solo se realiza con las series completas del estado de Oaxaca, es decir, aquellas que no tienen ninguna información mensual perdida, dando un total de 1518 series completas.

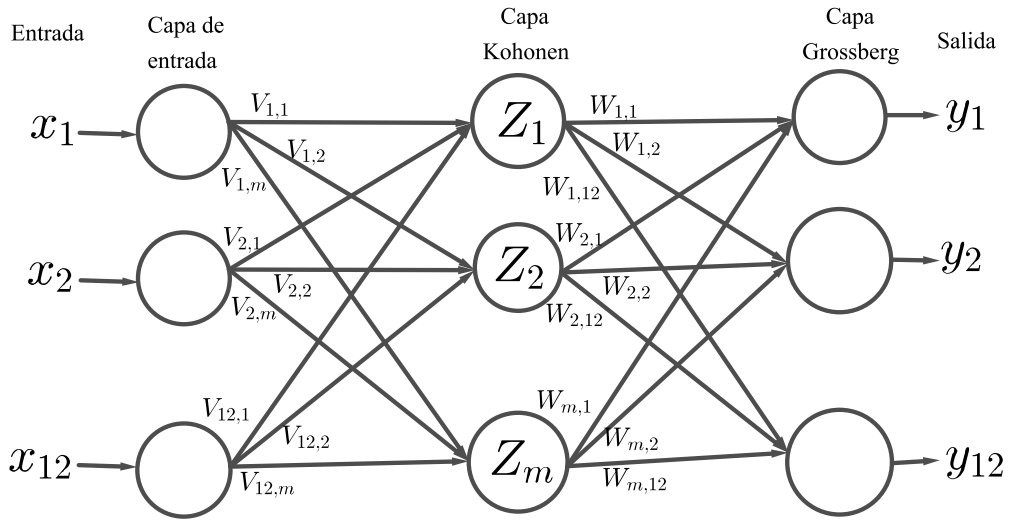


Figura 3.11: Estructura de la red de contrapropagación utilizada para la reconstrucción de precipitación

3.9. Contrapropagación

El algoritmo de contrapropagación no solo ayuda a realizar el agrupamiento de las estaciones climatológicas, también permite la reconstrucción de las series deseadas. La red para este modelo se puede visualizar en la Figura 3.11, los parámetros utilizados se enuncian a continuación:

- El conjunto de entrenamiento X ingresado a la red de contrapropagación es el mismo conjunto E que es ingresado a la red de Kohonen (ver 1 y 2).
- En la capa de entrada se utilizaron 12 neuronas debido a que cada elemento del conjunto X es de tamaño 12 y son ingresados directamente a la capa de entrada.
- El número de neuronas m en la capa de Kohonen es variable, es decir, se puede especificar en el programa el número de neuronas que se desea.
- En la capa de salida se tienen 12 neuronas ya que cada una representa el valor de la precipitación aproximada para el año ingresado a la red.
- Para encontrar la neurona ganadora se utiliza la distancia euclidiana.
- El valor del coeficiente de aprendizaje en la capa de Kohonen y la capa de Grossberg pueden ser el mismo valor en todo el proceso o, cambiar conforme cambia el número de iteraciones utilizando la ecuación (3.7).
- El número de iteraciones en la capa de Kohonen y la capa de Grossberg que son elegidos por el usuario.

Una vez que se entrenó la red de contrapropagación se procedió a la reconstrucción de las series de precipitación obtenidas de la base de datos medidos, la cual se conoce como fase de prueba de la red. Para esta fase, a la red de contrapropagación se ingresa el conjunto de prueba P que constituye de las series de precipitación del estado de Oaxaca,

y se construye de la misma forma que el conjunto T_p de la red de retropropagación (ver sección 3.8).

Cada serie $x \in T_p$ fue ingresada a la red de contrapropagación realizando los siguientes dos pasos:

1. La primera capa (capa de Kohonen) se asigno a un grupo i determinado, es decir, se activa la neurona Z_i de dicha capa, para $i = 1, 2, \ldots, m$.
2. Después, se pasó a la capa de Grossberg dando como salida la red reconstruida que es aquella formada por el peso W_m correspondiente a la neurona Z_m, para $i = 1, 2, \ldots, m$.

Para verificar la eficiencia de la red de contrapropagación se procedió igual que la red de retropropagación. Se calculó el coeficiente de correlación de Pearson ρ_{XY} y la diferencia absoluta DA_{XY} definidos por la ecuación (3.12) y (3.13), respectivamente. La eficiencia solo es calculada con aquellas series de precipitación del estado de Oaxaca que están completas, dando un total de 1518 series.

Capítulo 4

Resultados

Para llegar al objetivo planteado de este trabajo, se realizaron diferentes tareas como se puede observar en el diagrama de la Figura 3.4. Primero se realizó un análisis de la base de datos para después encontrar correlaciones entre las variables climatológicas. También se hizo el agrupamiento de las estaciones climatológicas y por último se aplicaron los algoritmos de reconstrucción. En el presente capítulo se presentan los resultados obtenidos en cada una de estas etapas del desarrollo.

4.1. Elección del rango de años

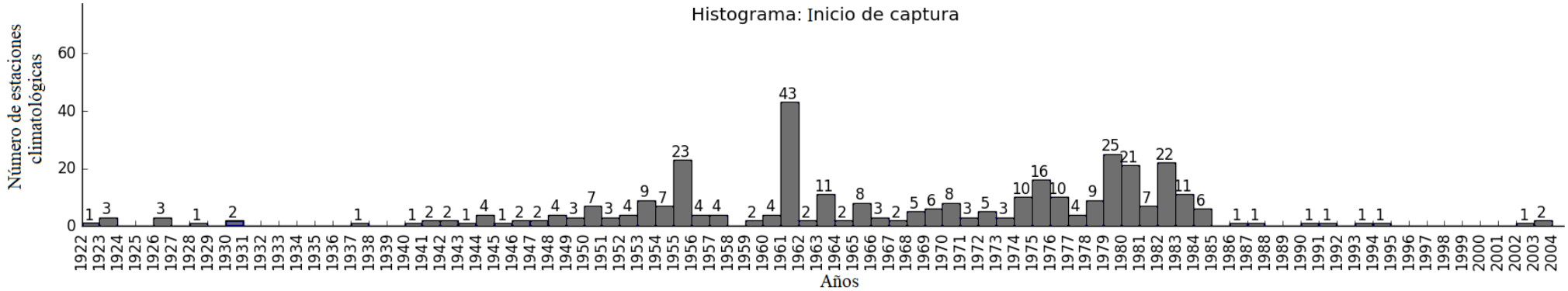
La Figura 4.1 (a), muestra el inicio de registros de datos en las estaciones. Se puede observar que la mayor concentración de estaciones que iniciaron registros, se encuentra en el centro de la gráfica, desde 1960 hasta 1985. El año promedio en que iniciaron los registros de las estaciones climatológicas es 1967.Los años en que finalizan los registros en las estaciones se muestran en la Figura 4.1 (b). Se puede observar que 98 estaciones terminan su registro en el 2009 y antes de 1980 muy pocas estaciones terminan su registro en ese periodo. Además, el promedio de años de fin de registro es de 1996.

Se debe tener en cuenta que la Organización Meteorológica Mundial (WMO, por sus siglas en inglés, World Meteorological Organization) recomienda contar contar con periodos de registros de al menos 50 años continuos, para detectar de manera confiable, tendencias y otros cambios en datos hidrológicos. Si solo tomamos en cuenta el promedio de inicio y fin de recolección de datos de las estaciones solo se tendrían 30 años. Debido a esto, se dirigieron esfuerzos para lograr tener el mayor número de series posibles con un periodo común de 50 años.

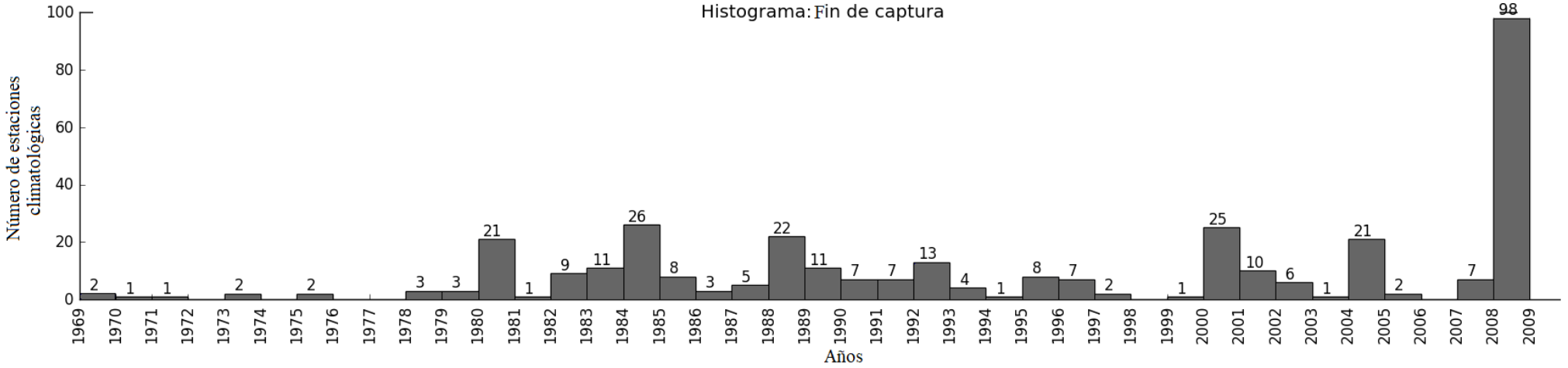
En la gráfica de la Figura 4.2 muestra el porcentaje de la cantidad de datos por año de todas las estaciones climatológicas. Se observar que, antes de 1950 se cuenta con una proporción muy pequeña (p_x<10 %) por lo que se consideró que el rango de datos antes de 1950 no sería viable para realizar la reconstrucción. Por otro lado, a partir de 1956 se cuenta con información mayor al 20 % por año. Además, se puede observar que en 2009

la información con la que se cuenta es mínima, p_x =1.1 %, por lo que, se decidió descartar este año de registros para su análisis.

Con base en lo anterior, se determinó que el rango donde hay mayor porcentaje es de 1957 al 2007, donde en cada año hay más del 24 % de información, por lo tanto, el rango de años en el que se realizará la reconstrucción de datos de precipitación es de 1957 a 2007, con un total de 51 años. Dado que la reconstrucción se realizó para precipitación mensual, en total se deben tener 879444 meses (12 meses por año de los 51 años del rango elegido por las 1437 estaciones climatológicas en total), pero hay 470075 meses en los que al menos un día no se registro la precipitación y 409369 meses con registro de precipitación de todos sus días.



(a)



(b)

Figura 4.1: La gráfica (a) representa el número de estaciones climatológicas que inician la captura de datos en un determinado año, mientras que (b) representa el fin de captura.

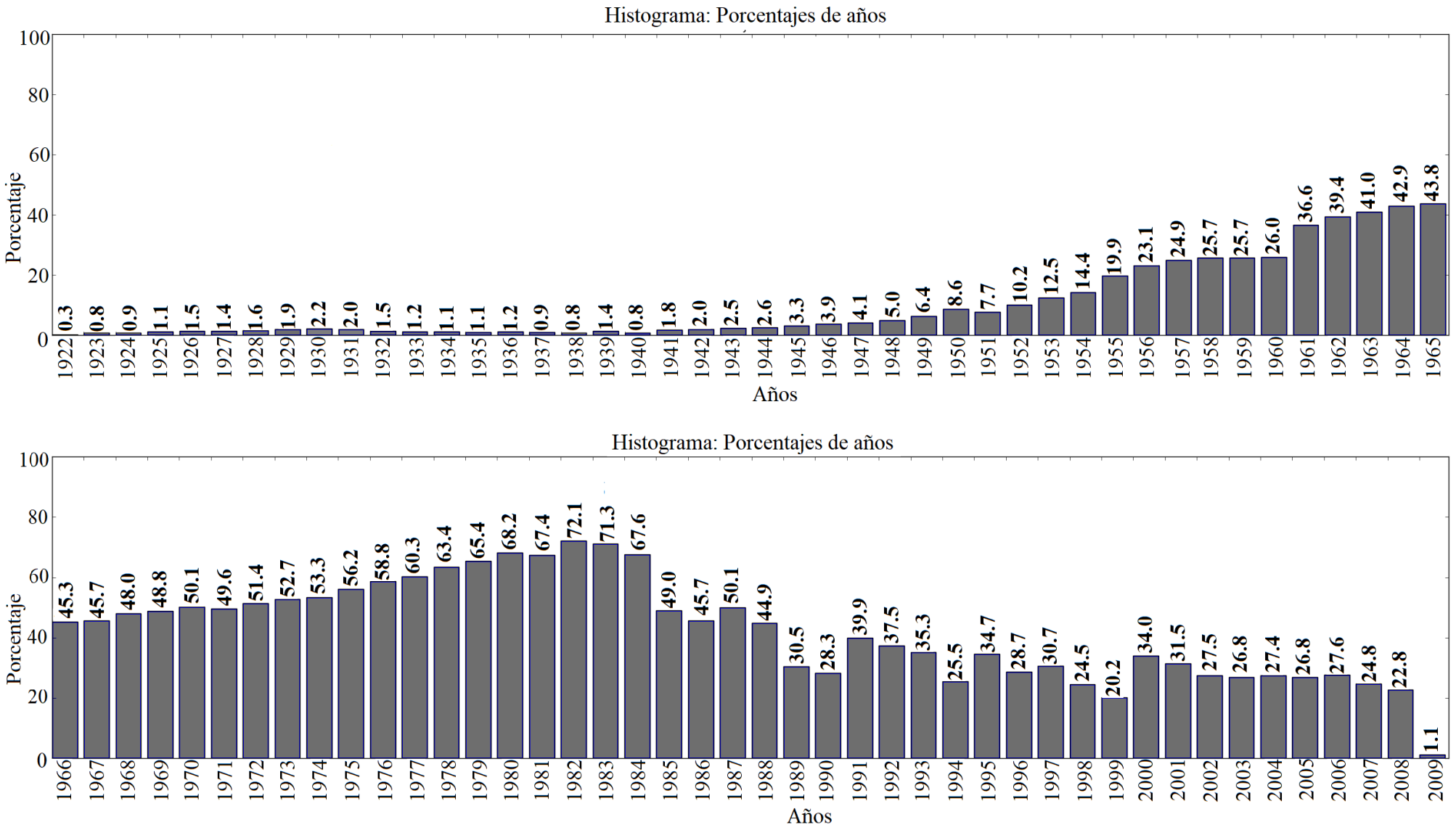


Figura 4.2: Porcentaje de los datos de precipitación con los que se cuenta en el estado de Oaxaca de 1922 al 2009.

4.2. Detección de datos atípicos

La prueba de Grubbs se aplicó a la información de las estaciones climatológicas del estado de Oaxaca y de las estaciones climatológicas de los estados colindantes (Chiapas, Veracruz, Puebla y Guerrero), dando un total de 1437 estaciones climatológicas. El valor de α=0.01, es decir se consideró un 99 % de nivel de confianza en los datos. En la Tabla 3.4 se pueden observar los valores de Grubbs que se utilizaron para la prueba. Se obtuvieron dos valores, un límite inferior $X_L^{l,m,N}$ y un límite superior $X_H^{l,m,N}$ para el mes m que se presenta para cada uno de los años en el intervalo de [1957,2007] para una estación l en particular, $l = 1, 2, \ldots, 1437$ y $m = 1, \ldots, 12$.

En total, se tienen 17244 límites inferiores y 17244 límites superiores ya que son 1437 estaciones climatológicas, por estación son 12 límites inferiores y 12 límites superiores. En la Tabla 4.1 se resume el resultado obtenido por estado. La información que contiene dicho cuadro es el total de datos que se encuentran fuera del rango, así como los datos que se encuentran por debajo de los intervalos inferiores y por arriba de los intervalos superiores.

Como resultado, se encontró que el 16.21 % de la información mensual quedó por debajo de su respectivo límite inferior, se observo que las precipitaciones mensuales dentro de este porcentaje son de 0 ml o cercanas. Mientras que el 0.10 % de los datos se encuentra por arriba de su respectivo límite superior. De los cinco estados, Guerrero es quien tiene más datos de precipitación mensual por debajo del límite inferior (24677 datos mensuales), pero es el estado con menor información mensual que queda fuera del límite superior (22 datos mensuales). En el caso del estado de Oaxaca, es el segundo estado con mayor información mensual que queda por debajo de sus respectivos límites inferiores con 17315 datos mensuales, pero es el estado que presenta más datos mensuales por arriba de sus respectivos límites superiores, con un total de 196 datos mensuales. Esta información se resume en la Tabla 4.1.

Debido a que solo el 16.31 % de la información de la precipitación mensual de las 1437 estaciones climatológicas quedaron fuera de sus respectivos límites se decidió descartarlos de la base de datos.

Tabla 4.1: Total de datos extremos por estado, es decir, aquellos datos que están por debajo o por arriba de su respectivo límite inferior o límite superior.

Estado	**Total de datos**	**Limite inferior**	**Limite superior**
Guerrero	24699	24677	22
Oaxaca	17315	17119	196
Chiapas	11151	11083	68
Puebla	8701	8651	50
Veracruz	4923	4850	73
Total	66789	66380	409

4.3. Análisis de componentes principales

Se puede observar en la Figura 4.3 que la correlación entre las distancias al Golfo de México y al Océano Pacífico con la precipitación es pequeña, por lo que se decidió realizar otro análisis sin tomar en cuenta dichas distancias. Sin embargo, la correlación entre el resto de las variables cuantitativas con la precipitación fue mínima cómo se puede observar en la Figura 4.4, por lo que no se tomaron en cuenta en el resto del procedimiento. Esto quiere decir que la reconstrucción de las series de precipitación se realizó con datos de precipitación, sin considerar ninguna otra variable.

El análisis obtenido se presenta en la Figura 4.3, se puede observar que las estaciones climatológicas que están más cerca al Golfo presentan mayor evaporación y también hay correlación entre las temperaturas. Por otro lado, se puede observar que la precipitación queda aislada del resto de las variables aunque está cerca de la temperatura mínima. Otra observación es que la distancia al pacífico no afecta demasiado al resto de las variables.

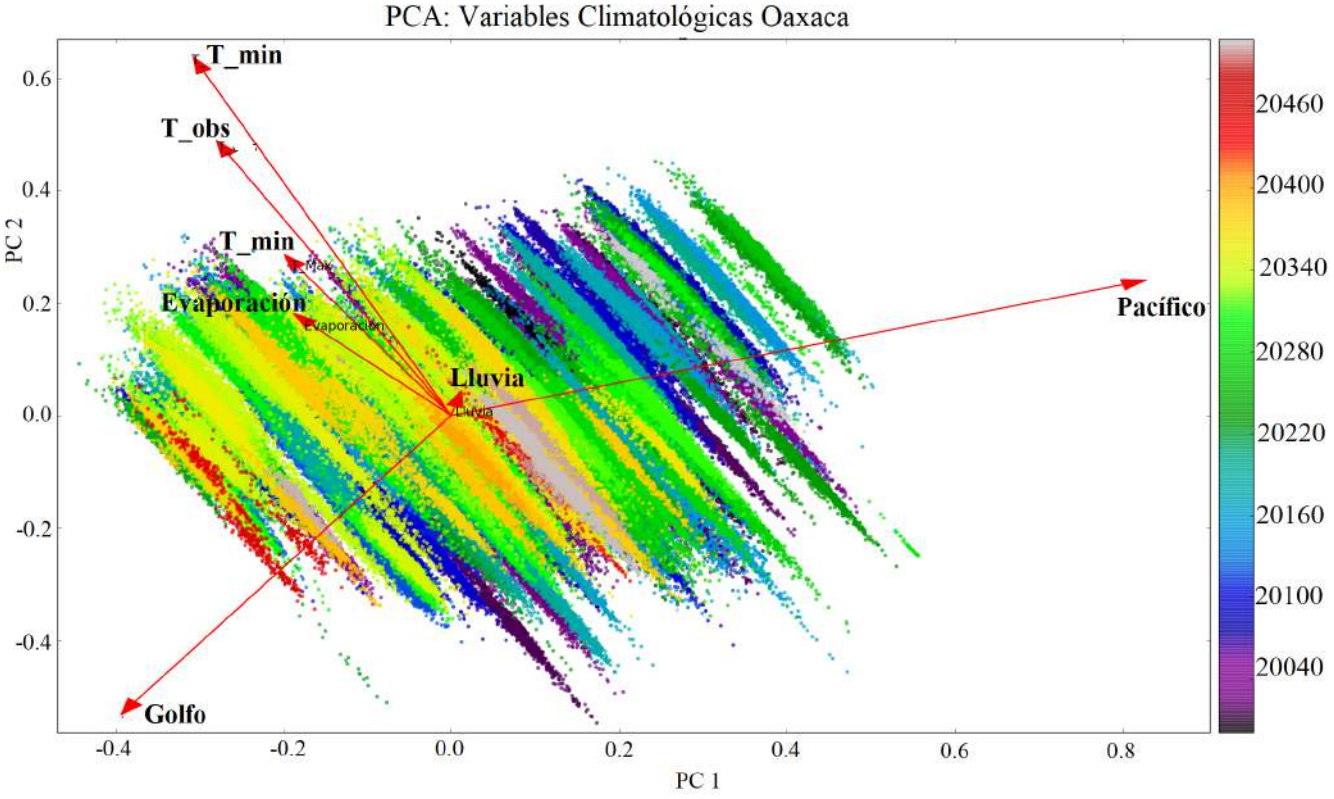


Figura 4.3: Análisis de componentes principales considerando la distancia de la estación climatológica al Océano Pacífico y al Golfo de México.

Debido a que la distancia al Golfo y al Pacífico no afecta demasiado al resto de las variables climatológicas se realiza el análisis de componentes principales sin tomar en cuenta estas distancias para observar que tanto afectan estás variables al análisis. Sin embargo, se puede observar en la Figura 4.4 que se obtienen resultados similares al análisis anterior. Por un lado se forman los grupos de temperaturas con evaporación, y en otro grupo el de la lluvia.

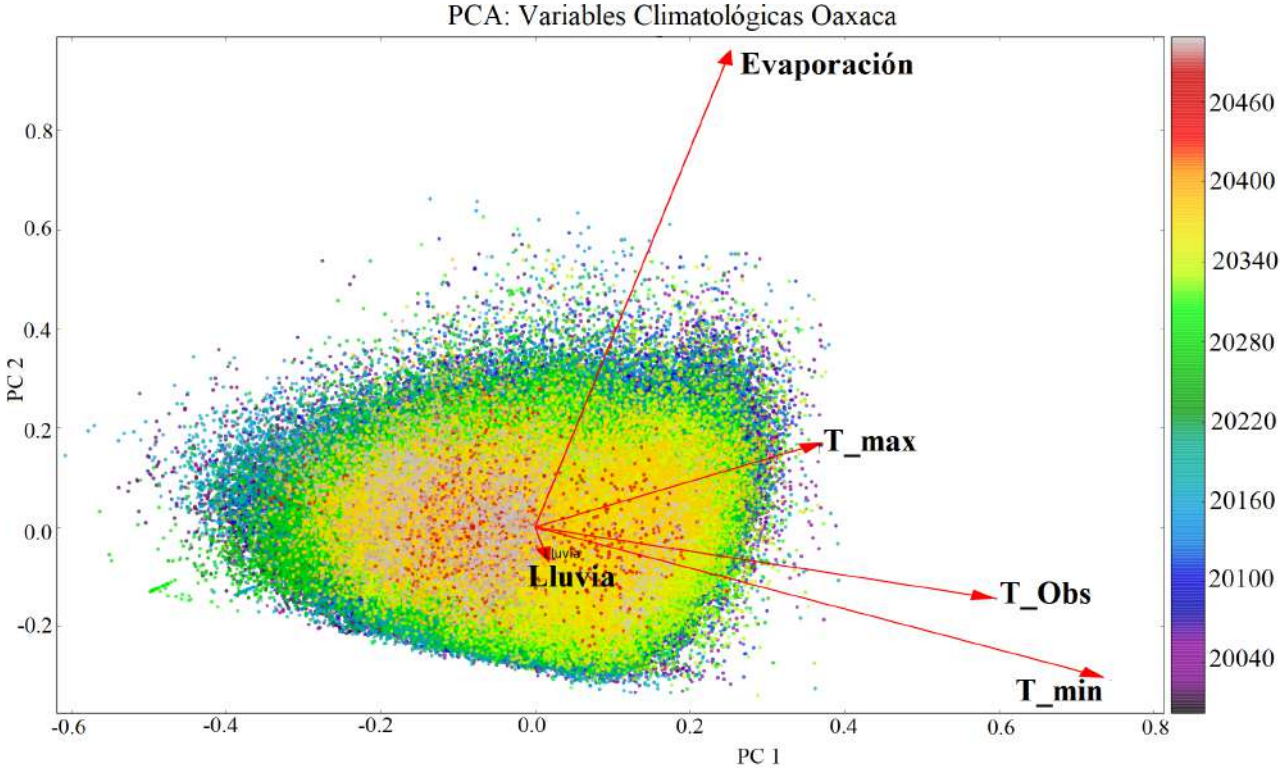


Figura 4.4: Análisis de componentes principales sin tomar en cuenta la distancia de la estación climatológica al Océano Pacífico y al Golfo de México.

Lo que se pretende con esté análisis de componentes principales, además de encontrar correlaciones entre las variables climatológicas, es que las variables con mayor correlación a la precipitación, se consideren en la elección de vecinos y reconstrucción de datos. Dado que se observa que el resto de las variables climatológicas no afectan significativamente a la precipitación, el resto del análisis solo se hará con los datos de precipitación.

En las siguientes secciones se muestra el resultado obtenido del agrupamiento así como los resultados obtenidos por los algoritmos de reconstrucción.

4.4. Agrupamiento de las estaciones climatológicas

Tras obtener una correlación baja entre la precipitación y el resto de variables climatológicas cuantitativas, se procedió a realizar el agrupamiento de las estaciones climatológicas del estado de Oaxaca y sus estados colindantes utilizando solo información de precipitación mensual. El motivo de utilizar la información de precipitación mensual de los estados colindantes es obtener mejores resultados al contar con un mayor número de datos disponibles para realizar la reconstrucción de las series de precipitación.

Es importante resaltar que, una serie de precipitación está formada por 12 datos, cada dato es la lluvia acumulada durante un mes, por lo que una serie contiene información de precipitación mensual. En total, se considera información de precipitación mensual de

Tabla 4.2: Parámetros utilizados y resultados del agrupamiento obtenidos por la red de Kohonen.

W	α	r	L	m	d	Tiempo	Grupo
0	0.01	0	1000	5	Euclidiana	10 min 21 seg	[48, 51, 163, 411, 764]
0	0.01	0	1000	10	Euclidiana	15 min 43 seg	[37, 39, 14, 110, 105, 80, 169, 363, 26, 494]
0	0.01	0	1000	15	Euclidiana	23 min 5 seg	[32, 25, 7, 33, 16, 47, 38, 104, 76, 86, 50, 178, 99, 156, 490]

1437 estaciones climatológicas pertenecientes a los estados de Oaxaca, Puebla, Guerrero, Veracruz y Chiapas.

Se obtuvieron 10371 series anuales con menos de 5 días de precipitación faltante, como se puede observar en la Figura 3.7, son series de 958 estaciones (ver Figura 3.8). Sin embargo, son 479 estaciones climatológicas que no se consideran en el conjunto obtenido, esto se debe a que las series anuales de esas estaciones tienen más de 5 días sin registros de precipitación. Para considerar a todas las estaciones climatológicas en el agrupamiento, se decidió agregar a las series anuales con el menor número de días no registrados, de las 479 estaciones climatológicas que no se están considerando; con lo cual se agregaron 958 series anuales al conjunto. Con esto, el conjunto E consta de $n = 11329$ series anuales.

Para las primeras pruebas realizadas, los pesos que conforman al conjunto W se inicializaron de forma aleatoria con una distribución uniforme $U(0, 1)$, distancia euclidiana y 200 neuronas en la capa de salida, es decir, se le dio la libertad a la red para generar los grupos. Sin embargo, al probar con diferentes valores del resto de los parámetros $(L, r, \alpha,)$ se obtuvo solo un grupo. Este resultado no se consideró correcto ya que la precipitación de todas las estaciones climatológicas no se pueden comportar de forma similar. Entonces, se optó por inicializar los pesos con el mismo valor. Después de varías pruebas se observó que el valor de inicio de los pesos con el que se generaron diferentes grupos fue cero.

Al inicializar los pesos de entrada de la red de Kohonen en cero, con 200 neuronas, y con el valor de r y α cambiando con respecto al número de iteraciones t (ecuación (3.7)), se obtuvo un solo grupo, sin importar el número de iteraciones con 1000, 3000 o 5000 iteraciones se obtuvo un solo grupo. Debido a estos resultados se decidió dejar el valor de r y α fijos en todas las iteraciones. Después de diferentes pruebas y combinaciones de parámetros se encontró que, con el valor de $r = 0$ y α=0.01 se obtuvieron varios grupos. Esto es, con los pesos inicializados en cero, $r = 0$, α=0.01, $L = 1000$, $m = 200$ se obtuvieron 102 grupos en un tiempo de 2 horas 6 minutos 24 segundos. Sin embargo, esto no es posible desde el punto de vista hidrológico, ya que no hay 102 climas diferentes en el país. Debido a eso se decidió restringir a la red de Kohonen con el número de neuronas, el número de neuronas en la capa de salida m, se restringió a 5, 10 o 15 neuronas.

En las Figuras 4.5, 4.6 y 4.7 se puede observar el agrupamiento de las 1437 estaciones climatológicas de los estados utilizados en la red de Kohonen (Oaxaca, Veracruz, Chiapas, Puebla y Guerrero) utilizando los parámetros que se describen en la Tabla 4.2. En este cuadro, la última columna es un vector donde cada elemento representa el número de estaciones climatológicas en cada grupo. El mayor tiempo de ejecución se obtuvo al agru-

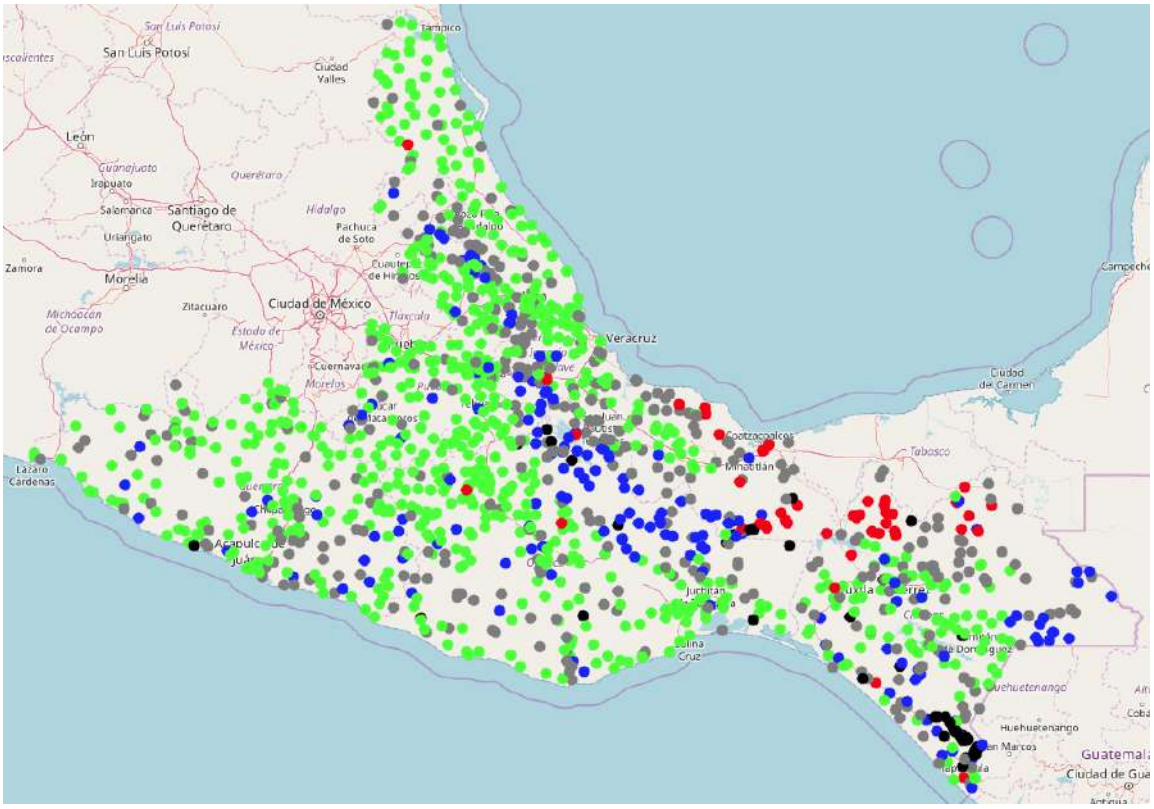


Figura 4.5: Resultado obtenido por la red de Kohonen al agrupar las 1437 estaciones climatológicas en 5 grupos. Los parámetros utilizados son: W=0, α=0.01, $r = 0$, L=1000, distancia euclidiana. En un tiempo de ejecución de 10 minutos con 21 segundos.

par a las estaciones climatológicas en 20 grupos. Los resultados mostrados corresponden a 1000 iteraciones. Con 3000 y 5000 iteraciones se obtuvieron resultados similares pero el tiempo de ejecución superiores al tiempo utilizado con 1000 iteraciones.

Algo interesante que se pudo observar en los tres agrupamientos fue que las estaciones climatológicas que están relativamente cercan no necesariamente pertenecen al mismo grupo, lo cual es un indicador de la gran variedad en climas que existen en el estado de Oaxaca.

En la Figura4.8 se observa una gráfica de barras que representa el número de estaciones climatológicas en cada grupo. Las barras de color azul indican el número de estaciones climatológicas que se concentraron en cada uno de los 5 grupos. Las barras de color anaranjado y amarillo son para el agrupamiento de 10 y 15 grupos obtenidos por la red de Kohonen. De los tres casos se puede observar que la mayor concentración de estaciones climatológicas se encuentran en el último grupo. Para el primer caso, donde las estaciones climatológicas se clasifican en cinco grupos se puede observar que en el grupo 5 se aglomeran 764 estaciones climatológicas que representan el 53.1 % del total, y la menor concentración se encuentra en el grupo 1 con 48 estaciones climatológicas que es el 3.3 % del total. Para el segundo caso, las estaciones climatológicas se clasifican en 10 grupos, como resultado se obtiene que en el grupo 10 se congregan 494 estaciones climatológicas que representa el 34.3 %, pero el grupo con menor número de estaciones climatológicas es

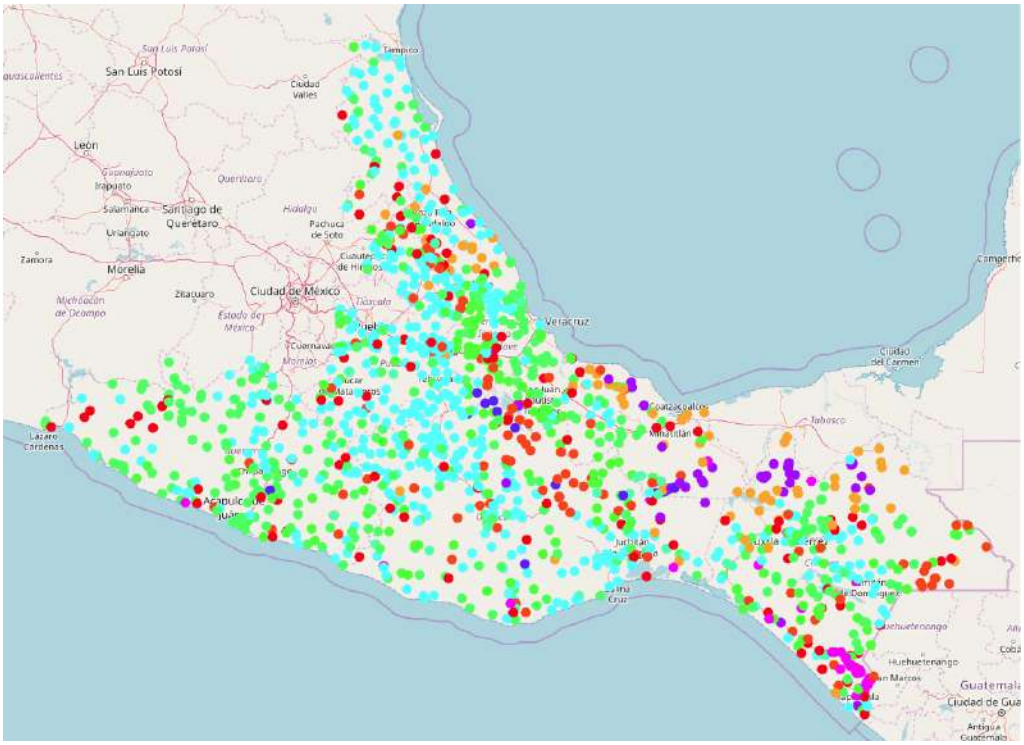


Figura 4.6: Resultado obtenido por la red de Kohonen al agrupar las 1437 estaciones climatológicas en 10 grupos. Los parámetros utilizados son: W=0, α=0.01, $r = 0$, L=1000, distancia euclidiana. En un tiempo de ejecución de 15 minutos con 43 segundos.

el 3 con solo el 0.9 %, que son 37 estaciones climatológicas. Por último, el agrupamiento en 15 grupos, el 34.1 % de las estaciones climatológicas se concentra en el grupo 15 que es el grupo con mayor número de estaciones climatológicas, mientras que el grupo con menor número de estaciones climatológicas es el grupo 3 con el 0.5 % de ellas, que son 32 estaciones climatológicas.

Aunque el agrupamiento de la red de Kohonen se realizó con series mensuales de precipitación que tuvieran menos de cinco días faltantes, este agrupamiento se puede extender al resto de las series que no fueron consideradas. Esta es la ventaja de la red de Kohonen, una vez que fue entrenada, es posible ingresar otra serie mensual y con los pesos obtenidos en el entrenamiento puede ser clasificada a un grupo determinado.

Cada uno de estos agrupamientos (5, 10 y 15 grupos) obtenidos por la red de Kohonen se utilizaron en el algoritmo de retropropagación para ajustar los pesos de esta red y así obtener la reconstrucción de series de precipitación. En el caso de la red de contrapropagación, ésta cuenta con una capa oculta donde utiliza la técnica de la red de Kohonen, por lo que el agrupamiento se realizó en el proceso. Los resultados de las redes neuronales utilizadas para la reconstrucción de series se mencionan en la siguiente sección.

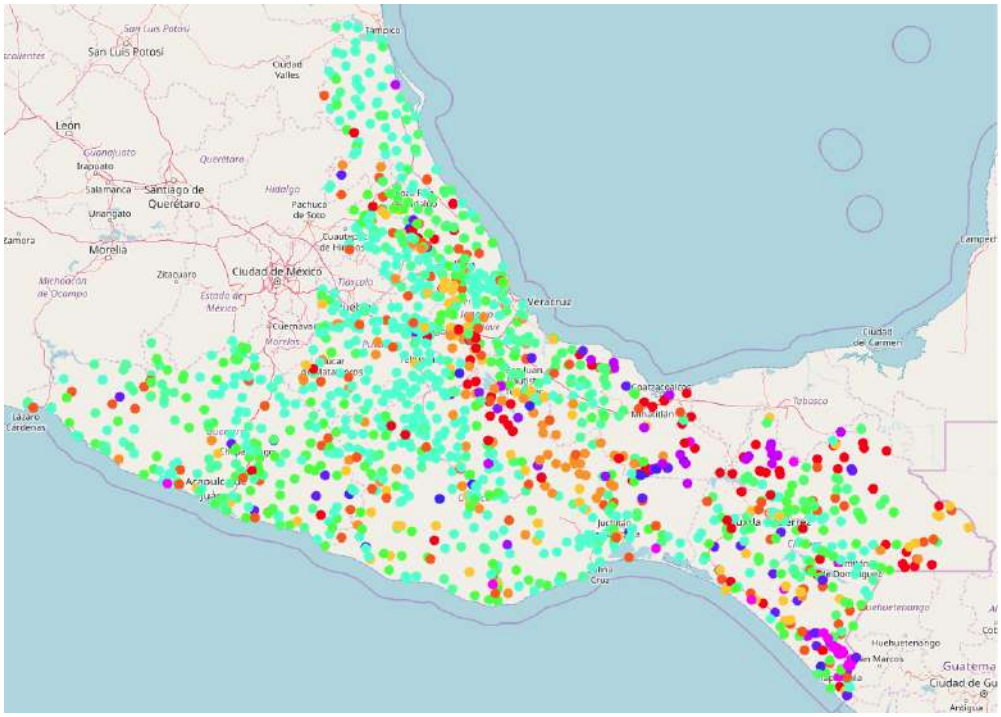


Figura 4.7: Resultado obtenido por la red de Kohonen al agrupar las 1437 estaciones climatológicas en 15 grupos. Los parámetros utilizados son: W=0, α=0.01, $r = 0$, L=1000, distancia euclidiana. En un tiempo de ejecución de 23 minutos con 5 segundos.

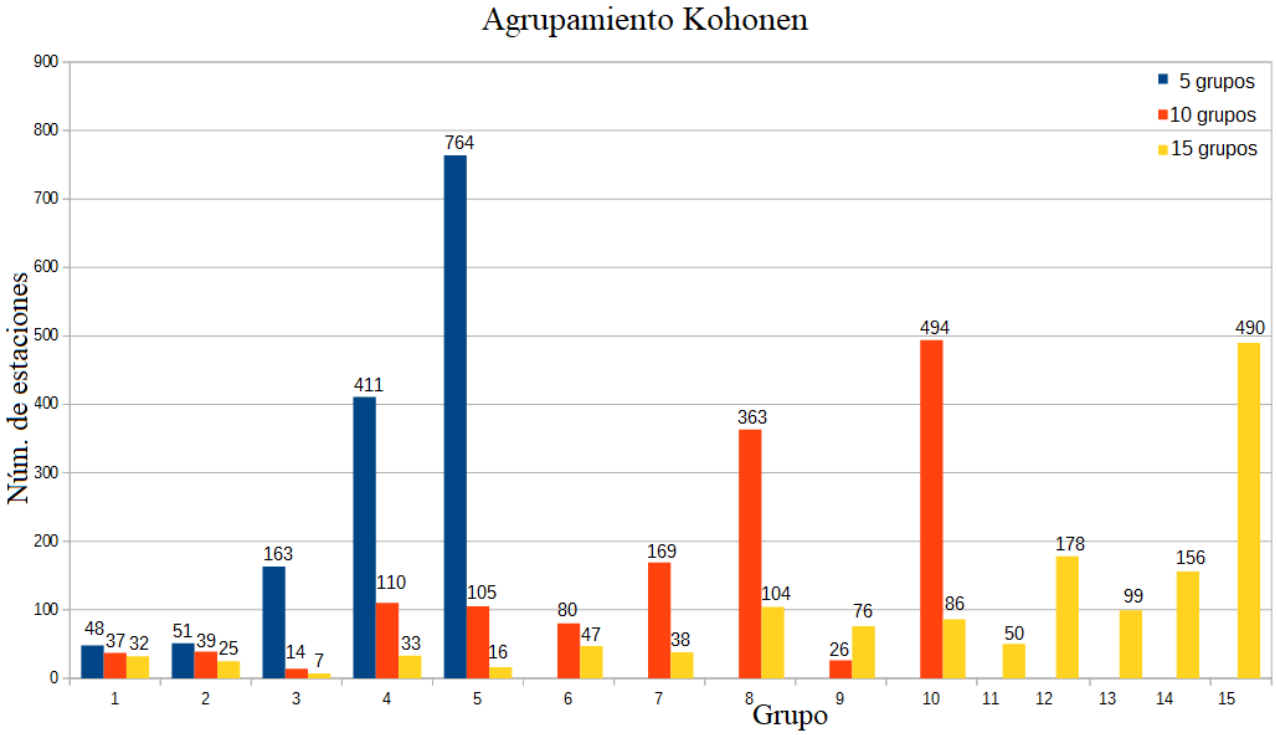


Figura 4.8: Representación del número de estaciones climatológicas en cada grupo utilizando un agrupamiento de 5, 10 y 15 grupos en la red de Kohonen.

Tabla 4.3: Casos con los que se entrenó la red de retropropagación. La segunda columna corresponde al número de vecinos NV, la tercera columna indica el número de grupos n, y la última columna indica el tamaño del conjunto de entrenamiento X que es ingresado a la red.

Caso de entrenamiento	NV	n	Tamaño del conjunto
1	3	5	3× 60
2	3	10	3× 120
3	3	15	3× 180
4	5	5	5× 60
5	5	10	5× 120
6	5	15	5× 180

4.5. Reconstrucción de las series

Tras observar que las variables climatológicas: temperatura observada, temperatura máxima, temperatura mínima, evaporación, distancia al Golfo de México y al Océano Pacífico, presentan una baja correlación con la precipitación, y después de realizar el agrupamiento de las estaciones, se procedió a la reconstrucción de series de precipitación. Primero se analizó el entrenamiento de retropropagación y después el de contrapropagación.

Algoritmo de retropropagación

El conjunto de entrenamiento X es de tamaño $NV \times N$ donde NV es el número de vecinos, $N = 12 * n$ con n el número de grupos formados en la red de Kohonen y 12 es por el total de meses en un año. Para las pruebas realizadas se utilizaron tres y cinco vecinos, es decir, $NV \in \{3, 5\}$. El agrupamiento fue de 5, 10 y 15 grupos con la red de Kohonen. Debido a esto, $n \in \{5, 10, 15\}$, esto quiere decir que se hicieron pruebas con seis casos de entrenamiento donde el tamaño del conjunto de entrenamiento dependió de cada caso como se describe en la Tabla 4.3.

Los parámetros utilizados fueron: función de activación, coeficiente de aprendizaje, número de iteraciones y el número de neuronas en la capa intermedia, que al combinarlos dan un total de 105 pruebas para cada uno de los seis casos de entrenamiento. Cada prueba se ejecutó 10 veces lo que da un total de 6300 pruebas para los seis casos.

En la Tabla 4.4 se pueden observar los parámetros utilizados y con los que se obtuvo el menor error MSE en cada una de las seis descritas. Se puede ver que en cinco de las seis pruebas con menor MSE se utilizó la función $h(x)$ que es Gaussiana, esto se debe a que la distribución de los datos de precipitación son similares a la distribución de una función Gaussiana y se alejan de las distribuciones hiperbólica o sigmoidal. Cabe mencionar que solo se muestra el mejor resultado obtenido para cada uno de los casos, más adelante se mencionan los primeros 10 mejores resultados para cada grupo generado por la red de Kohonen. Solo en el caso de entrenamiento 6 se obtuvo el mínimo MSE utilizando la

función tangente hiperbólica. Se puede ver que los tres casos con menor MSE se obtuvieron con cinco vecinos, además el menor fue con cinco grupos y el MSE más grande fue con 15 grupos. Otra cosa interesante es el número de neuronas en la capa intermedia. En estas pruebas se utilizaron 20, 30 o 50 neuronas, no menos de 20. Además, en ninguna de éstas el coeficiente de aprendizaje α fue menor a 0.05. Con respecto al tiempo de ejecución, no hay gran diferencia entre utilizar cinco o tres vecinos; en promedio la diferencia es de 5 minutos con 33 segundos.

Tabla 4.4: Pruebas del algoritmo de retropropagación con el menor error MSE para cada uno de los seis casos descritos en la tabla 4.3.

Caso de entrenamiento	NV[a]	n[b]	m[c]	Función[d]	α[e]	L[f]	MSE[g]	Tiempo[h]
4	5	5	20	$h(x)$	0.25	500000	3.3×10^{-6}	00:25:16
5	5	10	50	$h(x)$	0.10	500000	7.2×10^{-4}	00:47:39
6	5	15	50	$g(x)$	0.10	500000	29.1×10^{-4}	01:28:26
1	3	5	30	$h(x)$	0.10	500000	43.3×10^{-4}	00:21:06
2	3	10	20	$h(x)$	0.05	500000	119.4×10^{-4}	00:44:13
3	3	15	20	$h(x)$	0.05	500000	719.1×10^{-4}	01:05:05

[a] NV: número de vecinos.
[b] n: número de grupos.
[c] m: número de neuronas en la capa oculta.
[d] Función de activación utilizada en la red.
[e] α: coeficiente de aprendizaje.
[f] L: total de iteraciones.
[g] MSE: error cuadrático medio.
[h] Tiempo de ejecución del algoritmo expresado en horas:minutos:segundos.

Se ordenaron las pruebas realizadas de menor a mayor con respecto al MSE, para cada uno de los grupos obtenidos por la red de Kohonen (5, 10 y 15 grupos). En la Tabla 4.5 se describen los parámetros utilizados y el tiempo de ejecución de las 10 pruebas con menor MSE, para cuando las estaciones climatológicas se clasifican en 5 grupos, es decir, en estas pruebas n=5. Analizando el número de vecinos (NV) podemos ver que los 10 valores mas bajos del MSE corresponden a 5 vecinos. Hasta el lugar 18 se posiciona la primera prueba donde se ocuparon 3 vecinos obteniendo un MSE de 43.29×10^{-4} (m =20, función: $h(x)$, α =0.1, L =50000, Tiempo: 00:21:06).

El número de neuronas m utilizadas en estas las pruebas fueron: 2, 3, 5, 10, 20, 30, 50, pero en las 10 pruebas que se mencionan en la Tabla 4.5 utilizaron 10, 20, 30 o 50 neuronas. Es hasta la prueba posicionada en el lugar 14 donde se presenta la primera que utiliza 5 neuronas (función: $h(x)$, α=0.1, L= 50000, MSE= 16.67×10^{-4}, Tiempo= 00:25:00), en la posición 32 y 44 quedaron las primeras pruebas que utilizan 3 y 2 neuronas, respectivamente y cabe mencionar que en éstas también se ocuparon 5 vecinos.

También se observa que en 2 de las 10 pruebas con menor MSE fueron con la función tangente hiperbólica y el resto con la función Gaussina. Hasta la posición 33 se encuentra la primera prueba que obtuvo el menor MSE utilizando la función sigmoide con m= 10, α= 0.1, MSE de 79.52×10^{-4} en un tiempo de 29 minutos 58 segundos.

Tabla 4.5: Las 10 pruebas del algoritmo de retropropagación con el menor error MSE agrupando a las estaciones climatológicas en 5 grupos.

#	Caso de entrenamiento	NV	m	Función	α	L	MSE	Tiempo
1	4	5	20	$h(x)$	0.25	500000	3.34×10^{-6}	00:25:16
2	4	5	50	$h(x)$	0.10	500000	1.85×10^{-5}	00:25:35
3	4	5	50	$h(x)$	0.05	500000	6.90×10^{-5}	00:26:43
4	4	5	10	$h(x)$	0.25	500000	1.23×10^{-4}	00:23:49
5	4	5	30	$h(x)$	0.05	500000	4.74×10^{-4}	00:26:10
6	4	5	10	$g(x)$	0.25	500000	5.53×10^{-4}	00:30:03
7	4	5	50	$g(x)$	0.05	500000	5.67×10^{-4}	00:35:37
8	4	5	30	$h(x)$	0.10	500000	7.75×10^{-4}	00:26:19
9	4	5	20	$h(x)$	0.10	500000	10.38×10^{-4}	00:26:40
10	4	5	20	$h(x)$	0.05	500000	11.65×10^{-4}	00:26:59

El valor del coeficiente de aprendizaje α en estas pruebas fue de 0.05, 0.1 o 0.25, la prueba con el menor MSE y que utilizó α=0.001 quedó posicionada en el lugar 40 con parámetros NV=5, m =50, $h(x)$, y MSE de 101.75×10^{-4} en un tiempo de 25 minutos con 59 segundos. La prueba con el menor MSE y que utilizó α=0.01 quedó en el lugar 26 con parámetros: NV=5, m=20, función Gaussiana, obteniendo un MSE de 66.24×10^{-4} en 27 minutos y 31 segundos. Por último, la prueba con mayor tiempo de ejecución fue de 35 minutos con 37 segundos.

En la Figura 4.9 se presenta la comparación de las series de precipitación anual de cada grupo con su serie reconstruida obtenida por la red de retropropagación con el menor MSE que fue de 3.343×10^{-6} obtenida en un tiempo de 25 minutos con 16 segundos, con los parámetros: NV=5, n=5, m=20, función $h(x)$, α=0.25, L=500000. En las gráficas de la Figura 4.9 correspondientes a los grupos de 1 al 4, la serie reconstruida queda sobre la serie anual de precipitación original, esto quiere decir se tiene una adecuada reconstrucción de la serie, dado que tiene un comportamiento casi idéntico al de la serie original. En la gráfica del grupo 5 de dicha figura, solo en las colas se puede ver una pequeña diferencia entre la original y la reconstruida. En la última gráfica de la Figura 4.9 se muestra el comportamiento que tuvo el MSE en cada iteración, se puede ver que al principio se obtuvieron algunos brincos pero después de 100000 iteraciones la gráfica es decreciente.

Con respecto a los resultados obtenidos con el algoritmo de retropropagación para 10 grupos, se puede observar en la Tabla 4.6 los parámetros utilizados así como el error MSE y el tiempo obtenido de las 10 mejores pruebas. Se observa que las 10 pruebas con menor MSE obtenido fueron con 5 vecinos, al igual que los resultados obtenidos por la clasificación en 5 grupos. Hasta la posición 18 se tiene la primera prueba con el menor MSE utilizando 3 vecinos en un tiempo de 44 minutos y 13 segundos (m =30, función: $h(x)$, α=0.05, L=50000, MSE=119.43×10^{-4}).

El número de neuronas m utilizadas en la capa oculta de la red de retropropagación en las pruebas descritas en la Tabla 4.6 fue de 20, 30 o 50 neuronas, en particular, la prueba con menor MSE se obtuvo utilizando 50 neuronas. Las primeras pruebas que obtuvieron el

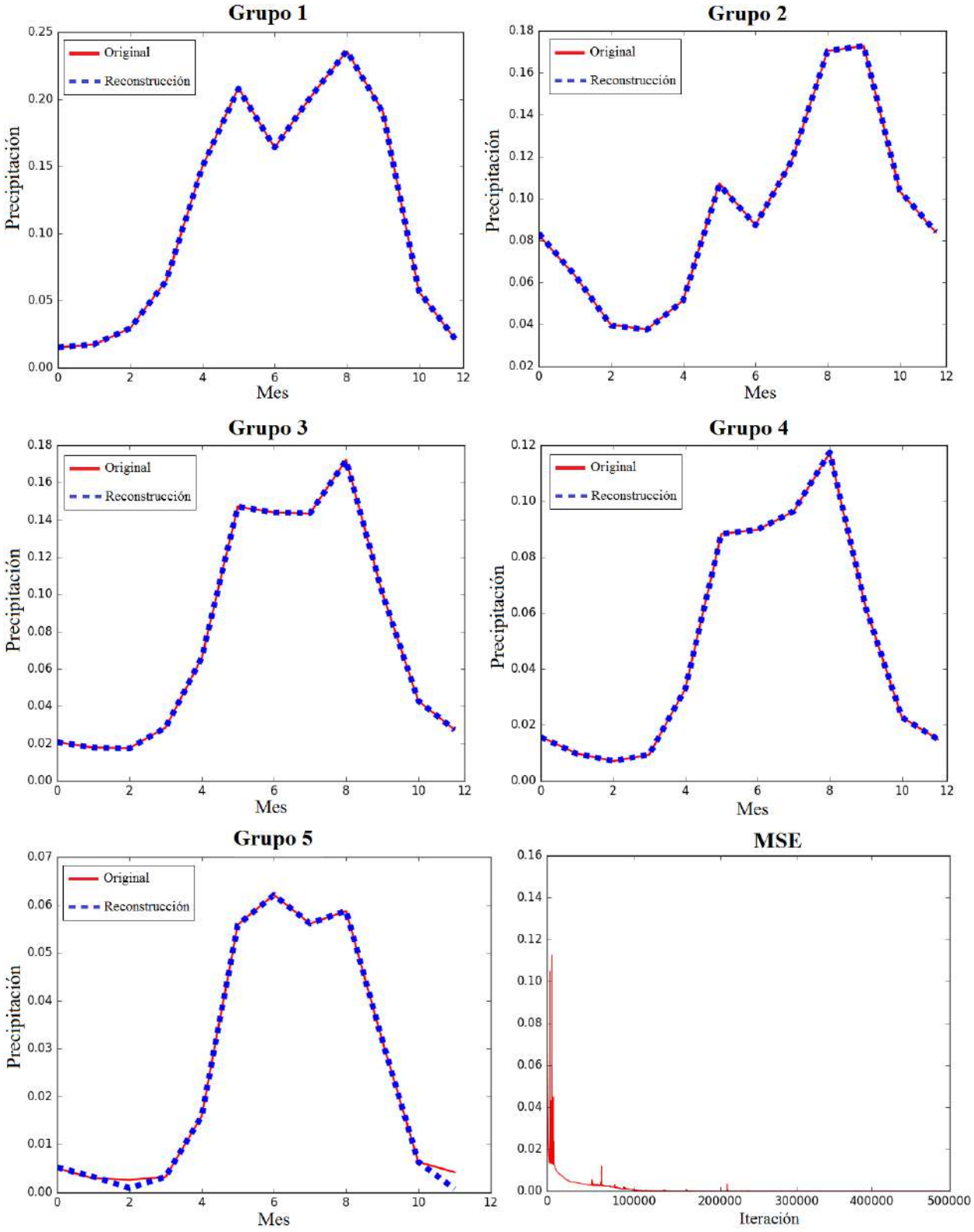


Figura 4.9: Gráfica de las series mensuales de precipitación que representa cada uno de los 5 grupos y sus series reconstruidas obtenida por la prueba con menor MSE en el algoritmo de retropropagación. Además de la representación gráfica del comportamiento del MSE en cada iteración del algoritmo.

Tabla 4.6: Las 10 pruebas del algoritmo de retropropagación con el menor error MSE agrupando a las estaciones climatológicas en 10 grupos.

#	Caso de entrenamiento	NV	m	Función	α	L	MSE	Tiempo
1	5	5	50	$h(x)$	0.10	500000	7.19×10^{-4}	00:47:39
2	5	5	50	$h(x)$	0.05	500000	16.68×10^{-4}	00:44:08
3	5	5	20	$h(x)$	0.10	500000	20.29×10^{-4}	00:40:26
4	5	5	50	$g(x)$	0.05	500000	31.42×10^{-4}	00:53:44
5	5	5	20	$g(x)$	0.10	500000	31.57×10^{-4}	00:58:06
6	5	5	20	$h(x)$	0.05	500000	58.33×10^{-4}	00:43:40
7	5	5	30	$g(x)$	0.05	500000	65.01×10^{-4}	00:50:48
8	5	5	30	$h(x)$	0.10	500000	67.70×10^{-4}	00:42:09
9	5	5	50	$h(x)$	0.01	500000	73.07×10^{-4}	00:46:03
10	5	5	20	$g(x)$	0.25	500000	79.40×10^{-4}	00:52:00

menor MSE con 2, 3, 5 o 10 neuronas quedaron en el lugar 52, 35, 14 y 15 respectivamente, es importante resaltar que en dichas pruebas también se utilizaron 5 vecinos. Para la función de activación, 6 de las 10 pruebas con menor MSE se obtuvo utilizando la función Gaussiana, las otras 4 se obtuvieron con la función tangente hiperbólica, y la primera prueba con menor MSE con la función sigmodal quedó en el lugar 23.

Los valores del coeficiente de aprendizaje utilizados fueron α de 0.01, 0.05, 0.1 o 0.25, donde la prueba con menor MSE corresponde a α=0.1. Dejando en el lugar 28 a la prueba con α de 0.001 en un tiempo de 44 minutos y 48 segundos (NV=5, m =20, función: $h(x)$, MSE= 154.41×10^{-4}). El tiempo de ejecución máximo para las pruebas realizadas con 10 grupos fue de 1 hora 1 minuto y 10 segundos.

En las Figuras 4.10 y 4.11 se pueden observar la comparación de las series de precipitación anual de cada grupo con su serie reconstruida obtenida con el menor MSE igual a 7.19×10^{-4} en un tiempo de ejecución de 47 minutos con 39 segundos utilizando los parámetros: NV =5, n =10, m=50, función $h(x)$, α =0.1, L =500000. Se puede observar que es mínima la diferencia entre las series que representan a los grupos 1, 2, 4, 7 y 9 con sus series reconstruidas, por otro lado las series de los grupos 8 y 10 son los que más presentan diferencia en comparación a los otros grupos. En la Figura 4.11 se presenta la gráfica del comportamiento del MSE durante las 5000 iteraciones, esta prueba inició con un error muy grande, cerca de 60 pero en la segunda iteración fue menor a 1.

Con respecto a los resultados obtenidos por la red de retropropagación aplicada a la clasificación de las estaciones climatológicas del estado de Oaxaca en 15 grupos, en la Tabla 4.7 se describen los parámetros que se utilizaron para obtener las 10 pruebas con menor MSE. El menor MSE obtenido fue de 29.13×10^{-4} en un tiempo de 1 hora con 28 minutos y 26 segundos utilizando los parámetros: NV =5, m=50, función $g(x)$, α =0.1, L=500000.

Cabe señalar que, en estas pruebas los 10 menores MSE se obtuvieron ocupando cinco vecinos al igual que en las pruebas aplicadas a los 5 y 10 grupos. La prueba con menor

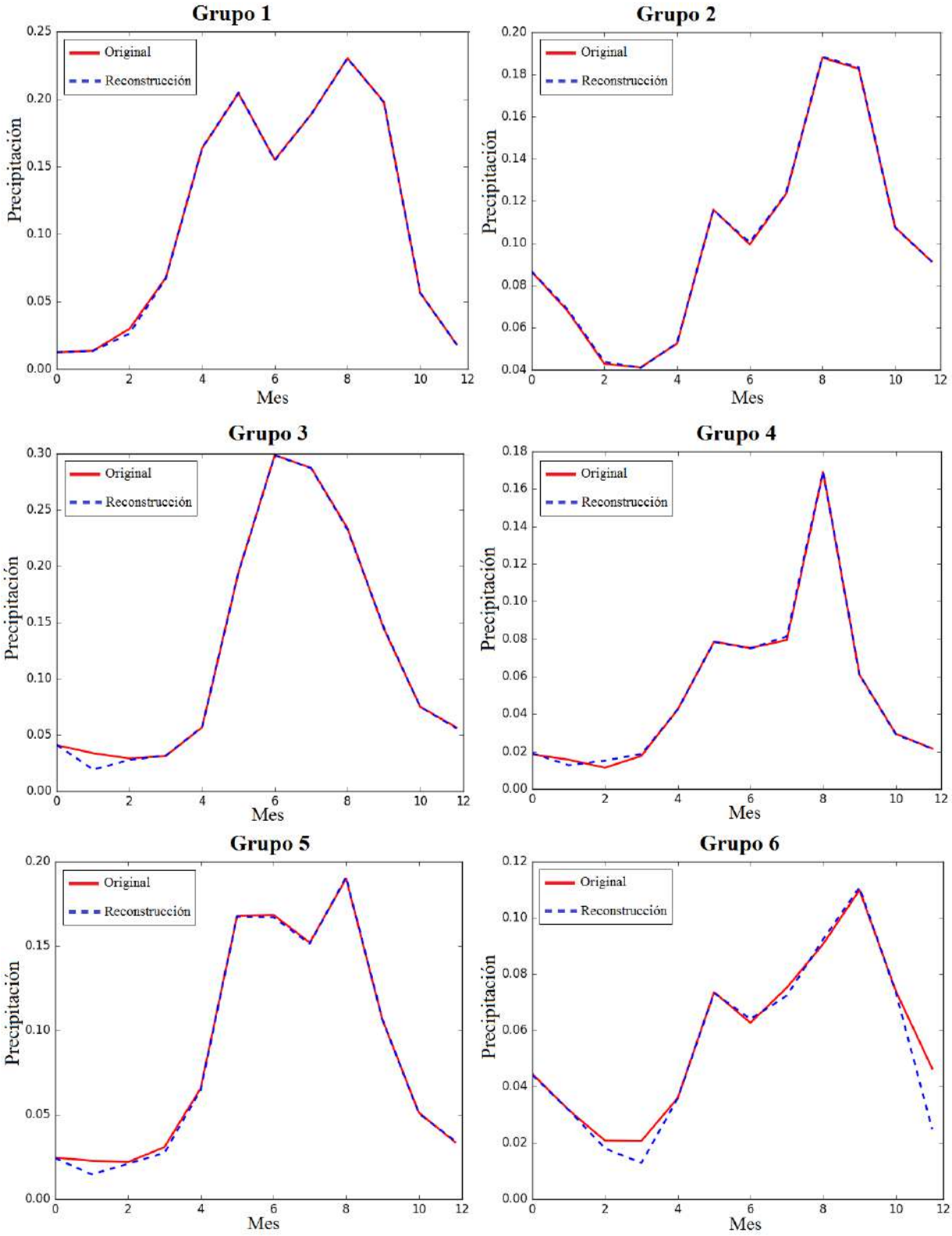


Figura 4.10: Comparación de la series mensuales de precipitación con las series reconstruidas, obtenidas por la red de retropropagación de los primeros 6 grupos del agrupamiento de las estaciones climatológicas del estado de Oaxaca en 10 grupos.

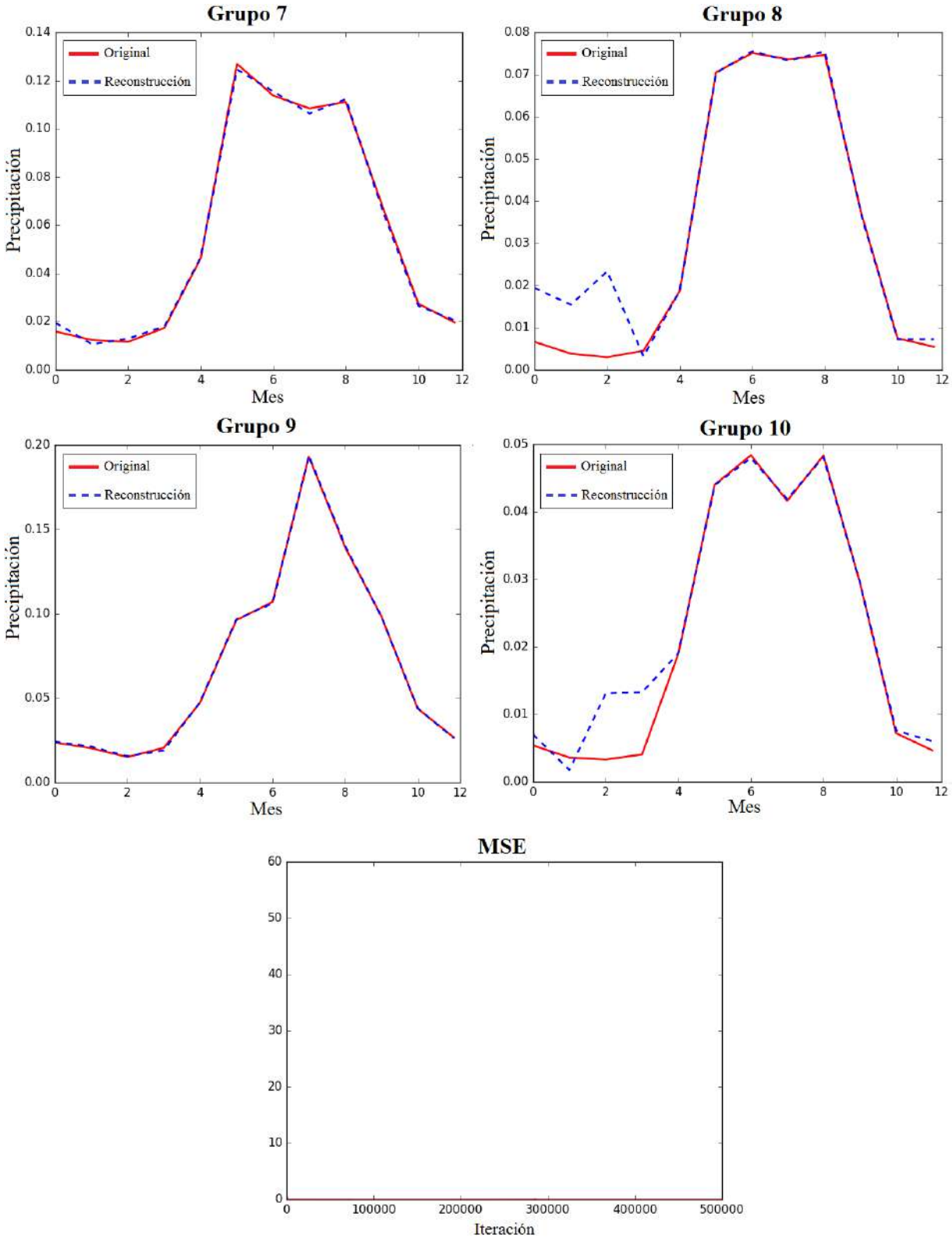


Figura 4.11: Comparación de las series mensuales de precipitación con las series reconstruidas, obtenidas por la red de retropropagación de los 4 últimos grupos del agrupamiento de las estaciones climatológicas del estado de Oaxaca en 10 grupos. Así como el comportamiento del MSE durante las iteraciones.

Tabla 4.7: Las 10 pruebas del algoritmo de retropropagación con el menor error MSE agrupando a las estaciones climatológicas en 15 grupos.

#	Caso de entrenamiento	NV	m	Función	α	L	MSE	Tiempo
1	6	5	50	$g(x)$	0.10	500000	29.13×10^{-4}	01:28:26
2	6	5	50	$g(x)$	0.05	500000	83.38×10^{-4}	01:21:36
3	6	5	30	$h(x)$	0.05	500000	100.23×10^{-4}	01:03:35
4	6	5	50	$h(x)$	0.05	500000	102.63×10^{-4}	01:09:08
5	6	5	50	$h(x)$	0.01	500000	115.40×10^{-4}	01:11:40
6	6	5	20	$g(x)$	0.05	500000	138.63×10^{-4}	01:16:36
7	6	5	20	$h(x)$	0.01	500000	163.45×10^{-4}	01:02:21
8	6	5	30	$h(x)$	0.10	500000	163.99×10^{-4}	01:02:42
9	6	5	30	$h(x)$	0.01	500000	169.21×10^{-4}	01:04:11
10	6	5	20	$h(x)$	0.1	500000	195.54×10^{-4}	01:03:12

MSE utilizando 3 vecinos quedó en el lugar 45. Además, estas 10 pruebas se obtuvieron con 50, 30 o 20 neuronas, más aún el menor MSE se obtuvo ocupando 50 neuronas.

Para este agrupamiento, el menor MSE se obtuvo utilizando la función tangente hiperbólica pero 7 de las 10 pruebas se obtuvieron utilizando la función Gaussiana. Es hasta la prueba en la posición 25 que se obtiene le menor MSE utilizando la función sigmoide. Pasando al análisis del coeficiente de aprendizaje, en estas 10 pruebas solo se ocupó α=0.01, 0.05, 0.1.

En la Figura 4.12, 4.13 y 4.14 se muestran las series de precipitación anual representantes de cada uno de los 15 grupos así como su serie aproximada de la prueba que obtuvo el menor MSE del algoritmo de retropropagación (prueba 1 de la Tabla 4.7). En los grupos 1 y 2 se obtuvieron las mejores aproximaciones de los 15 grupos, sin embargo, en los últimos tres grupos (13, 14 y 15) es donde podemos ver una diferencia notable entre la serie anual original y la aproximada. La última gráfica de la Figura 4.14 muestra el comportamiento del MSE durante las 500000 iteraciones, se puede observar que antes de las 100000 se presentan varias perturbaciones pero después el MSE tiene un comportamiento decreciente.

Una vez que se realizaron las diferentes pruebas con el conjunto de entrenamiento X, se eligió la prueba con el menor MSE para cada agrupamiento y se guardaron los pesos W de dichas pruebas los cuales son los elegidos para realizar la reconstrucción de las series de precipitación. En total se tienen 17901 series de precipitación anual del estado de Oaxaca que componen al conjunto T_p que es el conjunto que se quiere obtener al ingresar el conjunto de prueba a la red de retropropagación. Para cada agrupamiento, el conjunto de prueba P fue ingresado a la red mediante la aplicación del algoritmo de propagación. El conjunto P está compuesto de la información de las estaciones climatológicas vecinas de las series del estado de Oaxaca. Dado que, los valores menores de MSE se obtuvieron utilizando 5 vecinos, para cada serie del estado de Oaxaca se decidió almacenar 5 series al conjunto P, dando un total de 89505 series.

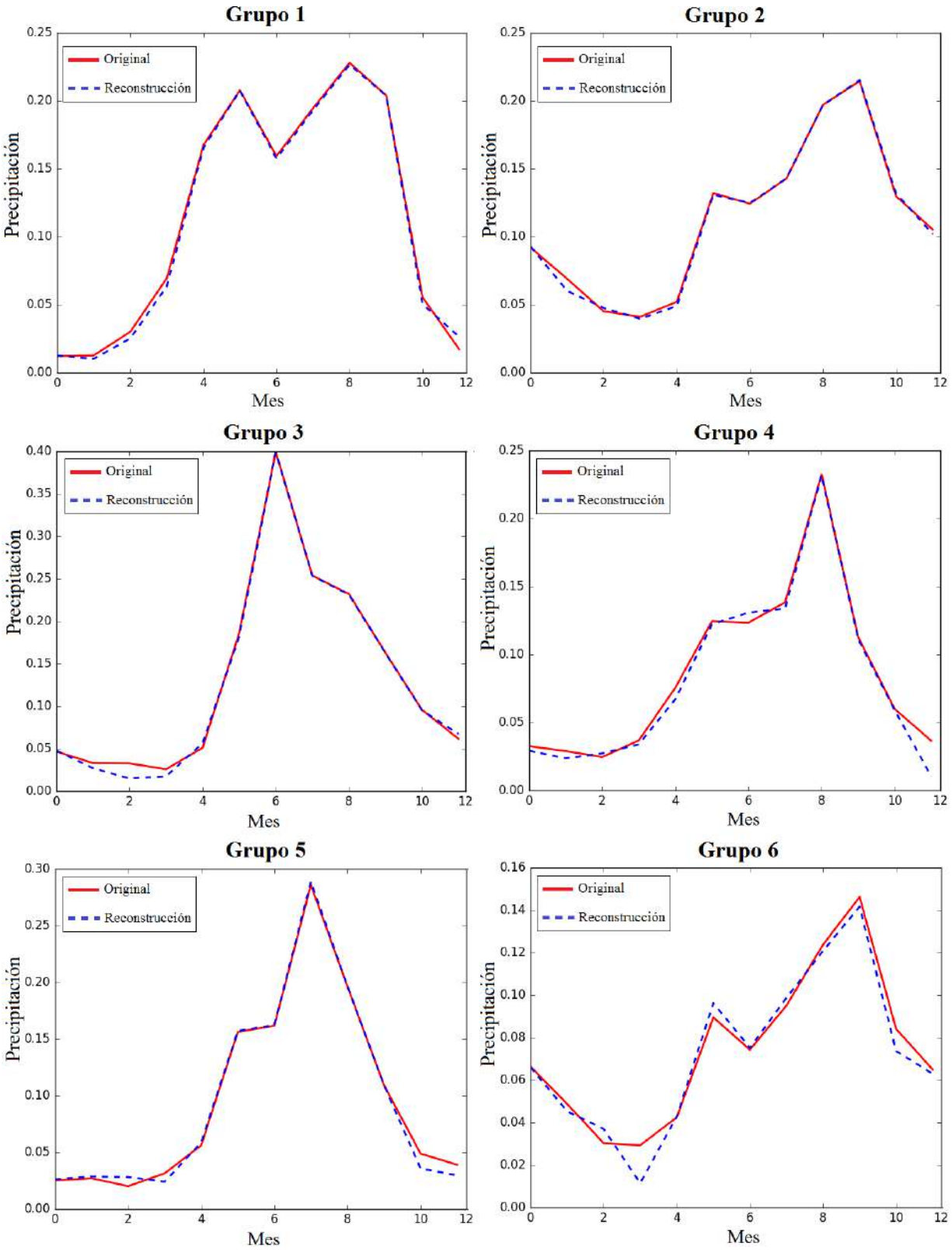


Figura 4.12: Comparación de las series mensuales de precipitación con las series reconstruidas, obtenidas por la red de retropropagación de los primeros 6 grupos del agrupamiento de las estaciones climatológicas del estado de Oaxaca en 15 grupos.

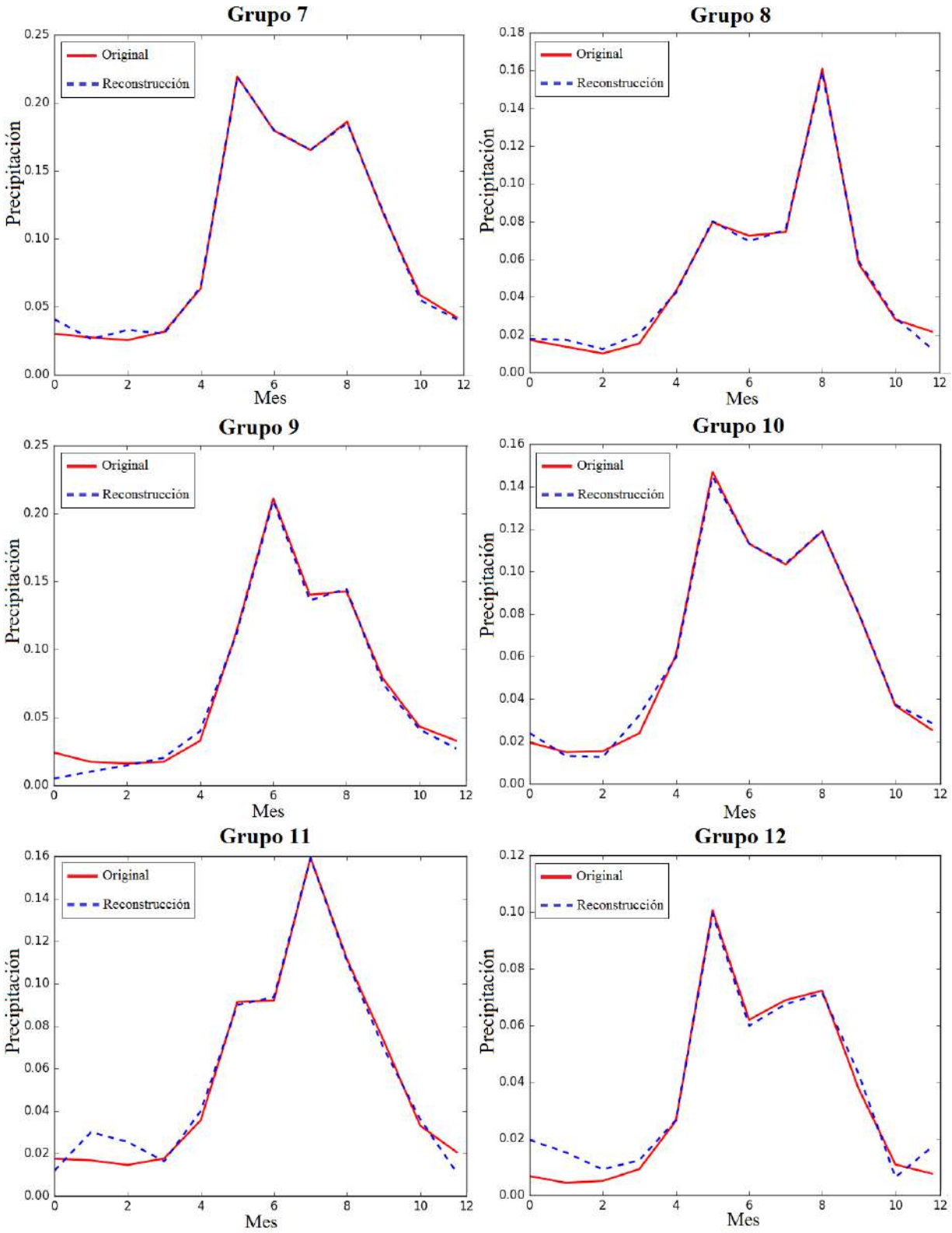


Figura 4.13: Comparación de las series mensuales de precipitación con las series reconstruidas, obtenidas por la red de retropropagación del grupo 7 al 12 del agrupamiento de las estaciones climatológicas del estado de Oaxaca en 15 grupos.

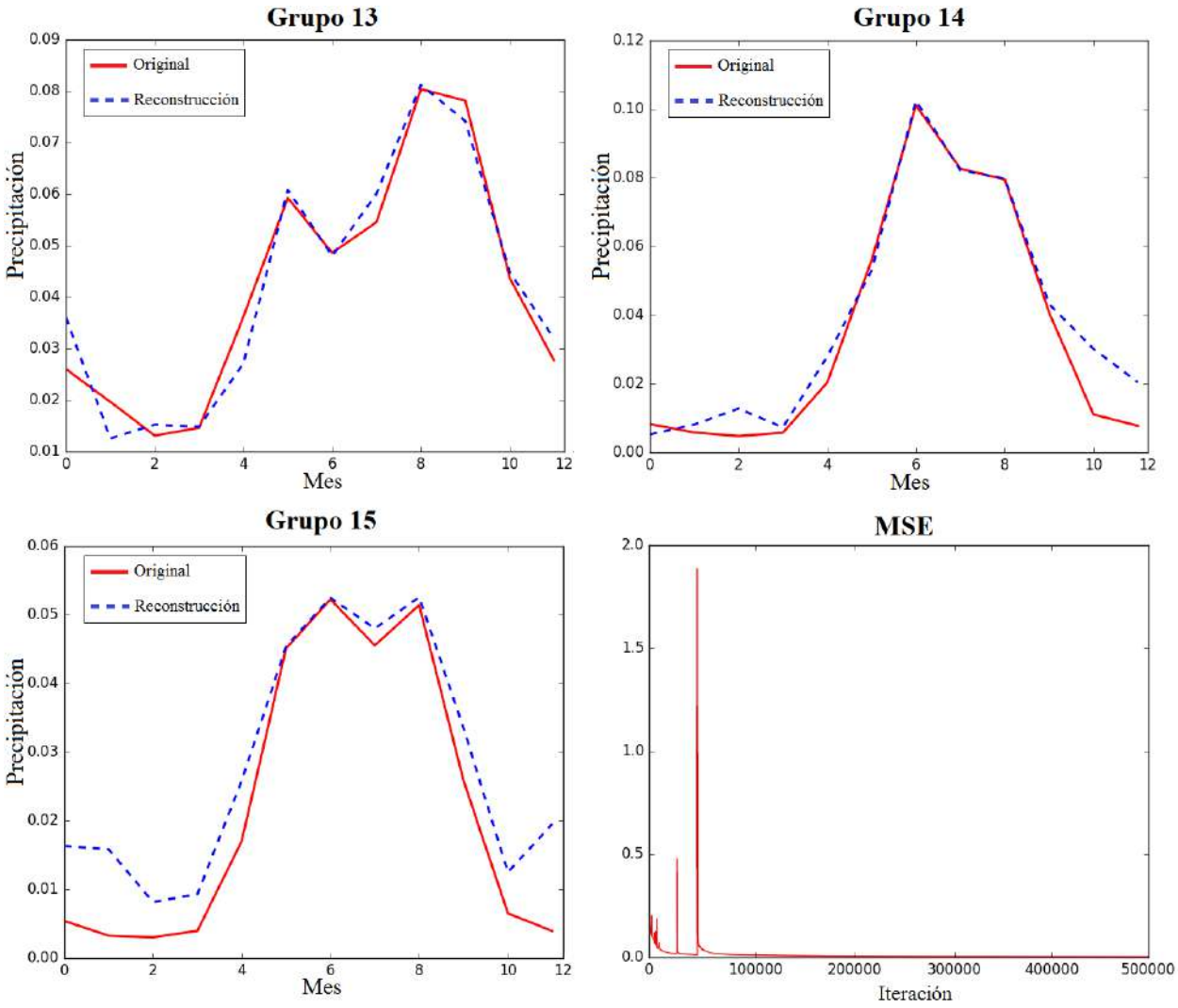


Figura 4.14: Comparación de las series mensuales de precipitación con las series reconstruidas, obtenidas por la red de retropropagación de los últimos 3 grupos del agrupamiento de las estaciones climatológicas del estado de Oaxaca en 15 grupos. También se puede observar la gráfica del comportamiento del menor MSE obtenido de todas las pruebas realizadas para dicho agrupamiento.

Una vez que se reconstruyeron las series se calculó el coeficiente de correlación de Pearson σ_{XY} y la diferencia absoluta DA_{XY}, entre la serie original con la reconstruida para determinar que tan eficiente fue el modelo propuesto. La comparación solo se realizó con las series completas del estado de Oaxaca, es decir, aquellas que no tienen ninguna información mensual perdida, las cuales son en total 1518 series.

Para la clasificación en 5 grupos la prueba con el menor MSE se obtuvo utilizando los parámetros: NV=5, n=5, m=20, función $h(x)$, α=0.25, L=500000 (ver Tabla 4.5). Para esta prueba se obtuvo un coeficiente de Pearson σ_{XY} promedio de 0.3725, pero el menor valor fue de -0.9011 el cual corresponde al año 1985 de la estación 20132 que tiene por nombre Santiago Astata (SMN) del municipio con el mismo nombre. Se observa en la Figura 4.15 (a) que cuando la serie original crece la serie reconstruida decrece, y cuando la serie original decrece la otra es creciente, debido a esto el valor de σ_{XY} es muy cercano a -1, es decir, existe una correlación negativa entre la serie original y la reconstruida. El máximo valor de σ_{XY} para este agrupamiento fue de 0.9659 correspondiente al año 1982 de la estación 20036 que tiene por nombre Santiago Progreso ubicada en el municipio San Juan Bautista Valle Nación, por lo que existe una correlación positiva entre la serie original y la reconstruida. Sin embargo, aunque la correlación es cercana a 1 es notable la diferencia entre ambas.

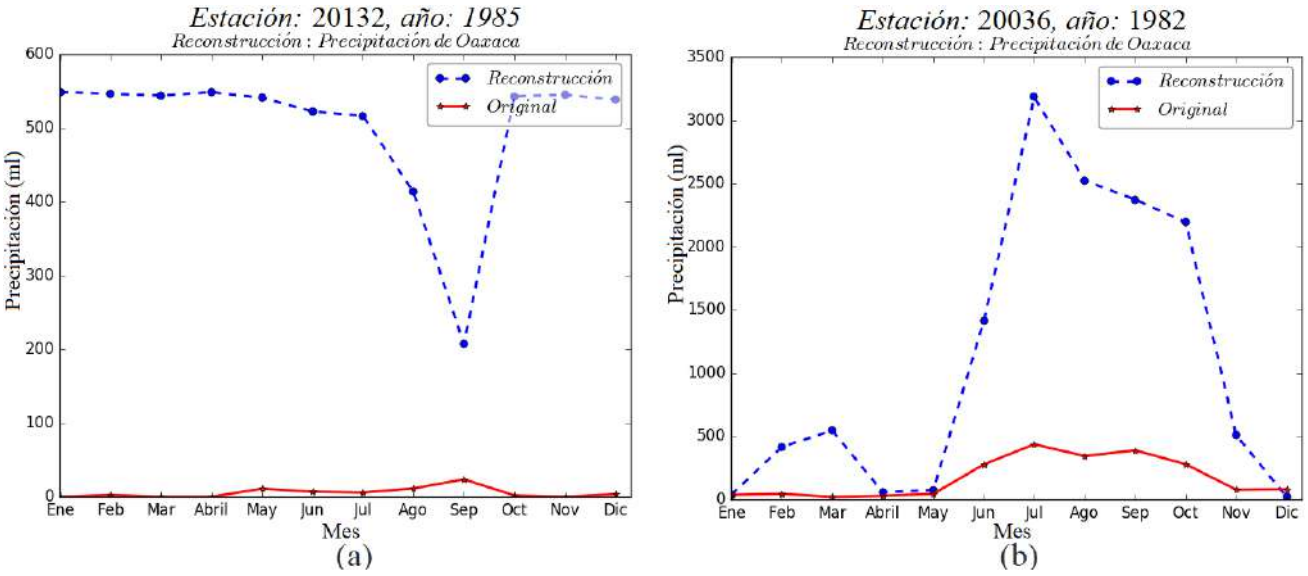


Figura 4.15: (a) son las series con σ_{XY}=-0.9011 y (b) corresponde a las series con σ_{XY}=0.9659, para la reconstrucción de series utilizando 5 grupos y 5 vecinos.

Por otro lado, considerando la precipitación normalizada, la reconstrucción con la máxima diferencia absoluta DA_{XY} fue de 5.88, esto quiere decir, que es la serie que presenta mayor diferencia con su serie original correspondiente y se refiere a 1982 de la estación 20041 que tiene por nombre Ixtlan de Juarez perteneciente al municipio con el mismo nombre. El mínimo DA_{XY} fue de 0.5699 correspondiente al año 2000 de la estación 20084 con nombre Papaloapan ubicada en el municipio San Juan Bautista Tuxtepec. La reconstrucción con el máximo y mínimo DA_{XY} se puede observa en la Figura 4.16.

De la clasificación en 10 grupos se sabe que el menor MSE obtenido fue de 0.00071,

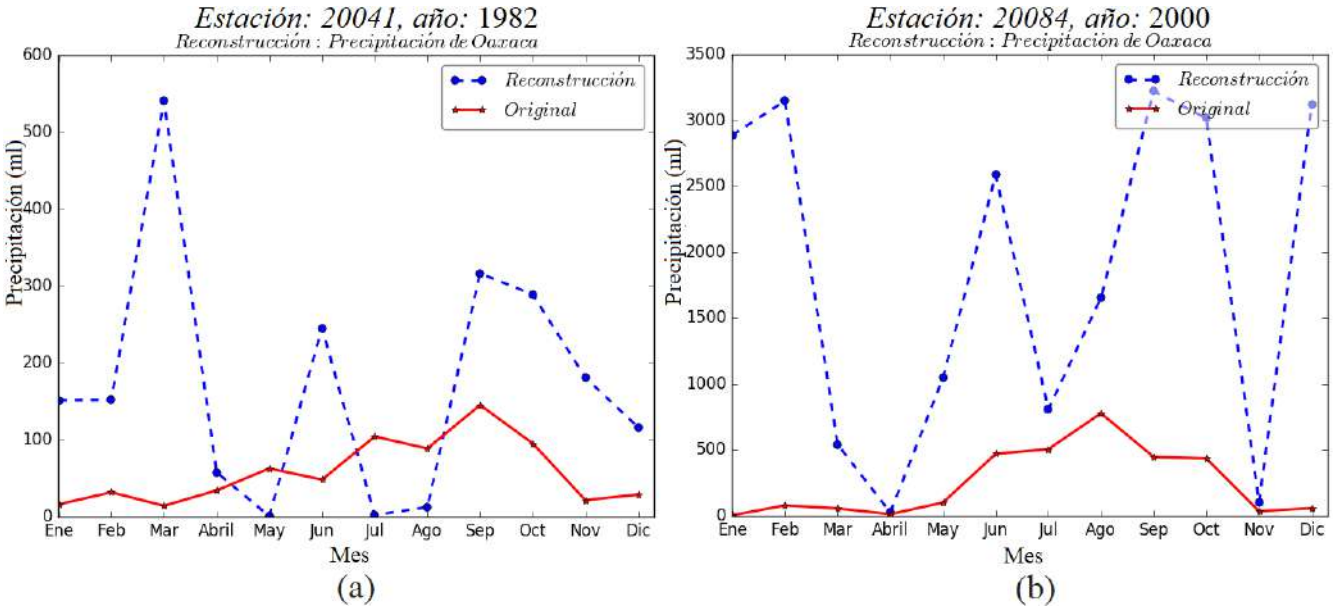


Figura 4.16: (a) representa las series con DA_{XY}=5.88 y (b) corresponde a las series con DA_{XY}=0.5699, para la reconstrucción de series utilizando 5 grupos y 5 vecinos.

los parámetros utilizados y el tiempo de ejecución se puede consultar en la Tabla 4.6. Para este caso el coeficiente de correlación mínimo fue de -0.3283 que corresponde al año 1969 de la estación 20012 que tiene por nombre "Campamento Vista Hermosa" ubicada en el municipio San Juan Comaltepec. Mientras que el máximo de σ_{XY} fue de 0.9844 correspondiente al año 1981 de la estación 20081 que tiene por nombre "San Lucas Ojitlan" ubicada en el municipio con el mismo nombre. En la Figura 4.17 se muestran las dos gráficas respectivamente.

Al calcular el valor de DA_{XY} para este agrupamiento el valor máximo fue de 3.176 que corresponde a la serie del año 1958 de la estación 20152 que tiene por nombre "Temascal" ubicada en el municipio de San Miguel Soyaltepec. La serie con menor DA_{XY}=0.0956 corresponde al año 1985 de la estación 20132 con nombre "Santiago Astata (SMN)" ubicada en el municipio que tiene el mismo nombre. En la Figura 4.18 se pueden observar las gráficas correspondientes al valor máximo (a) y mínimo (b) de DA_{XY}.

Con respecto al análisis de la clasificación en 15 grupos, se analizó que el menor MSE obtenido fue de 29.13×10^{-4}, los parámetros utilizados y el tiempo de ejecución se puede consultar en la Tabla 4.7. Para este caso el coeficiente de correlación σ_{XY} mínimo fue de -0.5166 que corresponde al año 1985 de la estación 20132 que tiene por nombre "Santiago Astata (SMN)" ubicada en el municipio que tiene el mismo nombre. Mientras que el máximo de σ_{XY} fue de 0.9936 correspondiente al año 1981 de la estación 20144 que tiene por nombre "Suchixtlahuaca-Tehuantepec" ubicada en el municipio de San Cristobal Suchixtlahuaca. En la Figura 4.19 se muestran las dos gráficas que obtuvieron el mínimo y el máximo σ_{XY} respectivamente; en la gráfica (a) se observa que las serie original y la reconstruida tienen el mismo comportamiento.

Pasando al análisis de DA_{XY} para los 15 grupos, el valor máximo fue de 1.6576 que corresponde a la serie del año 1969 de la estación 20016 que tiene por nombre "Cataluga"

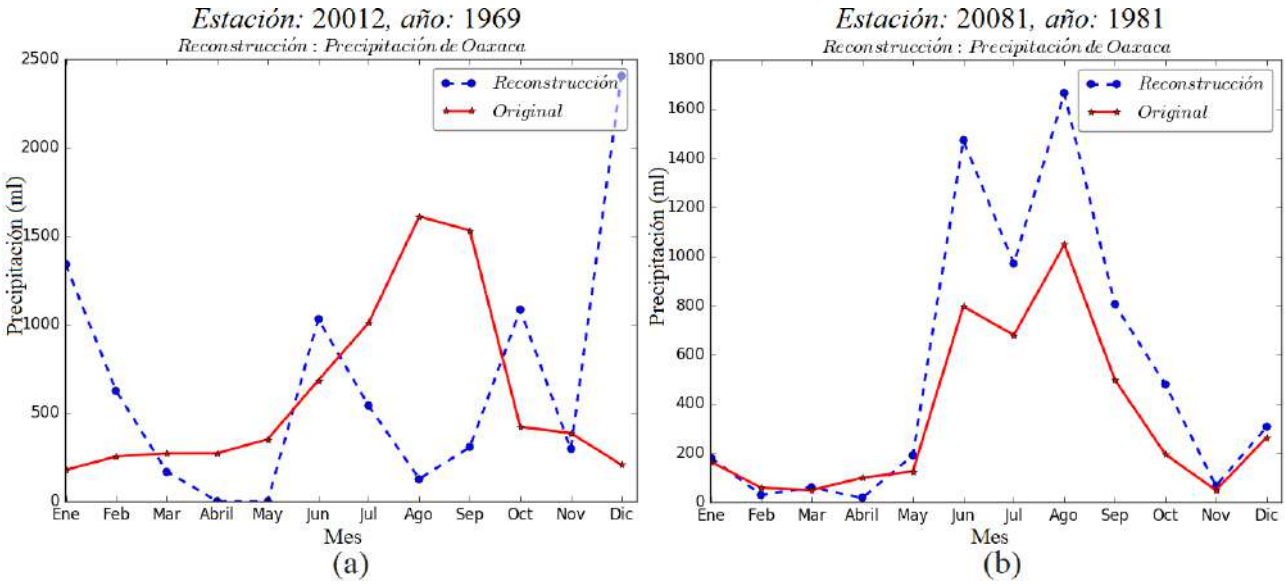


Figura 4.17: (a) corresponde a las series con σ_{XY}=-0.9011 y (b) las series con σ_{XY}=0.9659, para la reconstrucción de series utilizando 10 grupos y 5 vecinos.

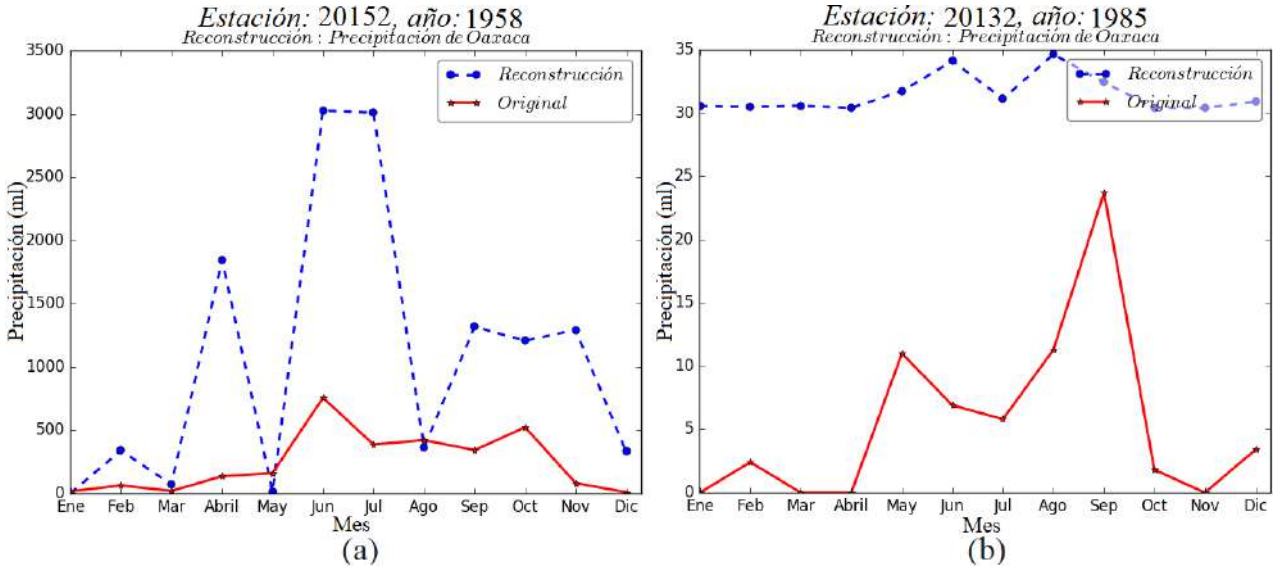


Figura 4.18: (a) corresponde a las series con DA_{XY}=3.176 y (b) a las series con DA_{XY}=0.0956, para la reconstrucción de series utilizando 10 grupos y 5 vecinos.

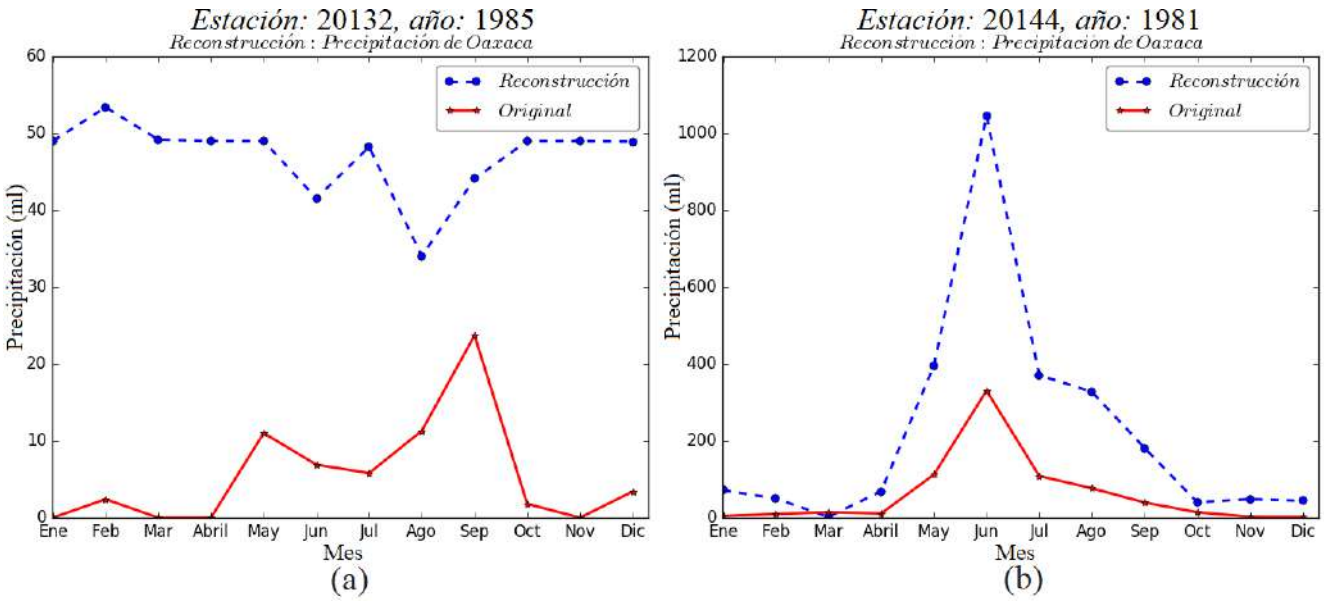


Figura 4.19: (a) corresponde a las series con σ_{XY}=-0.5166 y (b) a las series con $sigma_{XY}$=0.9936, para la reconstrucción de series utilizando 15 grupos y 5 vecinos.

ubicada en el municipio de Santa María Chilchotla. Por otro lado, la DA_{XY} mínima fue de 0.1240 que corresponde al año 1982 de la estación 20280 que tiene por nombre "Guelatao (CFE)" ubicada en el municipio de San Miguel del Río. En la Figura 4.20 se pueden observar las gráficas correspondientes al valor máximo (a) y mínimo (b) de DA_{XY}, respectivamente.

Algoritmo de contrapropagación

Para la reconstrucción de las series de precipitación del estado de Oaxaca utilizando el algoritmo de contrapropagación se consideraron dos casos: el caso de entrenamiento y el de prueba. Para el entrenamiento de la red se ingresó el mismo conjunto que se ingresa en el algoritmo de Kohonen.

En las primeras pruebas los conjuntos de pesos W y V se inicializaron de forma aleatoria con una distribución $U(0,1)$, sin embargo no se obtuvieron buenos resultados, la reconstrucción era la misma para todas las series de precipitación como se puede observar en las gráficas de la Figura 4.21, no importaba que tanto se modificaban el resto de los parámetros, siempre se obtuvieron los mismos resultados.

Debido a esto, se decidió inicializar el conjunto V de pesos igual a cero y el conjunto $W \sim U(0,1)$ o los dos conjuntos en cero. En la Tabla 4.8 se describen los parámetros de las pruebas que se realizaron en el entrenamiento de la red con la modificación de los pesos. Es importante resaltar que la pruebas se ejecutaban 10 veces cada una cuando los pesos se inicializaban de forma aleatoria. La prueba con menor MSE (51.89×10^{-4}) fue con los parámetros: m =200, α =0.01, β =0.1, Linput=1000, Lout=1000, V =0, $W \sim U(0,1)$.

Se puede resaltar que se obtuvieron mejores resultados al utilizar 100 o 200 neuronas,

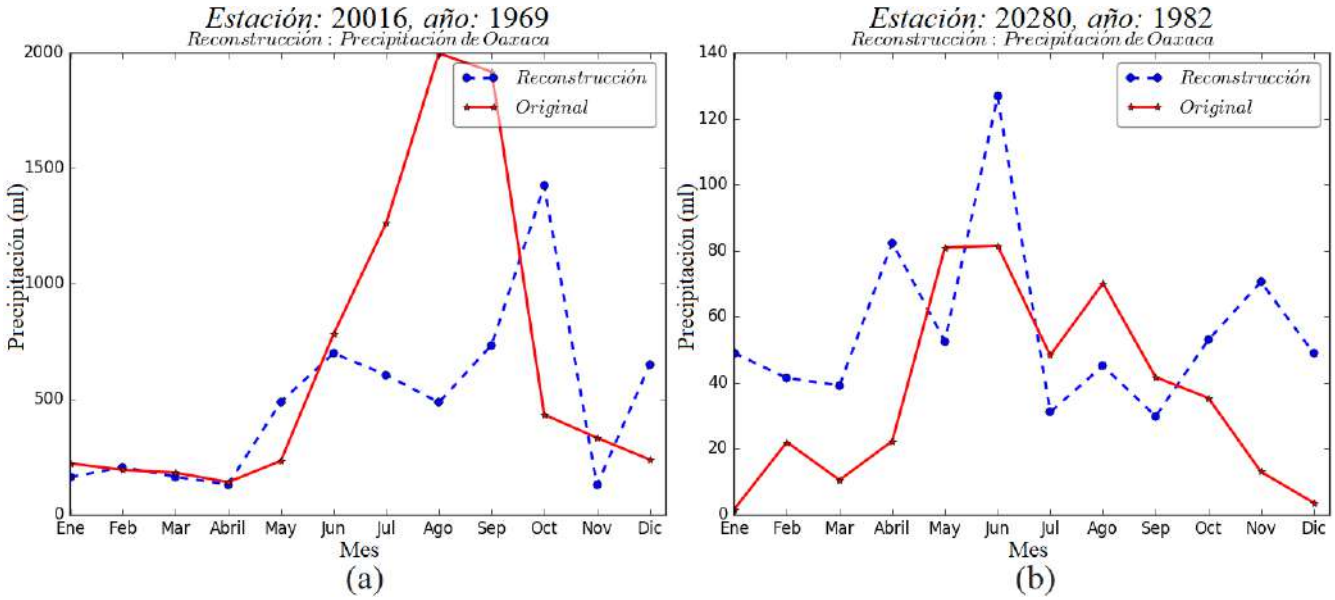


Figura 4.20: (a) corresponde a las series con DA_{XY}=1.6576 y (b) a las series con DA_{XY}=0.1240, para la reconstrucción de series utilizando 15 grupos y 5 vecinos.

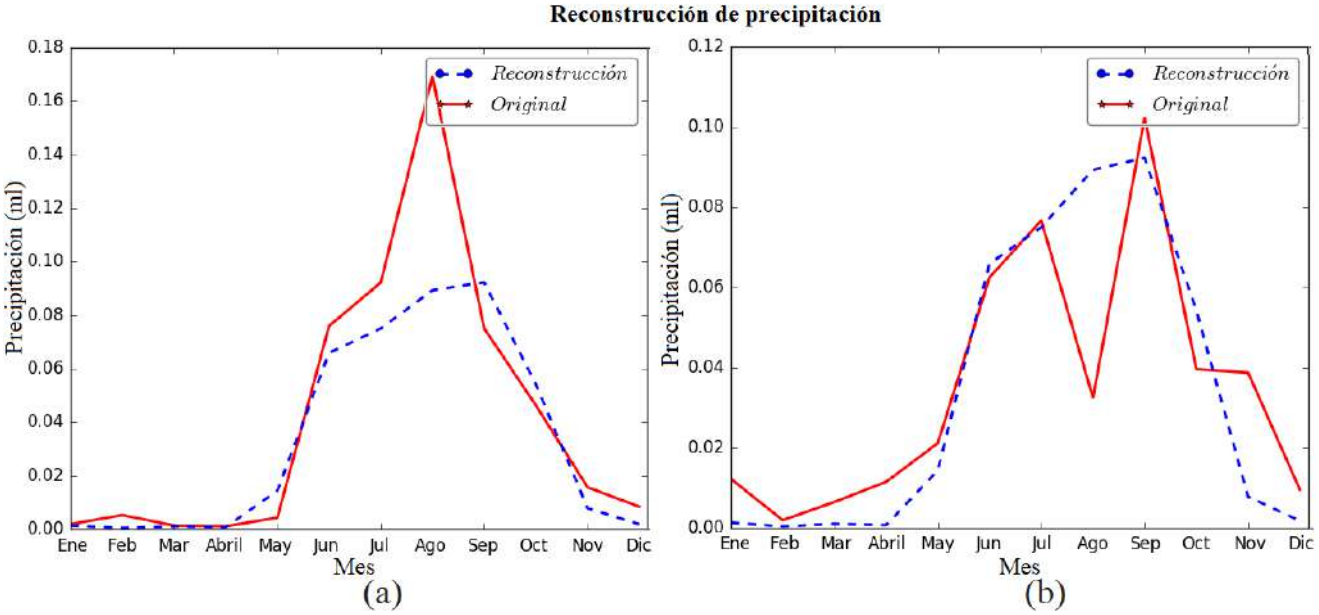


Figura 4.21: Resultados de las primeras pruebas del algoritmo de contrapropagación utilizando $V, W \sim U(0, 1)$.

o con el número de iteraciones de 1000. Sin embargo, al aumentar el número de neuronas o de iteraciones el tiempo de ejecución aumentaba, como se puede ver con la prueba que obtuvo el menor MSE, que se tardo cerca de 12 horas.

Tabla 4.8: Tabla de las pruebas realizadas con el algoritmo de contrapropagación.

m^a	α^b	β^c	Linput^d	Lout^e	V^f	W^g	MSE^h	Tiempo^i
200	0.01	0.10	1000	1000	0	[0,1]	51.89×10^{-4}	11:35:49
100	0.05	0.01	1000	1000	0	[0,1]	52.76×10^{-4}	06:40:22
150	0.01	0.10	1000	1000	0	[0,1]	53.21×10^{-4}	10:20:48
100	0.01	0.01	1000	1000	0	[0,1]	54.01×10^{-4}	07:20:21
100	0.01	0.01	1000	1000	0	0	54.01×10^{-4}	06:49:18
100	0.01	0.05	1000	1000	0	[0,1]	55.58×10^{-4}	06:09:13
100	0.01	0.05	1000	1000	0	0	55.58×10^{-4}	05:53:13
100	0.05	0.10	1000	1000	0	[0,1]	56.13×10^{-4}	07:27:18
100	0.01	0.10	1000	1000	0	[0,1]	57.97×10^{-4}	05:42:14
50	0.05	0.10	1000	1000	0	[0,1]	70.54×10^{-4}	03:25:53
50	0.01	0.10	1000	1000	0	[0,1]	70.96×10^{-4}	03:28:11
100	0.50	0.80	1000	1000	0	0	91.28×10^{-4}	06:35:03
100	0.20	0.80	1000	1000	0	0	91.48×10^{-4}	07:02:37
15	0.01	0.10	1000	1000	0	[0,1]	99.94×10^{-4}	01:07:36
200	0.10	0.10	250	500	0	[0,1]	287.71×10^{-4}	05:54:55

[a] m: Número de neuronas.
[b] α: coeficiente de aprendizaje de la capa oculta.
[c] β: coeficiente de aprendizaje de la capa de salida.
[d] Linput: número de iteraciones en la capa oculta.
[e] Lout: número de iteraciones en la capa de salida.
[f] V: valor del conjunto de pesos en la capa oculta.
[g] W: valor del conjunto de pesos en la capa de salida.
[h] MSE: error cuadrático medio.
[i] Tiempo de ejecución.

Para la prueba con menor MSE del conjunto de entrenamiento, se escogió a la serie de precipitación que obtuvo el máximo y el mínimo MSE, es decir aquella serie que su aproximación fue la peor y aquella que obtuvo la mejor aproximación. El máximo MSE fue de 0.2026 y lo obtuvo la serie de precipitación del año 1972 de la estación 30022 con nombre "Catemaco" ubicada en el municipio de Catemaco del estado de Veracruz (Figura 4.22 (a)). La serie con el mínimo MSE fue de 6.65×10^{-5} y corresponde al año 1964 de la estación 21083 con nombre "Tehuacan" ubicada en el municipio con el mismo nombre perteneciente al estado de Puebla (Figura 4.22 (b)).

Al tener la prueba con el menor MSE se utilizan los parámetros de esta prueba para ingresar el conjunto de prueba. El conjunto de prueba P está constituido por las series de precipitación anual del estado de Oaxaca al igual que el conjunto de prueba de la red de retropropagación. Una vez que se hizo la reconstrucción se calculó el coeficiente de correlación de Pearson σ_{XY} y la diferencia absoluta DA_{XY} entre las series originales con su reconstrucción obtenida. El mínimo valor de σ_{XY} fue de -0.6771 que corresponde al año 1979 de la estación 20154 con nombre "Santa María Teopoxco" ubicada en el municipio

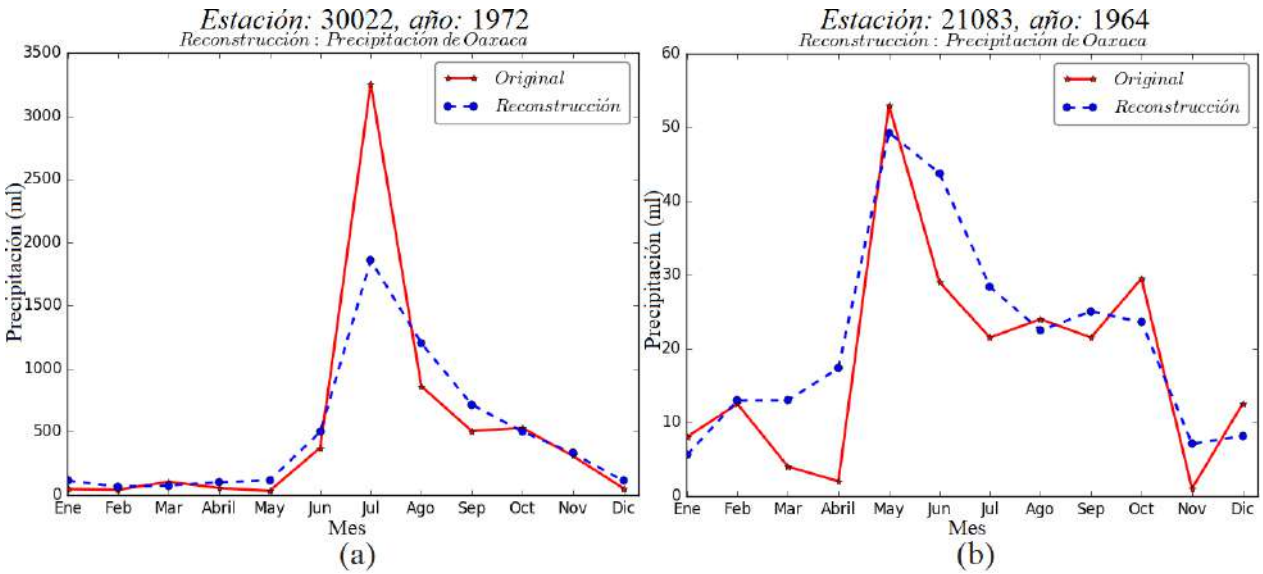


Figura 4.22: (a) Peor y (b) mejor reconstrucción de la prueba que obtuvo el menor MSE=51.89×10^{-4}, del conjunto de entrenamiento.

con el mismo nombre. En la Figura 4.23 (a) se puede observar claramente que hay una relación negativa entre las dos series. El máximo σ_{XY} es de 0.997 que corresponde al año 1988 de la estación 20303 con nombre "Tonameca" ubicada en el municipio de Santa María Tonameca. En la Figura 4.23 (b) se puede observar que la serie reconstruida sigue el mismo comportamiento de la serie original.

Por otro lado, la diferencia absoluta máxima fue de DA_{XY}=1.2491 obtenida de la serie de precipitación del año 1992 de la estación 20143 con nombre "Suchixtlahuaca-Quiotep-" ubicada en el municipio de San Cristobal Suchixtlahuaca. En la gráfica de la Figura 4.24 (a) se puede ver una gran diferencia entre la serie reconstruida y la serie original, por el contrario en la gráfica de la Figura 4.24 (b) se observa una similitud entre las series con el mínimo DA_{XY}. Este valor mínimo corresponde a la serie de precipitación del año 1998 de la estación 20364 con nombre "Huitzo" ubicada en el municipio de San Pablo Huitzo.

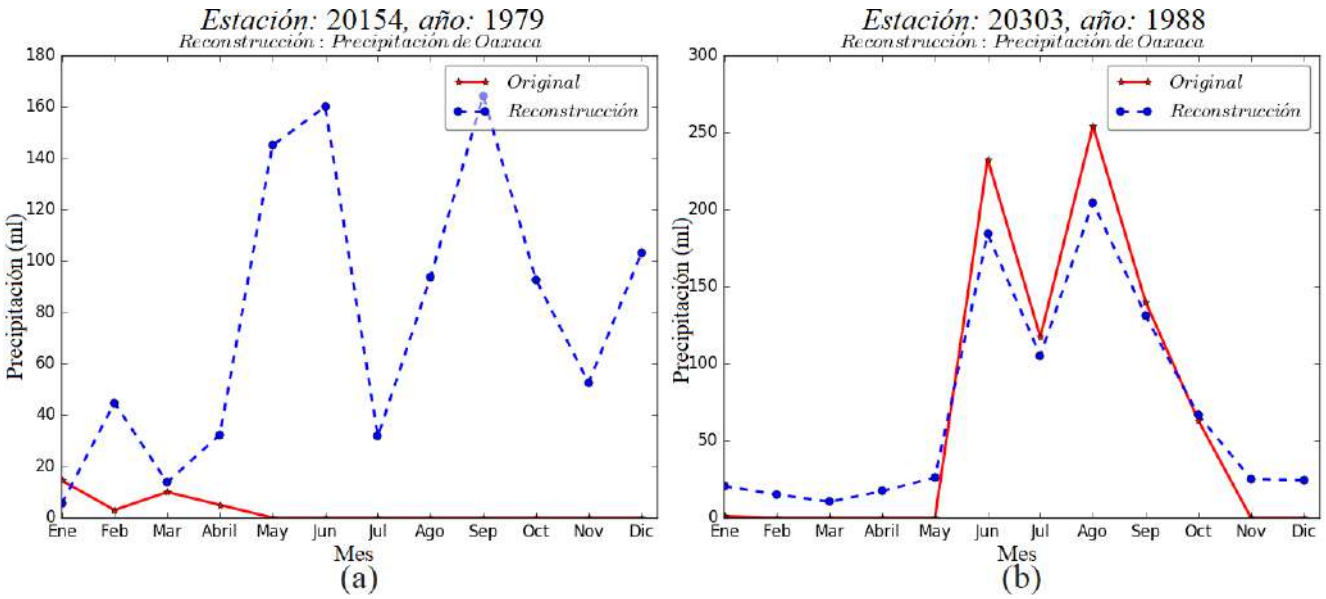


Figura 4.23: Series de precipitación con σ_{XY}=-0.6771 (a) y σ_{XY}=0.997 (b) de la reconstrucción utilizando el algoritmo de contrapropagación y los parámetros de la prueba con MSE=51.89×10^{-4}.

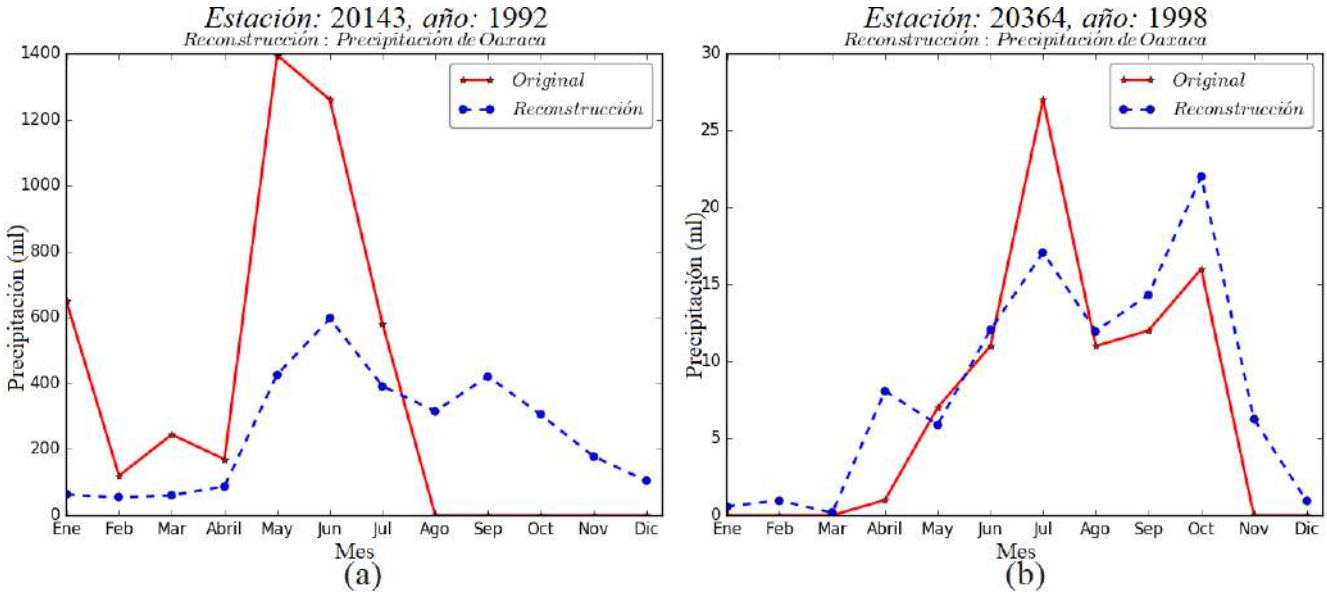


Figura 4.24: Series de precipitación con DA_{XY}=1.24911 (a) y DA_{XY}=1.2491 (b) de la reconstrucción utilizando el algoritmo de contrapropagación y los parámetros de la prueba con MSE=51.89×10^{-4}.

Conclusiones

El objetivo principal de este trabajo se centró en realizar la reconstrucción de las series de precipitación con información mensual obtenida de las estaciones climatológicas del estado de Oaxaca.

Del análisis estadístico para la elección del periodo de registros utilizados en la reconstrucción de las series de precipitación, se determinó que el intervalo de años adecuado es [1957,2007]. Con la prueba de Grubbs, se detectaron datos atípicos con un nivel se significancia de α=0.01, descartando en total el 16.31 % de los datos de precipitación mensual (16.21 % quedando por debajo del límite inferior y 0.1 % por arriba del límite superior).

En el análisis de componentes principales se incluyeron las variables: temperatura observada, temperatura mínima, temperatura máxima, evaporación, precipitación, distancias entre las estaciones climatológicas al Golfo de México y al Océano Pacífico. Con los resultados obtenidos, se determinó que ninguna de estas variables está correlacionada significativamente con la precipitación, por lo que se decidió realizar la reconstrucción con solo esta variable.

La regionalización de estaciones en 5, 10 y 15 grupos, fue la base para reconstruir las series de precipitación del estado de Oaxaca mediante el algoritmo de retropropagación, debido a que las estaciones climatológicas vecinas de cada estación, se eligieron de dichos grupos. De este análisis se determina que, para obtener un adecuado agrupamiento es necesario que los pesos del algoritmo se inicialicen en cero.

Se generaron seis casos de entrenamiento para la red de retropropagación, que eran de los tres agrupamientos generados de la red de Kohonen por dos tipos de vecindades utilizados: 3 y 5 vecinos. El conjunto de entrada del entrenamiento de la red se conformó de los representantes de cada uno de los grupos. De las pruebas realizadas se obtuvieron mejores resultados utilizando 5 estaciones climatológicas vecinas y utilizando la función Gaussiana.

Como resultado de las pruebas de reconstrucción con la red de contrapropagación, se observó que se obtienen buenos resultados al inicializar en cero al conjunto de pesos de la capa oculta y utilizando 200 neuronas. Sin embargo, la desventaja del algoritmo es el tiempo de ejecución utilizado.

Es importante destacar que no hay estudios realizados con respecto a la reconstrucción de series de precipitación del estado de Oaxaca utilizando redes neuronales artificiales, en particular las dos redes utilizadas en este trabajo.

Como trabajo futuro queda por realizar más pruebas, por ejemplo, con la red de contrapropagación utilizar más vecinos, más neuronas, menos iteraciones, u otros tipos de funciones de activación.

Apéndice A

Manual de usuario del software desarrollado

El software utilizado en este trabajo consta de cuatro programas los cuales se realizaron bajo el lenguaje de programación Python 2.7, con solo contar con dicho lenguaje es suficiente para ejecutar los programas desde la terminal con el comando: *python nombre_programa.py*. Si se cuenta con el sistema operativo Windows, se puede descargar e instalar WinPython 2.7 el cual se obtiene de `https://winpython.github.io/` y que incluye a la interfaz *Spyder*. El motivo de dividir los programas se debe a que por contar con una base de datos grande hay algunos algoritmos que tardan varios minutos o hasta horas.

El primer programa tiene por nombre *main_PCA.py* el cual al ejecutar despliega el siguiente menú:

```
Elegir la opción deseada:
  1. Generar datos de lluvia mensual.
  2. Datos atípicos.
  3. Distancias de las estaciones meotorológicas.
  4. PCA
Opción:
```

La primera opción permite obtener la base de datos de precipitación mensual del estado de Oaxaca y sus estados vecinos el cual los almacena en el archivo que tiene por nombre *Lluvia_mensual_No_normalizada*. La segunda opción permite obtener los datos atípicos de la base de datos aplicando el método de Grubss y los almacena en el archivo de texto *Lluvia_mensual_No_Atipicos_norma_porcentaje* donde "porcentaje" es el valor del porcentaje de confianza. Dentro de esta opción se muestra otro menú de opciones como se puede observar en la Figura A.1, que permite elegir el porcentaje de confianza que requiere el método de grubbs, los porcentajes disponibles son 1 %, 5 % y 10 %.

La tercera opción permite obtener la distancia de cada estación meteorológica a la frontera con el Golfo de México y el Océano Pacífico y los resultados los almacena en el archivo de texto *DistanciaMar.txt*. Para esto es importante contar con las coordenadas geográficas de diferentes puntos en dichas fronteras, los cuales se obtuvieron utilizando

```
Elegir la opción deseada:
  1. Generar datos de lluvia mensual.
  2. Datos atípicos.
  3. Distancias de las estaciones meotorológicas.
  4. PCA
Opción: 2
------------------------------
Método de Grubss:
   Elegir el porcentaje de confianza:
            1 porciento
            5 porciento
            10 porciento
Porcentaje:
```

Figura A.1: Menú para obtener los datos atípicos de la base de datos.

el programa ArcGis y que tienen por nombre *Golfo.csv* y *Pacifico.csv*. La última opción muestra el análisis de componentes principales aplicados a las variables climatológicas cuantitativas. Al elegir esta opción se despliega el siguiente menú:

```
Elegir la opción deseada:
  1. Generar datos de lluvia mensual.
  2. Datos atípicos.
  3. Distancias de las estaciones meotorológicas.
  4. PCA
Opción: 4
------------------------------
Elegir la opción deseada:
           1. Generar datos para el PCA
           2. Obtener PCA
Opción:
```

La primera opción permite generar la base de datos que es ingresado al análisis de componentes principales y la segunda opción permite obtener dicho análisis.

El segundo programa ejecuta el algoritmo de Kohonen que permite obtener el agrupamiento de las estaciones climatológicas y que tiene por nombre *main_SOM.py*. Al ejecutar el programa se muestra el siguiente menú:

```
Elegir la opción deseada:
  1. Generar datos de entrenamiento.
  2. Agrupar conjunto de entrenamiento.
  3. Generar conjunto de prueba.
  4. Agrupar conjunto de prueba
Opción:
```

La opción 1 genera el conjunto de entrenamiento que es aquel que contiene las series de precipitación anual con 5 o menos días faltantes, considerando a las series del estado de Oaxaca y sus estados vecinos. La segunda opción agrupa a dichas series, para esto, al seleccionar la opción 2 se despliega una serie de comandos donde se pide ingresar los parámetros necesarios en el algoritmo de Kohonen como se muestra en la Figura A.2. Los

parámetros a ingresar son: número de neuronas, coeficiente de aprendizaje α, el número de presentaciones o iteraciones, por último se decide si los pesos serán fijos (1) o aleatorios (2), en caso de ser fijos, el valor real ingresado será el valor inicial de todos los pesos.

```
Elegir la opción deseada:
  1. Generar datos de entrenamiento.
  2. Agrupar conjunto de entrenamiento.
  3. Generar conjunto de prueba.
  4. Agrupar conjunto de prueba
Opción: 2
Ingresar los parámetros:

          Ingrese el número de neuronas: 5
          Ingrese el coeficiente de aprendizaje: 0.01
          Ingrese el número de presentaciones: 1000
          Valores de pesos:
             1. fijos
             2. Aleatorios

Opción pesos: 1
Valor fijo de los pesos: 0
```

Figura A.2: Menú del programa *main_SOM.py* que muestra como se deben ingresar los parámetros al elegir la opción 2.

Si en el menú principal se elige la opción 3, entonces se genera la base de datos para el conjunto de prueba, en este caso, solo se consideran a las series de todas las estaciones climatológicas del estado de Oaxaca. Por último, al elegir la opción 4 se agrupan dichas series y para eso se deben ingresar los mismos parámetros que se ingresaron al elegir la opción 2 (ver Figura A.3).

```
Elegir la opción deseada:
  1. Generar datos de entrenamiento.
  2. Agrupar conjunto de entrenamiento.
  3. Generar conjunto de prueba.
  4. Agrupar conjunto de prueba
Opción: 4
Ingresar los valores utilizados en el entrenamiento:

          Ingrese el número de neuronas: 5
          Ingrese el coeficiente de aprendizaje: 0.01
          Ingrese el número de presentaciones: 1000
          Valores de pesos:
             1. fijos
             2. Aleatorios

Opción pesos: 1
Valor fijo de los pesos: 0
```

Figura A.3: Menú del programa *main_SOM.py* que muestra como se deben ingresar los parámetros al elegir la opción 4.

En el tercer programa tiene por nombre *main_Retro.py* y es el algoritmo de retropropagación que realiza la reconstrucción de las series de precipitación del estado de Oaxaca.

Al ejecutar el programa se muestra el siguiente menú:

```
Elegir la opción deseada:
  1. Generar datos de entrenamiento.
  2. Reconstruir conjunto de entrenamiento.
  3. Generar conjunto de prueba.
  4. Reconstruir conjunto de prueba
Opción:
```

La opción 1 genera la base de datos para el entrenamiento de la red, para eso se debe contar con el agrupamiento obtenido del conjunto de entrenamiento al ejecutar el programa *main_ SOM.py* (opción 1 y 2), ya que dicha base la conforman los representantes de cada grupo. La opción 2 del algoritmo de retropropagación realiza la reconstrucción del conjunto de entrenamiento por lo que se deben ingresar los parámetros: número de neuronas en la capa oculta, coeficiente de aprendizaje, número de iteraciones, se elige la función de activación de tres opciones, el número de grupos utilizados en la red de Kohonen y el número de vecinos (ver Figura A.4).

```
Elegir la opción deseada:
  1. Generar datos de entrenamiento.
  2. Reconstruir conjunto de entrenamiento.
  3. Generar conjunto de prueba.
  4. Reconstruir conjunto de prueba
Opción: 2
          Ingresar:

           Número de neuronas en la capa oculta:20
           Coeficiente de aprendizaje: 0.1
           Número de iteraciones: 5000
           Elegir función de activación:
            1. Sigmoide
            2. Tangh(z/2)
            3. Gaussina
            opc: 3
           Número de grupos:5
           Número de vecinos:5
```

Figura A.4: Menú del programa *main_ Retro.py* que muestra como se deben ingresar los parámetros al elegir la opción 2.

La opción 3 genera el conjunto de prueba cuyos elementos son las series de precipitación anual de las estaciones climatológicas del estado de Oaxaca. Por último, la opción 4 obtiene la reconstrucción de este conjunto y para eso se deben ingresar los parámetros ingresados al elegir la opción 2 como se puede observar en la Figura A.5.

El último programa ejecuta la reconstrucción de las series de precipitación del estado de Oaxaca utilizando el algoritmo de contrapropagación y tiene por nombre *main_ Contra.py*. Al ejecutar el programa se genera el menú que se muestra en la Figura A.6 desplegando cuatro opciones. La opción genera la base de datos para el conjunto de entrenamiento que en realizad es el mismo conjunto que genera el algoritmo de Kohonen. Al seleccionar la opción 2, se despliegan otros comandos que piden ingresar los parámetros necesarios para realizar la reconstrucción de las series, dado que se tienen dos capas cada una de ellas

```
Elegir la opción deseada:
  1. Generar datos de entrenamiento.
  2. Reconstruir conjunto de entrenamiento.
  3. Generar conjunto de prueba.
  4. Reconstruir conjunto de prueba
Opción: 4
          Ingresar:

           Número de neuronas en la capa oculta:20
           Coeficiente de aprendizaje: 0.1
           Número de iteraciones: 5000
           Elegir función de activación:
            1. Sigmoide
            2. Tangh(z/2)
            3. Gaussina
            opc: 3
           Número de grupos:5
           Número de vecinos:5
```

Figura A.5: Menú del programa *main_ Retro.py* que muestra como se deben ingresar los parámetros al elegir la opción 4.

tienen parámetros distintos, es por ello que se ingresan primero los parámetros de la capa intermedia y después los de la capa de salida (ver Figura A.7).

```
Elegir la opción deseada:
  1. Generar datos de entrenamiento.
  2. Reconstruir conjunto de entrenamiento.
  3. Generar conjunto de prueba.
  4. Reconstruir conjunto de prueba
Opción:
```

Figura A.6: Menú del programa *main_ Contra.py*.

La opción 3 genera el conjunto de entrenamiento que constituyen las series de precipitación del estado de Oaxaca, y la opción 4 realiza la reconstrucción de las series por lo que se deben ingresar los parámetros que se consideraron al realizar el entrenamiento de la red.

```
Elegir la opción deseada:
  1. Generar datos de entrenamiento.
  2. Reconstruir conjunto de entrenamiento.
  3. Generar conjunto de prueba.
  4. Reconstruir conjunto de prueba
Opción: 2
          Ingresar elementos en la capa oculta:

           Número de neuronas:50
           Coeficiente de aprendizaje : 0.1
           Número de iteraciones: 1000
           Valores de pesos:
             1. fijos
             2. Aleatorios

           Opción pesos: 1
            Valor fijo de los pesos: 0
          Ingresar elementos en la capa de salida:

           Coeficiente de aprendizaje : 0.25
           Número de iteraciones: 1000
           Valores de pesos:
             1. fijos
             2. Aleatorios

           Opción pesos: 1
            Valor fijo de los pesos: 0
```

Figura A.7: Menú del programa *main_ Contra.py* al elegir la opción 2, se deben ingresar los parámetros en la capa intermedia y de salida de la red.

```
Elegir la opción deseada:
  1. Generar datos de entrenamiento.
  2. Reconstruir conjunto de entrenamiento.
  3. Generar conjunto de prueba.
  4. Reconstruir conjunto de prueba
Opción: 4
Ingresar los valores utilizados en el entrenamiento:

          Ingresar elementos en la capa oculta:

           Número de neuronas:50
           Coeficiente de aprendizaje : 0.1
           Número de iteraciones: 1000
           Valores de pesos:
             1. fijos
             2. Aleatorios

           Opción pesos: 1
            Valor fijo de los pesos: 0
          Ingresar elementos en la capa de salida:

           Coeficiente de aprendizaje : 0.25
           Número de iteraciones: 1000
          Valores de pesos:
             1. fijos
             2. Aleatorios

           Opción pesos: 1
            Valor fijo de los pesos: 0
```

Figura A.8: Menú del programa *main_ Contra.py* al elegir la opción 4, se deben ingresar los parámetros en la capa intermedia y de salida de la red.

Bibliografía

Acock, M. C. y Pachepsky, Ya A., (2000). *Estimating Missing Weather Data for Agricultural Simulations Using Group Method of Data Handling.* Journal of Applied Meteorology, 39(7):1176-1184.

Almeida Román, M. d. l. P., (2010). *Instructivos de procesamiento de información hidrometeorológica.*

Alvarado Medellin, P. (2007). *Modelación estocástica de los escurrimientos de la cuenca del río Amajac, Hidalgo, México.* PhD thesis, Institución de enseñanza e investigación en ciencias agrícolas, Texcoco, edo. de México.

Álvarez-Olguín, G. y Escalante-Sandoval, C. (2016). *Análisis de frecuencias no estacionario de series de lluvia anual.* Tecnología y Ciencias del Agua, VII(1):71-88.

Álvarez-Olguín, G. y Escalante-Sandoval, C. (2017). *Modes of Variability of Annual and Seasonal Rainfall in Mexico.* JAWRA Journal of the American Water Resources Association, 53(1):144-157.

Antelo, M. R y Long, M. E. F. (2014). *Estimación de datos faltantes de precipitación diaria para las distintas ecorregiones de la República Argentina.* In editor, editor, 2do. Encuentro de Investigadores en Formación en Recursos Hídricos: resúmenes de trabajos, page 58.

Barnett, V. y Lewis, T. (1984). *Outliers in statistical data.* Biometrical Journal, 30(7).

Beckman, R. J. y Cook, R. D. (1983). *Outlier. . . . . s.* Technometrics, 25(2):119-149.

Bustamante, A. M. (2013). *Series de tiempo: Una aplicación a registros hidrométricos en una cuenca del estado de Oaxaca.* Universidad Tecnológica de la Mixteca, Huajuapan de León, Oaxaca.

Chandola, V., Banerjee, A. y Kumar, V. (2007). *Outlier detection: A survey.* ACM Computing Surveys.

Coulibaly, P. y Evora, N. (2007). *Comparison of neural network methods for infilling missing daily weather records.* Journal of Hydrology, 341(1-2)27-41.

Creutin, J., Andrieu, H. y Faure, D. (1997). *Use of a weather radar for the hydrology of a mountainous area. part II: radar measurement validation.* Journal of Hydrology, 193(1-4):26-44.

Devi, S. R., Arulmozhivarman, P., Venkatesh, C. y Agarwal, P. (2016). *Performance comparison of artificial neural network models for daily rainfall prediction.* International Journal of Automation and Computing, 13(5):417-427.

Faucher, M., Burrows, W. R., y Pandolfo, L. (1999). *Empirical-statistical reconstruction of surface marine winds along the western coast of Canada.* Climate Research, 11(3):173-190.

Filippini, F., Galliani, G., y Pomi, L. (1970). *The estimation of missing meteorological data in a network of automatic stations.* WIT Transactions on Ecology and the Environment, 4.

Freeman, J. y Skapura, D. (1991). *Neural Networks: Algorithms, Applications, and Programming Techniques.* Addison-Wesley Computation and Neural Systems Series. Addison-Wesley.

Grubbs, F. E. (1969). *Procedures for detecting outlying observations in samples.* Technometrics, 11(1):1-21.

Grubbs, F. E. y Beck, G. (1972). *Extension of sample sizes and percentage points for signicance tests of outlying observations.* Technometrics, 14(4):847-854.

Hawkins, D. M. (1980). *Identification of outliers.* Volume 11. Springer.

James, G., Witten, D., Hastie, T. y Tibshirani, R. (2014). *An Introduction to Statistical Learning: with Applications in R.* Springer Texts in Statistics. Springer New York.

Jeffrey, S. J., Carter, J. O., Moodie, K. B. y Beswick, A. R. (2001). *Using spatial interpolation to construct a comprehensive archive of australian climate data.* Environmental Modelling & Software, 16(4):309-330.

Khalil, M., Panu, U. y Lennox, W. (2001). *Groups and neural networks based streamflow data infilling procedures.* Journal of Hydrology, 241(3):153-176.

Kim, J.-W. y Pachepsky, Y. A. (2010). *Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamow simulation.* Journal of Hydrology, 394(3-4):305-314.

Kundzewicz, Z., Robson, A., Data, W. C. y Programme, M. (2000). *Detecting trend and other changes in hydrological data.* Geneva: Secretariat of the World Meteorological Organization.

Kuzmanovski, I. y Novič, M. (2008). *Counter-propagation neural networks in matlab.* Chemometrics and Intelligent Laboratory Systems, 90(1):84-91.

Linacre, E. (1992). *Climate Data and Resources: A Reference and Guide.* Routledge.

Lowry, W. (1972). *Compendium of lecture notes in climatology class IV meteorological personnel.* WMO (Series). Secretariat of the WMO, Geneva, first edition.

Lucio, P. S., Conde, F. C., Cavalcanti, I. F. A., Serrano, A., Ramos, A. M. y

Cardoso, A. O. (2007). *Spatiotemporal monthly rainfall reconstruction via artificial neural network - case study: south of Brazil.* In Advances in Geosciences, páginas 67-76.

Luk, K., Ball, J. y Sharma, A. (2000). *A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting.* Journal of Hydrology, 227(1-4):56-65.

Peat, J. y Barton, B. (2005). *Medical statistics: A guide to data analysis and critical appraisal.* John Wiley & Sons.

Pérez Aguila, R. (2005). *Una introducción al cómputo neuronal artificial.* El Cid Editor.

Pérez, F. E. L. (2012). *Técnicas de agrupamiento para la regionalización de caudales en la Mixteca Oaxaqueña.* Universidad Tecnológica de la Mixteca, Huajuapan de León, Oaxaca.

Ramos-Calzado, P., Gómez-Camacho, J., Pérez-Bernal, F. y Pita-López, M. (2008). *A novel approach to precipitation series completion in climatological datasets: application to Andalusia.* International Journal of Climatology, 28(11):1525-1534.

Ruíz Corral, J. A., Díaz Padilla, G., García Nieto, H., Silva Serna, M. M. y Medina García, G. (2007). *Estadísticas climatológicas básicas del estado de Guanajuato (1961-2003)*

Ruíz Corral, J. A., Medina García, G., Rigoberto, M. S., Díaz Padilla, G. y Víctor, S. A. (2006). *Estadísticas climatológicas básicas del estado de Baja California Sur (período 1961-2003).*

Serrano Altamirano, V., Silva Serna, M. M., Cano García, M. n., Medina García, G. y Ruíz Corral, J. A. (2005). *Estadísticas climatológicas básicas del estado de Oaxaca (período 1961-2003).*

Solís, K. J. B. y Rivera, K. I. P. (2004). *Revisión de metodologías para extensión y relleno de datos en series históricas, y su aplicación a los ríos de el Salvador.*

Teegavarapu, R. S. y Chandramouli, V. (2005). *Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records.* Journal of Hydrology, 312(1):191-206.

Uribe, I. A. (2010). *Guía metodológica para la selección de técnicas de depuración de datos.*

Willmott, C. J., Robeson, S. M. y Feddema, J. J. (1994). *Estimating continental and terrestrial precipitation averages from rain-gauge networks.* International Journal of Climatology, 14(4):403-414.

Xia, Y., Fabian, P., Stohl, A. y Winterhalter, M. (1999). *Forest climatology: estimation of missing values for Bavaria, Germany.* Agricultural and Forest Meteorology, 96(1):131-144.

Young, K. C., (1992). *A three-way model for interpolating for monthly precipitation values.* Monthly Weather Review, 120(11):2561-2569.