UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

FACULTAD DE CONTADURÍA, ADMINISTRACIÓN E INFORMÁTICA
MAESTRÍA EN OPTIMIZACIÓN Y CÓMPUTO APLICADO

Algoritmo Jerárquico Aglomerativo Paralelo para la Agrupación (Clustering).

T E S I S

QUE PARA OBTENER EL GRADO DE
MAESTRÍA EN OPTIMIZACIÓN Y CÓMPUTO APLICADO

PRESENTA

RICARDO MONJE LÓPEZ

DIRECTOR DE TESIS
DR. JOSÉ CRISPÍN ZAVALA DÍAZ

CO-DIRECTOR
DRA. NELVA NELY ALMANZA ORTEGA

REVISORES:
DR. JOSÉ ALBERTO HERNÁNDEZ AGUILAR
DR. MARTÍN HERIBERTO CRUZ ROSALES
DR. JOAQUÍN PÉREZ ORTEGA

CUERNAVACA, MORELOS ABRIL, 2021

Cuernavaca, Morelos a 15 de enero del 2021.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN Y POSGRADO DE LA FCAeI
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Computo Aplicado, del estudiante Monje López Ricardo, con matricula 10023095, con el título **Algoritmo Jerárquico Aglomerativo Paralelo para la Agrupación (Clustering).** Por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación y Posgrado tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dr. José Crispín Zavala Díaz
Profesor- investigador
Facultad de Contaduría Administración e Informática

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

JOSE CRISPIN ZAVALA DIAZ | Fecha:2021-01-27 18:56:37 | Firmante

dCW03xrIb3deOx+wezoqIsjdeMGLZ4a5RvEtwaboAQAYinUdG9bhKoetZT9tCsKQMMEchM5pDpgJ/WmpRJCPpgIpfV4vaCeWzZCdhveAExoD5Mi0Hb9aUdVFxvPoaGgka0WCL
EefZGzNhon0kc47K6koDtHytO6NecnWhhUeMJ3U/31LKSlaWQeMjsxzVuPBj59TjLa3vjQtTWPkNNFMAi1yWWkAEVE5rvb3I1SebxPhKo/IowslsKD+GVPKeHc1tpwnRJDcG4iNm
us/GL+csng8fMfiJQN6xKYa0AsKZwzyKlan1oc2p4n6glh3mi8hp07eLHgxI+qyOd3VsfeDew==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

4YDZno

https://efirma.uaem.mx/noRepudio/pRJz6SiCWlAg3YdemWyaeWtuuBcBppJB

Cuernavaca, Morelos a 15 de enero del 2021.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN Y POSGRADO DE LA FCAeI
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Computo Aplicado, del estudiante Monje López Ricardo, con matricula 10023095, con el título **Algoritmo Jerárquico Aglomerativo Paralelo para la Agrupación (Clustering).** Por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación y Posgrado tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dra. Nelva Nely Almanza Ortega
Profesor- investigador
Instituto Tecnológico de Tlalnepantla

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

NELVA NELY ALMANZA ORTEGA | Fecha:2021-01-27 19:16:51 | Firmante

S039BkgmtN2kO8FSq0KLG2tFrrTAG0E0XtqnGjej3XY1rQSAu4/bWikqrblWqEcoXess3p6d/k71xDgIR/5NFQG1VwOf/YGrPb9B6ngzyxmj5NYfRbiWpAdwrA9rDgNQ1V7Ua+UPrLAaiqRw+rlntd+XYfTshNcNcPwhjpPaeagNUYJKY1jegYNrmhvqyiO2KmLOYVXQWyV+OtUK3GLuVUI+on27iUCY9d6qXSo3sZVKjrbjZf4mHvkemK6TB0smOdKZwggxtFhT3dcU1WUq3pZMYMGp1EDLv8RLnXBgQVJYIPNmMZHlOcIcQBBDLdXwzOJIuCcLOrGrihiULrM+jQ==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

1gt6mx

https://efirma.uaem.mx/noRepudio/vpK7302389PmbA9zB9jpzCrL6VyVSD9E

Cuernavaca, Morelos a 15 de enero del 2021.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN Y POSGRADO DE LA FCAeI
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Computo Aplicado, del estudiante Monje López Ricardo, con matricula 10023095, con el título **Algoritmo Jerárquico Aglomerativo Paralelo para la Agrupación (Clustering).** Por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación y Posgrado tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dr. José Alberto Hernández Aguilar
Profesor- investigador
Facultad de Contaduría Administración e Informática

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

JOSE ALBERTO HERNANDEZ AGUILAR | Fecha:2021-01-27 21:41:16 | Firmante

a4YDeDWqt95eO9depJrk1DSFAWvv1lpDe+lyMj/V7fybATHZLysHxzqsRDMWOHJfBRioUoSymkvLwAh3UKmMRsNtZEoqELjixRYHtRkvkh67aTKf9zk9qCQFW2HDpXbtUdSKinn+OI+N/CV7HC8aV+RXNMK0wdfXC5LTksBPWKj2bP+qU4tIZAEII6P7Tr21UMg4Pbbn32M2LwZP2imxzcCf4whCOZv3irkNWF0Qxcw6eMC8DQ7BynQGGXlTfFE+9D5Cp5oO8BmjXQ4a+4P09PfqTzfwZ83crDJp0QoO3o0soRVjvaeONicvRCFdrDXBW0IyM9H5XYW7/aRZcqGw7Q==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

E54iQ0

https://efirma.uaem.mx/noRepudio/mqQ6dhlK3lRSqzumIeiR3wd2XVdIefXz

Cuernavaca, Morelos a 15 de enero del 2021.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN Y POSGRADO DE LA FCAeI
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Computo Aplicado, del estudiante Monje López Ricardo, con matricula 10023095, con el título **Algoritmo Jerárquico Aglomerativo Paralelo para la Agrupación (Clustering).** Por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación y Posgrado tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dr. Martín Heriberto Cruz Rosales
Profesor- investigador
Facultad de Contaduría Administración e Informática

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

MARTIN HERIBERTO CRUZ ROSALES | Fecha:2021-01-31 18:10:48 | Firmante

p4BBUZr7p8bCjCq1Kawt5AMMhnNv/pbwx1up6/QJfZJcRzJHlUO6+ifqhM86gx5b3YyssnCAMX9uf2SIScz5TAD8UnRkwt3KjUsJj+KTxteck0s6TuiYTjiL/0m3HVFimR7gc2bC+OoNZVdMyjAp3gmcArrEg0fLmsWcfy3F4L/KZMbBp7YjvGC7x4z9Y7nEAEWOg/We5hcBmAM9tobX63tdEVCedfN6HlLsPQeMIr7uX5RZVFA0by4ogEg+1fUZ5AJV6JGSuFhgtR1Tav+Pr8Vw8DPhx7ZQUggHMOEmheGEWAygmnAnSw0T5xVCn90wicQA1HZu0iAltF/jLOqrGw==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

NsKByb

https://efirma.uaem.mx/noRepudio/C4UeOKAbcSlL7PJqr50Wl24taLRhvx9a

Cuernavaca, Morelos a 15 de enero del 2021.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN Y POSGRADO DE LA FCAeI
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Computo Aplicado, del estudiante Monje López Ricardo, con matricula 10023095, con el título **Algoritmo Jerárquico Aglomerativo Paralelo para la Agrupación (Clustering).** Por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación y Posgrado tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dr. Joaquín Pérez Ortega
Profesor- investigador
Centro Nacional de Investigación y Desarrollo Tecnológico

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

JOAQUÍN PÉREZ ORTEGA | Fecha:2021-04-18 21:10:55 | Firmante

NJB51mdKoa5xfOh+y6rU/wVzvvc4158Jt7AI9fpuVAjxl4rkKlZikG5FAoxNvivopu6XpvqGB0pOZywnR/zAf9iFNnMGJxE/IdPMjD4ntRxqWpzmpZHlO6gSKQsxQFQrmzb4KZU7FrgZ3u5gIfevUj0JwufB1rKrx89U+QfSEor16QXsXqqgAbFh9CgTs87bHjG6IqSegKbf5dPjmZeD/VQMDVMHPKY9QOyOL/V60h5gtjRZWD3PPVnxRcOuW+po2sl6gZBsW5wwCM4IPjZe7zZi/2L/kEV/yXo+LdMbMOcnHMSDA3iJxR4Wqvtxx2WR5gHyGC0pScwM1TYCuRmwGA==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

48PNdA

https://efirma.uaem.mx/noRepudio/FnVIgdQDUG5RHw0CI5D1F3tyGDzUP82Z

AGRADECIMIENTOS

Especialmente a Dios por darme la fuerza, voluntad y paciencia en esta etapa de mi vida.

A mis padres, esposa e hijo por su gran apoyo incondicional, estar a mi lado y sobretodo depositar su confianza en mí.

A mi asesor el Dr. José Crispín Zavala Díaz y mi Co-asesora la Dra. Nelva Nely Almanza Ortega por su paciencia y enseñanza en este proyecto realizado, y a todos los docentes por haberme permitido aprender un poco de lo mucho que saben.

Al Consejo de Ciencia y Tecnología (CONACYT) por el apoyo económico brindado durante la realización de este proyecto y a la Facultad de Contaduría Administración e Informática.

RESUMEN

En esta tesis de investigación se trabajó con la disminución de la complejidad del método de Ward de n^4 a n^3, mediante programación paralela, mejorando los tiempos de proceso sin disminuir la calidad de agrupación.

El método de Ward es uno de los diversos métodos jerárquicos de análisis de cluster, el cual utiliza el criterio de la suma de error al cuadrado.

El método de Ward es importante para la agrupación debido a que es capaz de agrupar objetos similares, aunque estén dispersos en el espacio de búsqueda, una de sus aplicaciones ha sido el análisis de la resistividad en rocas, análisis de yacimiento por presencia o ausencia de ciertos taxones, actividades que financian los préstamos, datos meteorológicos, etc. El método de Ward realiza una búsqueda exhaustiva entre los objetos, lo que lo hace un problema combinatorio, con una complejidad de n^4 debido a esto lo hace complicado para agrupar gran cantidad de objetos.

La programación paralela utiliza múltiples procesadores disponibles para resolver un problema. Se distingue de la programación secuencial en que varias operaciones pueden ocurrir simultáneamente.

Con la implementación de este algoritmo con programación paralelase logró el objetivo principal que es disminuir la complejidad del método. Dando como resultado un mejor tiempo de solución, y con la agrupación igual a las instancias de prueba que se obtuvieron de la literatura, con datos reales y de prueba.

Con esta mejora en la disminución de la complejidad se pueden resolver problemas con mayor número de objetos en menor tiempo.

Abstract

In this research thesis, we worked with the reduction of the complexity of Ward's method from $n^4$to n^3, through parallel programming, improving process times without reducing the quality of grouping.

Ward's method is one of several hierarchical methods of cluster analysis, which uses the criterion of the sum of squared error.

Ward's method is important for grouping because it is capable of grouping similar objects even though they are scattered in the search space, one of its applications has been the analysis of resistivity in rocks, reservoir analysis by presence or absence of certain taxa, activities that finance loans, weather data, etc. Ward's method performs an exhaustive search between the objects, which makes it a combinatorial problem, with a complexity of n^4 due to this it makes it difficult to group large amounts of objects.

"Parallel programming is the use of multiple computational resources to solve a problem". It is distinguished from sequential programming in that several operations can occur simultaneously.

With the implementation of this algorithm with parallel programming, the main objective was achieved, which is to reduce the complexity of the method. Resulting in a better solution time, and with the grouping equal to the test instances that were obtained from the literature, with real and test data.

With this improvement in the reduction of complexity, problems with a greater number of objects can be solved in less time.

CONTENIDO

Pág.

LISTA DE FIGURAS

Pág.

LISTA DE TABLAS

Pág.

CAPÍTULO I. INTRODUCCIÓN

Los avances en la tecnología de detección, almacenamiento y el crecimiento espectacular en aplicaciones tales como la búsqueda en Internet, imágenes digitales y video vigilancia han creado muchos conjuntos de datos de gran volumen y gran dimensión (Jain, 2010).

El análisis de estos grandes conjuntos de datos y la aplicación de técnicas para el reconocimiento de patrones, también se le conoce como aprendizaje y se divide en dos principales categorías (Jain, 2010).

I. **Supervisado (clasificación):** Sólo datos etiquetados (patrones de entrenamiento con etiquetas de categoría conocidas).
II. **No supervisado (agrupamiento):** Sólo incluye datos no etiquetados.

"La agrupación desempeña un papel destacado en aplicaciones de Minería de Datos tales como exploración de datos científicos, recuperación de información y minería de texto, aplicaciones de bases de datos, análisis web, marketing, diagnóstico médico, biología computacional y muchas otras. La Minería de Datos se suma a la agrupación de las complicaciones de conjuntos de datos muy grandes. Además, la agrupación en clusters es objeto de investigación activa en varios campos, como estadísticas, reconocimiento de patrones y aprendizaje automático" (Berkhin, 2006).

El objetivo de la agrupación de datos, es descubrir patrones en las agrupaciones naturales de un conjunto puntos u objetos (Jain, 2010). Una definición operacional de agrupamiento puede establecerse de la siguiente manera: Dada una representación de *n* objetos, agruparlos en *k* grupos, basándose en una medida de similitud tal que los objetos dentro del mismo grupo sean similares pero diferentes a los otros grupos (Jain, 2010).

En la Figura 1, se muestran las técnicas de agrupamiento se dividen ampliamente en particional y jerárquico (Berkhin, 2006).

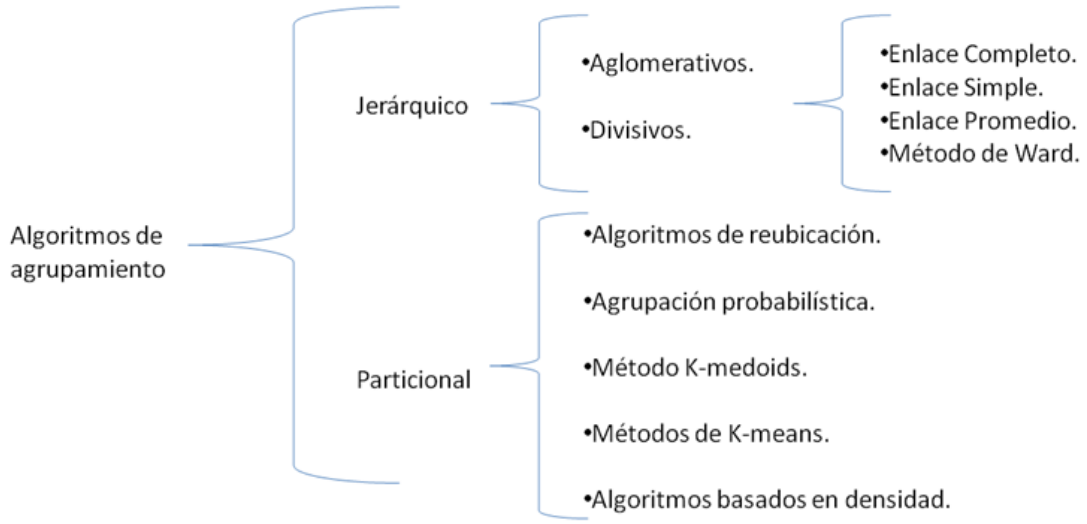

Figura 1. Algoritmos de agrupamiento.

La agrupación jerárquica crea una jerarquía de cluster (un árbol de clusters), también conocido como dendrograma. Tal enfoque permite explorar datos en diferentes niveles de granularidad y se subdivide en aglomeración y división. Los conceptos básicos de la agrupación jerárquica incluyen la fórmula de Lance-Williams (Berkhin, 2006) y el método de Ward que utiliza la suma de error al cuadrado.

La agrupación jerárquica revela la relación genética entre cada observación, porque agrupa iterativamente las observaciones "similares" para formar un árbol jerárquico, hasta que todas las observaciones se reúnen en un solo grupo si es aglomerativo. El resultado se puede visualizar en diferentes niveles del árbol jerárquico, lo que proporciona vistas sintéticas globales y detalladas del conjunto de datos analizado (Dumont et al, 2018).

Los algoritmos de partición intentan descubrir clusters mediante la reubicación iterativa de puntos entre subconjuntos, o intentan identificar clusters como áreas altamente pobladas con datos. Los algoritmos del primer tipo se encuentran en la

sección Métodos de reubicación de particiones. Se categorizan en agrupamientos probabilísticos, métodos k-medoids, y métodos k-means, los algoritmos de partición del segundo tipo se encuestan en la sección Partición basada en densidad. Intentan descubrir componentes densos de datos conectados, que son flexibles en términos de su forma (Berkhin, 2006).

1.1. Planteamiento del problema

El método de Ward es el único entre los métodos de agrupamiento aglomerado que utiliza el criterio de suma de error al cuadrado, esto da como resultado grupos que minimizan la dispersión dentro del grupo en cada fusión. La suma de error al cuadrado, la varianza mínima u otros criterios relacionados son problemas de optimización NP-completa (Murtagh, & Legendre, 2014). Esto implica que no es posible una solución en un tiempo polinomial. Sólo la búsqueda exponencial en el espacio de la solución, ya que es exhaustiva, garantizará una solución óptima (Murtagh, & Legendre, 2014).

La complejidad del método de Ward es n^4, lo que lo hace muy complejo y esto conlleva a que no se puede trabajar con un número de n tan grande por el espacio y tiempo de búsqueda que se lleva el método.

1.2. Hipótesis

H1: Es posible generar un algoritmo jerárquico Ward paralelo para agrupar en un menor tiempo que un secuencial.

H0: No es posible generar un algoritmo jerárquico Ward paralelo para agrupar en un menor tiempo que un secuencial.

1.3. Objetivos

1.3.1 Objetivo General

Implementar el algoritmo de Ward en paralelo para reducir el tiempo de ejecución.

1.3.2 Objetivos Específicos

- Desarrollar un algoritmo jerárquico-aglomerativo de agrupamiento con programación en paralelo.
- Medir los tiempos de ejecución del algoritmo modificado utilizando instancias referenciadas en artículos.
- Evaluar los resultados de nuestro algoritmo en tiempo y calidad de agrupamiento.

1.4. Justificación

Debido a que el método de Ward tiene una complejidad alta, no existen instancias o resultados con un número de objetos alto, el número mayor encontrado con datos es de 312 objetos y dos atributos, cabe mencionar que estas instancias son para el método Ward, por tal motivo se pretende desarrollar un algoritmo paralelo que resuelva este problema y poder aplicarlo en datos reales como los que se manejan en plantación de algún cultivo (frijol, maíz, etc.), que contenga mayor número de objetos y de variables (atributos).

1.5. Alcances y Limitaciones

1.5.1 Alcances

Al término de esta tesis de investigación se obtendrá un algoritmo capaz de agrupar objetos con similitud de un grupo, pero diferente a otro, sin importar el número de objetos, atributos y su posición en el espacio de búsqueda, esto quiere decir que podrá agrupar con más de 2 atributos, no solo *x* y *y*, el espacio de búsqueda se refiere a la posición del objeto en el plano cartesiano.

1.5.2 Limitaciones

a) La investigación se realizará en 2 años.
b) Falta de instancias, especialmente para cultivos.
c) Los datos son de un artículo de cultivo de Brasil (único).
d) Falta de resultados paralelos en la literatura.

CAPÍTULO II. MARCO TEÓRICO

Método particional.

En el clustering particional el objetivo es obtener una partición de los objetos en grupos o clusters, de tal forma que todos los objetos pertenezcan a alguno de los *k* clusters posibles de acuerdo con una función objetivo y su grado se similitud, y que por otra parte los clusters sean disjuntos disimilares al resto de los grupos. Uno de los problemas en aplicaciones prácticas es el desconocimiento del valor de *k* adecuado (Larranaga et al, 2012).

Un esquema de reubicación consiste en reasignar puntos de forma iterativa entre los *k* clusters. A diferencia de los métodos jerárquicos tradicionales, en los que los conglomerados no se vuelven a visitar después de su construcción, los algoritmos de reubicación mejoran gradualmente los conglomerados (Berkhin, 2006).

1. **Método K-Medoids:** Es una solución fácil ya que cubre cualquier tipo de atributo y los K-Medoids tienen resistencia integrada contra valores atípicos ya que los puntos periféricos del cluster no los afectan. Cuando se seleccionan medoides, los conglomerados se definen como subconjuntos de puntos cercanos a los medoides respectivos, y la función objetivo se define como la distancia promediada u otra medida de disimilitud entre un punto y su medoide (Berkhin, 2006).
2. **Método K-Means:** Es la herramienta de agrupamiento más popular utilizada en aplicaciones científicas e industriales. El nombre proviene de representar cada uno de los *k* agrupamientos por la media (o promedio ponderado) de sus puntos, el llamado centroide.

   Si bien esto obviamente no funciona bien con los atributos categóricos, tiene un buen sentido geométrico y estadístico para los atributos numéricos (Berkhin, 2006).

3. **Algoritmos basados en densidad:** La implementación de esta idea para la partición de un conjunto finito de puntos requiere conceptos de densidad, conectividad y límite. Están estrechamente relacionados con los vecinos más cercanos de un punto. Un cluster, definido como un componente denso conectado, crece en cualquier dirección que conduzca la densidad. Por lo tanto, los algoritmos basados en densidad son capaces de descubrir grupos de formas arbitrarias. También esto proporciona una protección natural contra valores atípicos.

Método Jerárquico.

El algoritmo jerárquico crea una jerarquía de cluster (un árbol de clusters), también conocido como dendrograma Figura 3. Cada nodo de cluster contiene clusters secundarios; Los métodos de agrupación jerárquica se clasifican en aglomerativos (ascendentes) y divisivos (descendentes) (Berkhin, 2006).

1. **Aglomerativo:** Comienza con clusters de un punto y fusiona recursivamente dos o más clusters más apropiados.
2. **Divisivo:** Comienza con un grupo de todos los puntos de datos y divide recursivamente el cluster más apropiado.

En la Figura 2, se muestra un ejemplo de la agrupación jerárquica.

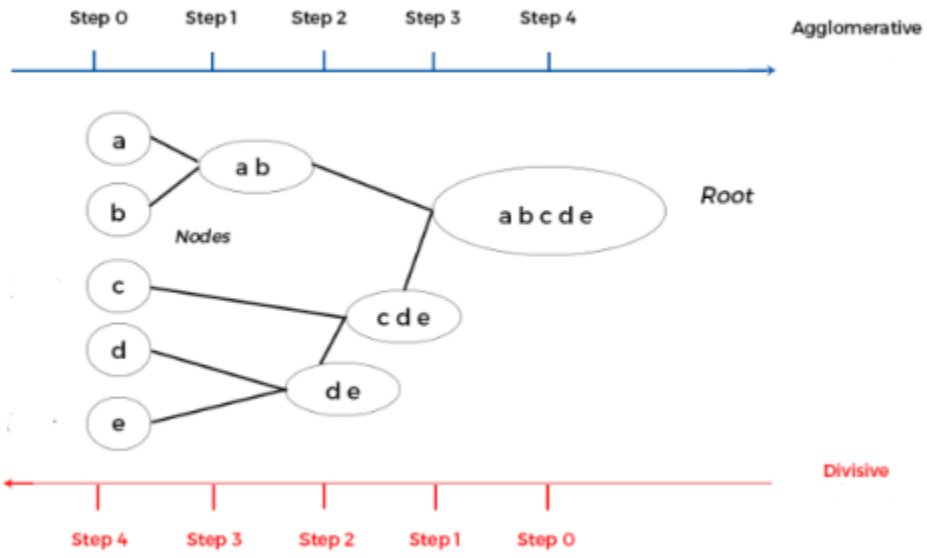

Figura 2. Ejemplo de los métodos aglomerativo y divisivo.

Dendrograma: En la Figura 3, se muestra el dendrograma y se construye a partir de las hojas y combina las agrupaciones hasta el tronco (James et al, 2013).

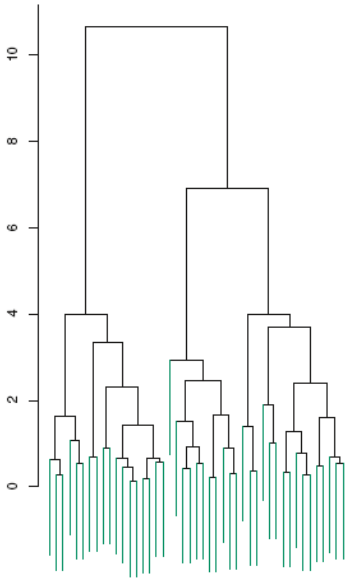

Figura 3. Dendrograma.

En el eje Y se coloca el valor de la distancia y en el eje X el número de clusters, las ramas que están unidas en la parte inferior del eje Y, son más similares que las que se unen más arriba (James et al, 2013).

Corte del dendrograma: Para tener un número de *k* óptimo, se utiliza el método del codo, esto quiere decir que donde se encuentre la distancia de unión con mayor salto se da el corte, como se muestra en la Figura 4 (Dumont et al, 2018).

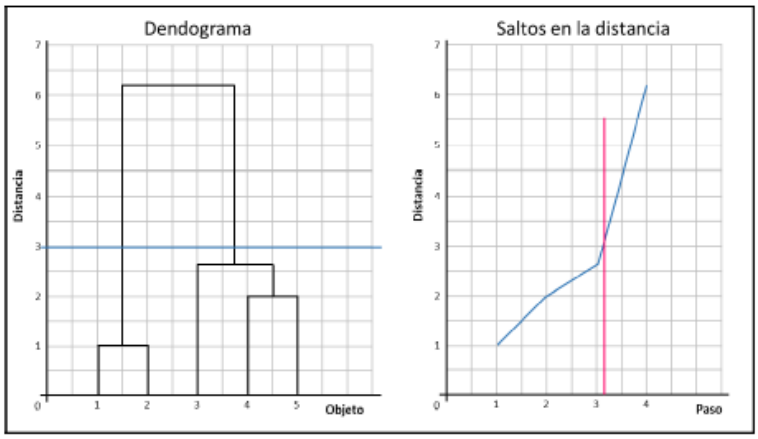

Figura 4. Corte del dendrograma.

Método jerárquico aglomerativo.

Este método puede utilizar diferentes tipos de enlace (Completo o máximo, simple o mínimo, promedio y de Ward), todos ellos trabajan de diferente manera por lo cual se generan diferentes resultados (dendrograma), por eso es importante conocer el problema para poder seleccionar el enlace más adecuado. Por ejemplo, en la Tabla 1 se muestran los datos de 5 cluster a agrupar.

Tabla 1. Valores de los clusters en el espacio euclidiano.

	X	Y
clusterA	1	1
clusterB	2	1
clusterC	4	2
clusterD	1	4
clusterE	4	3

En la Figura 5, se visualiza la distribución de los clusters a agrupar en el espacio euclidiano, según con los datos de la Tabla 1.

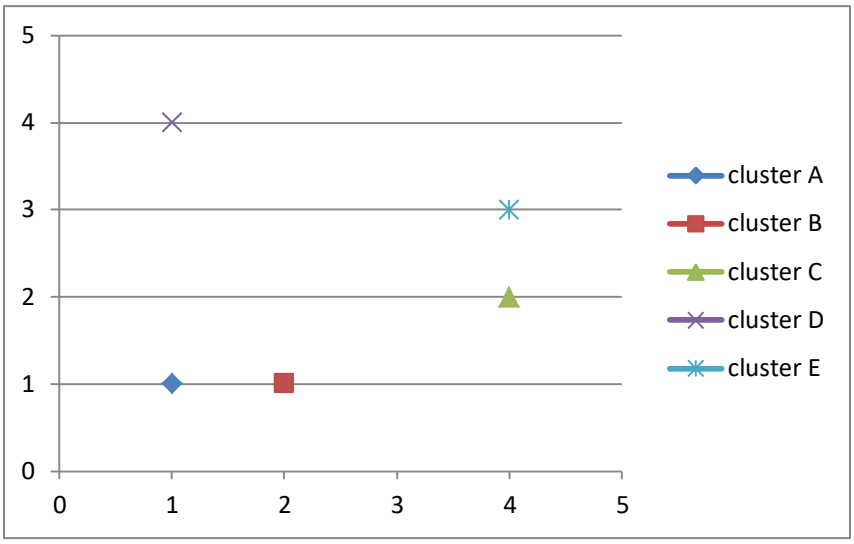

Figura 5. Los 5 clusters en el espacio para su agrupación.

A. **Completo/Máximo:** De la unión de dos clusters (C_i, C_j), se compara la distancia con los clusters restantes respectivamente y se toma el mayor.

$$D\,((C_i, C_j), (C_k)) = \text{máx}\,\{\,D\,(C_i, C_k), D\,(C_j, C_k))\} \qquad (1)$$

Donde C_k es cada uno de los clusters que interactúan con C_i y C_j, D la distancia.

Ejemplo 1. Se tiene la Tabla 2 que muestra la matriz con la distancia euclidiana, se toman los dos clusters más similares.

Tabla 2. Distancia euclidiana al cuadrado.

	A	B	C	D	E
A	0	1	10	9	13
B		0	5	10	8
C			0	13	1
D				0	10
E					0

Paso 1. Unión del cluster A y cluster B.

$D\ ((C_i, C_j), (C_k)) = \text{máx}\ \{D\ (C_A, C_C), D\ (C_B, C_C)\} = (\mathbf{10},5)$

$D\ ((C_i, C_j), (C_k)) = \text{máx}\ \{D\ (C_A, C_D), D\ (C_B, C_D)\} = (9,\mathbf{10})$

$D\ ((C_i, C_j), (C_k)) = \text{máx}\ \{D\ (C_A, C_E), D\ (C_B, C_E)\} = (\mathbf{13},8)$

Como se observa en la Tabla 3, en cada comparación se toma el valor con mayor distancia y se actualiza la matriz.

Tabla 3. Distancia euclidiana actualizada paso 1 completo.

	A,B	C	D	E
A,B	0	10	10	13
C		0	13	1
D			0	10
E				0

Paso 2. Unión del cluster C y cluster E.

$D\ ((C_i, C_j), (C_k)) = \text{máx}\ \{D\ (C_C, C_{A,B}), D\ (C_E, C_{A,B})\} = (10,\mathbf{13})$

$D\ ((C_i, C_j), (C_k)) = \text{máx}\ \{D\ (C_C, C_D), D\ (C_E, C_D)\} = (\mathbf{13},10)$

Como se observa en la Tabla 4, en cada comparación se toma el valor con mayor distancia y se actualiza la matriz.

Tabla 4. Distancia euclidiana actualizada paso 2 completo.

	A,B	C,E	D
A,B	0	13	10
C,E		0	13
D			0

Paso 3. Unión del cluster A, B y cluster D.

$D\ ((C_i,C_j), (C_k)) = \text{máx}\ \{D\ (C_{A,B}, C_{C,E}), D\ (C_D, C_{C,E})\} = (13,\mathbf{13})$

Como se observa en la Tabla 5, en cada comparación se toma el valor con mayor distancia y se actualiza la matriz.

Tabla 5. Distancia euclidiana actualizada paso 3 completo.

	A,B,D	C,E
A,B,D	0	13
C,E		0

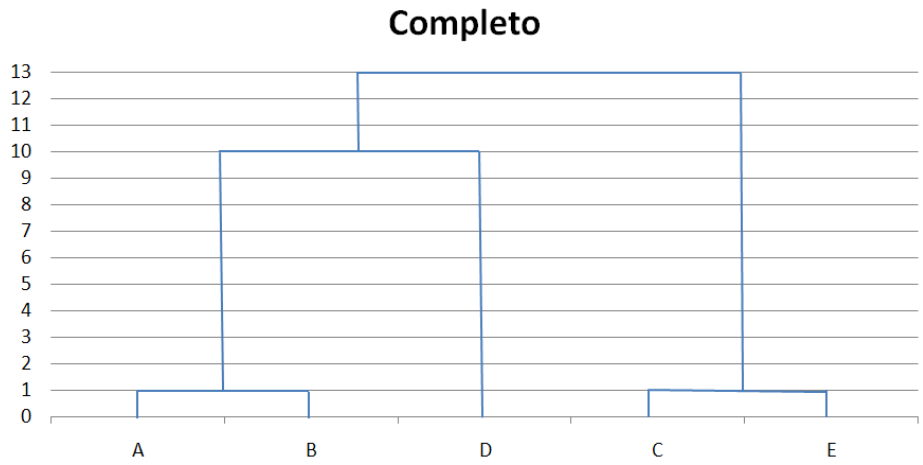

Figura 6. Dendrograma de enlace completo.

El dendrograma de enlace completo que se muestra en la Figura 6 es resultado de la agrupación de los clusters de la Figura 5 y las distancias obtenidas es de la unión de los clusters en cada paso de la Tabla 2, Tabla 3, Tabla 4 y Tabla 5.

B. **Simple/Mínimo:** De la unión de dos clusters (C_i, C_j), se compara la distancia con los clusters restantes respectivamente y se toma el menor.

$$D\,((C_i,C_j),(C_k)) = \min\{\,D\,(C_i,C_k),\,D\,(C_j,C_k)\}\qquad(2)$$

Donde C_k es cada uno de los clusters que interactúan con C_i y C_j. D es la distancia.

Ejemplo 2. Se tiene la Tabla 2 con matriz de distancia euclidiana, se toman los dos clusters más similares.

Paso 1. Unión del cluster A y cluster B.

$D\ ((C_i, C_j), (C_k)) = \min \{ D\ (C_A, C_C), D\ (C_B, C_C) \} = (10,\underline{\mathbf{5}})$

$D\ ((C_i, C_j), (C_k)) = \min \{ D\ (C_A, C_D), D\ (C_B, C_D) \} = (\underline{\mathbf{9}},10)$

$D\ ((C_i, C_j), (C_k)) = \min \{ D\ (C_A, C_E), D\ (C_B, C_E) \} = (13,\underline{\mathbf{8}})$

Como se observa en la Tabla 6, en cada comparación se toma el valor con menor distancia y se actualiza la matriz.

Tabla 6. Distancia euclidiana actualizada paso 1 simple.

	A,B	C	D	E
A,B	0	5	9	8
C		0	13	1
D			0	10
E				0

Paso 2. Unión del cluster C y cluster E.

$D\ ((C_i, C_j), (C_k)) = \min \{ D\ (C_C, C_{A,B}), D\ (C_E, C_{A,B}) \} = (\underline{\mathbf{5}},8)$

$D\ ((C_i, C_j), (C_k)) = \min \{ D\ (C_C, C_D), D\ (C_E, C_D) \} = (13,\underline{\mathbf{10}})$

Como se observa en la Tabla 7, en cada comparación se toma el valor con menor distancia y se actualiza la matriz.

Tabla 7. Distancia euclidiana actualizada paso 2 simple.

	A,B	C,E	D
A,B	0	5	9
C,E		0	10
D			0

Paso 3. Unión del cluster A, B y cluster C, E.

$D\ ((C_i, C_j), (C_k)) = \min \{ D\ (C_{A,B}, C_D), D\ (C_{CE}, C_D) \} = (\underline{\mathbf{9}},10)$

Como se observa en la Tabla 8, en cada comparación se toma el valor con menor distancia y se actualiza la matriz.

Tabla 8. Distancia euclidiana actualizada paso 3 simple.

	A,B,C,E	D
A,B,C,E	0	9
D		0

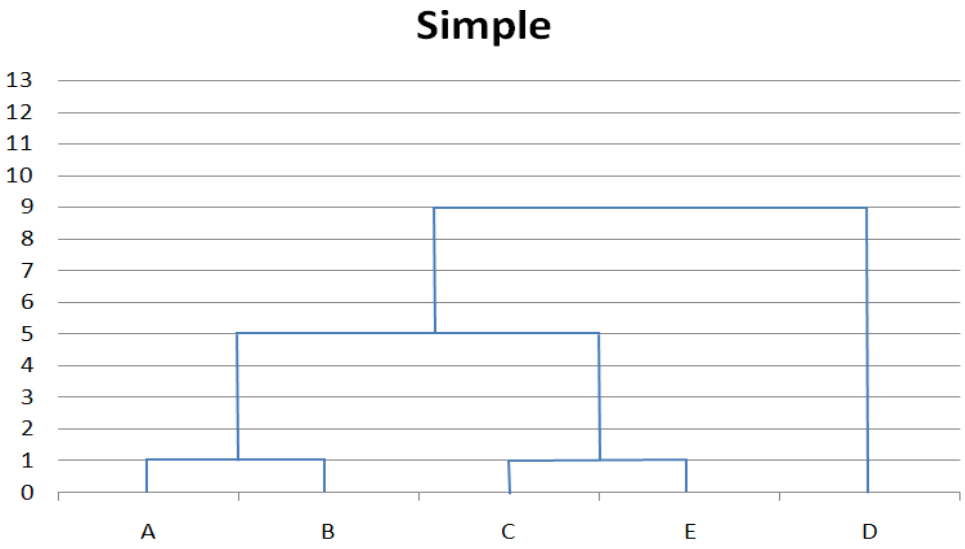

Figura 7. Dendrograma de enlace simple.

El dendrograma de enlace simple de la Figura 7 muestra el resultado de la agrupación de los clusters de la Figura 5, y las distancias obtenidas de la unión de los clusters en cada paso de la Tabla 6, Tabla 7 y Tabla 8. A diferencia del enlace completo, en este caso, se toma la distancia de cada unión que es menor, además de la forma de agrupar, ya que en este primero agrupa *A*,*B* con *C*,*E* y al último *D*.

C. Promedio: De la unión de dos clusters (C_i, C_j), se suman las distancias de los clusters restantes respectivamente y se divide entre dos.

$$D((C_i, C_j), (C_k)) = [D(C_i, C_k) + D(C_j, C_k)]/2. \quad (3)$$

Donde C_k es cada uno de los clusters que interactúan con C_i y C_j. D la distancia.

Ejemplo 3. De las distancias euclidianas, de la matriz de la Tabla 2, se toman los dos clusters más similares.

Paso 1. Unión del cluster A y cluster B.

$D((C_i, C_j), (C_k)) = D(C_A, C_C) + D(C_B, C_C)\} = [(10+5)]/2 = \underline{\mathbf{7.5}}$

$D((C_i, C_j), (C_k)) = D(C_A, C_D) + D(C_B, C_D)\} = [(9+10)]/2 = \underline{\mathbf{9.5}}$

$D((C_i, C_j), (C_k)) = D(C_A, C_E) + D(C_B, C_E)\} = [(13+8)]/2 = \underline{\mathbf{10.5}}$

Como se observa en la Tabla 9, en cada comparación se suman los dos valores y se dividen entre 2, se actualiza la matriz.

Tabla 9. Distancia euclidiana actualizada paso 1promedio.

	A,B	C	D	E
A,B	0	7.5	9.5	10.5
C		0	13	1
D			0	10
E				0

Paso 2. Unión del cluster C y cluster E.

$D((C_i, C_j), (C_k)) = D(C_C, C_{A,B}) + D(C_E, C_{A,B})\} = [(7.5+10.5)]/2 = \underline{\mathbf{9}}$

$D((C_i, C_j), (C_k)) = D(C_C, C_D) + D(C_E, C_D)\} = [(13+10)]/2 = \underline{\mathbf{11.5}}$

Como se observa en la Tabla 10, en cada comparación se suman los dos valores y se dividen entre 2, se actualiza la matriz.

Tabla 10. Distancia euclidiana actualizada paso 2 promedio.

	A,B	C,E	D
A,B	0	9	9.5
C,E		0	11.5
D			0

Paso 3. Unión del cluster A,B y cluster C,E.

$D\ ((C_i, C_j), (C_k)) = D\ (C_{A,B}, C_D) + D\ (C_{C,E}, C_D)\ \} = [(9.5+11.5)]/2 =$ **10.5**

Como se observa en la Tabla 11, en cada comparación se suman los dos valores y se dividen entre 2, se actualiza la matriz.

Tabla 11. Distancia euclidiana actualizada paso 3 promedio.

	A,B,C,E	D
A,B,C,E	0	10.5
D		0

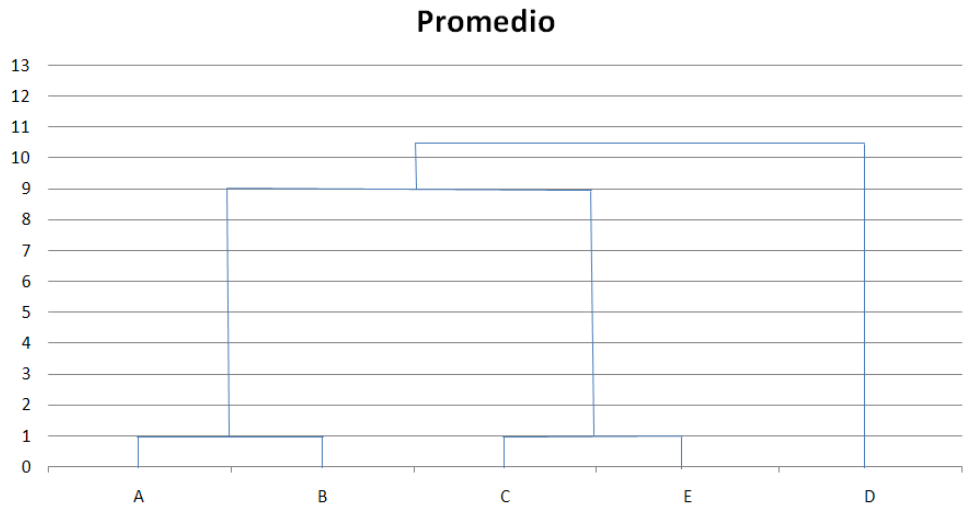

Figura 8. Dendrograma de enlace promedio.

El dendrograma de enlace promedio de la Figura 8 es el resultado de la agrupación de los clusters de la Figura 5, y las distancias obtenidas es de la unión de los clusters en cada paso de la Tabla 9, Tabla 10 y Tabla 11. A diferencia de los enlaces anteriores, en este caso la distancia de cada unión, es intermedio y la agrupación es como el enlace simple.

Ward: El método de Ward es interesante porque busca grupos en el espacio euclidiano (Murtagh, & Legendre, 2014), esto quiere decir que hace una búsqueda exhaustiva de un punto con el resto.

(Lance & Williams,1967) establecieron una forma para la actualización de las disimilitudes después de una aglomeración. La que considere la posibilidad de que los grupos i y j se aglomeren para formar el grupo i ∪ j, y luego considera la posibilidad de redefinir la disimilitud con respecto a un grupo externo, k. Se tiene la expresión 4 (Murtagh, & Legendre, 2014).

$$d(i \cup j, k) = a(i).d(i,k) + a(j).d(j,k) + b.d(i,j) + c|d(i,) - d(j,k)| \quad \textbf{(4)}$$

Dónde d es la disimilitud utilizada, que no tiene que ser una distancia euclidiana para comenzar y |.| denota el valor absoluto.

La fórmula de recurrencia de Lance-Williams considera las diferencias y no las diferencias al cuadrado.

Entonces Ward surge de la formula actualizada de Lance-Williams y se dividen en 2 fórmulas Ward1, Ward2, la diferencia de estas dos es que para Ward1 la entrada de la distancia es al cuadrado, y para Ward2 la entrada no es al cuadrado (Murtagh, & Legendre, 2014).

a) **Ward1:** Inicialmente, fue (Wishart, 1969) quien escribió el algoritmo de Ward en términos de la fórmula de actualización de Lance-Williams que se muestra en la expresión 5 (Murtagh, & Legendre, 2014).

$$\delta\left(i \cup i', i''\right) = \frac{\omega_i + \omega_i''}{\omega_i + \omega_{i'} + \omega_{i''}} \delta(i, i'') + \frac{\omega_i' + \omega_i''}{\omega_i + \omega_{i'} + \omega_{i''}} \delta(i', i'') - \frac{\omega_i''}{\omega_i + \omega_{i'} + \omega_{i''}} \delta(i, i') \textbf{(5)}$$

Donde δ = Distancia (euclidiana), ω_i= Cardinalidad del cluster, todos empiezan con valor 1 y conforme se van uniendo se van sumando.

La expresión (5) ayuda a obtener la disimilaridad de todos los objetos, esto quiere decir que, si se tienen 5 clusters, el cluster 1 se va a comparar con el resto de los clusters (2,3,4,5), posteriormente el cluster 2 con el resto (3,4,5), esto hasta llegar al cluster 5. Cuando ya se tiene la disimilaridad de todos, se toman los dos clusters más similares, se hace la unión y se actualizan sus atributos, con este proceso se obtiene el centroide, el cual será un nuevo cluster. La expresión 6 (Murtagh, & Legendre, 2014) muestra el cálculo de los centroides.

$$(C_i \cup C_j) = \frac{|C_i| C_i + |C_j| C_j}{|C_i| + |C_j|} = C_n \qquad \textbf{(6)}$$

Donde |C| es la cardinalidad del cluster,C_n es el nuevo centroide o nuevos atributos de la unión de los dos clusters.

Posteriormente se repite hasta que el valor de *n* sea igual a 1 (*n*=1).

b) **Ward2:** En cada paso de aglomeración, la suma extra de cuadrados causada por los grupos de aglomeración se minimiza, exactamente como se mostró para el algoritmo Ward1. La expresión 7 muestra cómo se actualizan los clusters (Murtagh, & Legendre, 2014).

$$\delta(i \cup i', i'') = \left(\begin{array}{c} \frac{\omega_i + \omega_i''}{\omega_i + \omega_{i'} + \omega_{i''}} \delta^2(i, i'') + \frac{\omega_i' + \omega_i''}{\omega_i + \omega_{i'} + \omega_{i''}} \delta^2(i', i'') \\ - \frac{\omega_i''}{\omega_i + \omega_{i'} + \omega_{i''}} \delta^2(i, i') \end{array} \right) 1/2 \qquad \textbf{(7)}$$

Donde δ^2 = Distancia (euclidiana), ω_i= Cardinalidad del cluster todos empiezan con valor 1 y conforme se van uniendo se van sumando.

Al contrario de Ward1, las diferencias de entrada son distancias euclidianas (no cuadradas).

Ejemplo del método Ward2: Se tienen las matrices que se muestran en la Tabla 12 de matriz de atributos y Tabla 13 de distancia euclidiana.

Tabla 12. Matriz de atributos.

	Atributo1	Atributo2	Cardinalidad
A	1	1	1
B	2	1	1
C	4	2	1
D	1	4	1
E	4	3	1

Tabla 13. Distancia euclidiana no cuadrada.

	A	B	C	D	E
A	0	1	10	9	13
B		0	5	10	8
C			0	13	1
D				0	10
E					0

Paso 1: Se utiliza la expresión 7 de Ward2, para obtener cuáles clusters son más disimilares:

$$\delta(A \cup B, C) = \begin{pmatrix} \frac{1+1}{1+1+1}\delta^2(10) + \frac{1+1}{1+1+1}\delta^2(5) \\ -\frac{1}{1+1+1}\delta^2(1) \end{pmatrix} 1/2$$

$$\delta(A \cup B, C) = \begin{pmatrix} \frac{2}{3}(10) + \frac{2}{3}(5) \\ -\frac{1}{3}(1) \end{pmatrix} 1/2$$

$$\delta(A \cup B, C) = \begin{pmatrix} 6.66 + 3.33 \\ -0.33 \end{pmatrix} 1/2$$

$$\delta(A \cup B, C) = (9.66)1/2$$

$$\underline{\delta(A \cup B, C) = \mathbf{3.10}}$$

Este es el resultado de la unión de A,B referente a C, y esto se repite cambiando a C por D luego por E, cuando se termina se cambia la B por C $\underline{(A \cup C, B)}$ y el proceso es el mismo hasta llegar al último dato, en este caso sería cuando se llegue al E en lugar de la A, a continuación están los ejemplos y sus resultados:

A ∪ B, C =3.10

A ∪ B, D =3.51

A ∪ B, E =3.69

A∪ C, B =0.81

A ∪ C, D =3.36

A ∪ C, E =2.44

A ∪ D, B =2.08

A ∪ D, C =3.51

A ∪ D, E =3.51

A ∪ E, B =1.29

A ∪ E, C =1.73

A ∪ E, D =2.88

B ∪ C, A =2.38

B ∪ C, D =3.69

B ∪ C, E =2.08

B ∪ D, A =1.82

B ∪ D, C =2.94

B ∪ D, E =2.94

B ∪ E, A =2.58

B ∪ E, C =1.15

B ∪ E, D =3.26

C ∪ D, A =2.88

C ∪ D, B =2.38

C ∪ D, E =1.73

C ∪ E, A = 3.87

C ∪ E, B =2.88

C ∪ E, D = 3.87

D ∪ E, A =3.36

D ∪ E, B =2.94

D ∪ E, C =2.44

Como se observa los clusters más disimilares son *C* y *E.*

Paso 2: Se actualizan los atributos del cluster *C* y *E*, cabe mencionar que la fórmula se ocupa para cada atributo, en este ejemplo se ocupa dos veces uno para el atributo 1 y otro para el atributo 2.

$$(C_C \cup C_E) = \frac{|1|4 + |1|4}{|1| + |1|} = C_1$$

$$(C_C \cup C_E) = \frac{4 + 4}{2} = C_1$$

$$(C_C \cup C_E) = \frac{8}{2} = C_1$$

$$(C_C \cup C_E) = 4 = C_1$$

Para el atributo 1 del nuevo centroide el valor es 4.

$$(C_C \cup C_E) = \frac{|1|2 + |1|3}{|1| + |1|} = C_2$$

$$(C_C \cup C_E) = \frac{2+3}{2} = C_2$$

$$(C_C \cup C_E) = \frac{5}{2} = C_2$$

$$(C_C \cup C_E) = 2.5 = C_2$$

Para el atributo 2 del nuevo centroide, el valor es 2.5. Con estos valores, se actualizan los atributos sumando la cardinalidad de los dos clusters y se obtiene la distancia euclidiana como se muestra en la Tabla 14 y Tabla 15.

Tabla 14. Matriz de atributos actualizada paso 2 Ward2.

	Atributo1	Atributo2	Cardinalidad
A	1	1	1
B	2	1	1
C,E	4	2.5	2
D	1	4	1

Tabla 15. Distancia euclidiana no cuadrada actualizada paso 2 Ward2.

	A	B	C,E	D
A	0	1	11.25	9
B		0	6.25	10
C,E			0	11.25

D				0

Se repiten los pasos 1 y 2.

Paso 3: Se utiliza la expresión 7 de Ward2, para obtener cuales clusters son más disimilares:

A ∪ B, C = 3.55

A ∪ B, D = 3.51

A ∪C,E, B = 1.54

A ∪ C,E, D = 3.18

A ∪ D, B = 2.08

A ∪ D, C,E = 3.51

B ∪ C,E, A = 2.71

B ∪ C,E, D = 3.44

B ∪ D, A = 1.82

B ∪ D, C,E = 2.85

C,E ∪ D, A = 3.18

C,E ∪ D, B = 2.62

Como se observa los clusters más disimilares son *A*, *B*.

Paso 4: Se actualizan los atributos del cluster *A* y *B*.

$$(C_A \cup C_B) = \frac{|1|1 + |1|2}{|1| + |1|} = C_1$$

$$(C_A \cup C_B) = \frac{1 + 2}{2} = C_1$$

$$(C_A \cup C_B) = \frac{3}{2} = C_1$$

$$(C_A \cup C_B) = 1.5 = C_1$$

Para el atributo 1 del nuevo centroide, el valor es 1.5

$$(C_A \cup C_B) = \frac{|1|1 + |1|1}{|1| + |1|} = C_2$$

$$(C_A \cup C_B) = \frac{1 + 1}{2} = C_2$$

$$(C_A \cup C_B) = \frac{2}{2} = C_2$$

$$(C_A \cup C_B) = 1 = C_2$$

Para el atributo 2 del nuevo centroide, el valor es 1. Con esto valores se actualizan los atributos sumando la cardinalidad de los dos clusters y se obtiene la distancia euclidiana como se muestra en la Tabla 16 y Tabla 17.

Tabla 16. Matriz de atributos actualizada paso 4 Ward 2.

	Atributo1	Atributo2	Cardinalidad
A,B	1.5	1	2
C,E	4	2.5	2
D	1	4	1

Tabla 17. Distancia euclidiana no cuadrada actualizada paso 4 Ward2.

	A,B	C,E	D
A,B	0	8.5	9.25
C,E		0	11.25
D			0

Paso 5: Se utiliza la expresión 7 de Ward2, para obtener cuales clusters son más disimilares:

A,B ∪ C,E, D = 3.25

A,B ∪ D, C,E = 3.13

C,E ∪ D, A,B =2.80

Como se observa los clusters más disimilares son *A,B* Y *C,E*.

Paso 6: Se actualizan los atributos del cluster *A,B* y *C,E.*

$$(C_{A,B} \cup C_{C,E}) = \frac{|2|1.5 + |2|4}{|2| + |2|} = C_1$$

$$(C_{A,B} \cup C_{C,E}) = \frac{3 + 8}{4} = C_1$$

$$(C_{A,B} \cup C_{C,E}) = \frac{11}{4} = C_1$$

$$(C_{A,B} \cup C_{C,E}) = 2.75 = C_1$$

Para el atributo 1 del nuevo centroide, el valor es 2.75

$$(C_{A,B} \cup C_{C,E}) = \frac{|2|1 + |2|2.5}{|2| + |2|} = C_2$$

$$(C_{A,B} \cup C_{C,E}) = \frac{2 + 5}{4} = C_2$$

$$(C_{A,B} \cup C_{C,E}) = \frac{7}{4} = C_2$$

$$(C_{A,B} \cup C_{C,E}) = 1.75 = C_2$$

Para el atributo 2 del nuevo centroide, el valor es 1.75. Con esto valores se actualizan los atributos sumando la cardinalidad de los dos clusters y se obtiene la distancia euclidiana como se muestra en la Tabla 18 y Tabla 19.

Tabla 18. Matriz de atributos actualizada paso 6 Ward2.

	Atributo1	Atributo2	Cardinalidad
A,B,C,E	1.5	1	4
D	1	4	1

Tabla 19. Distancia euclidiana no cuadrada actualizada paso 6 Ward2.

	A,B,C,E	D
A,B,C,E	0	9.25
D		0

Y con esto se termina el proceso y se obtiene la agrupación que se muestra en la Figura 9.

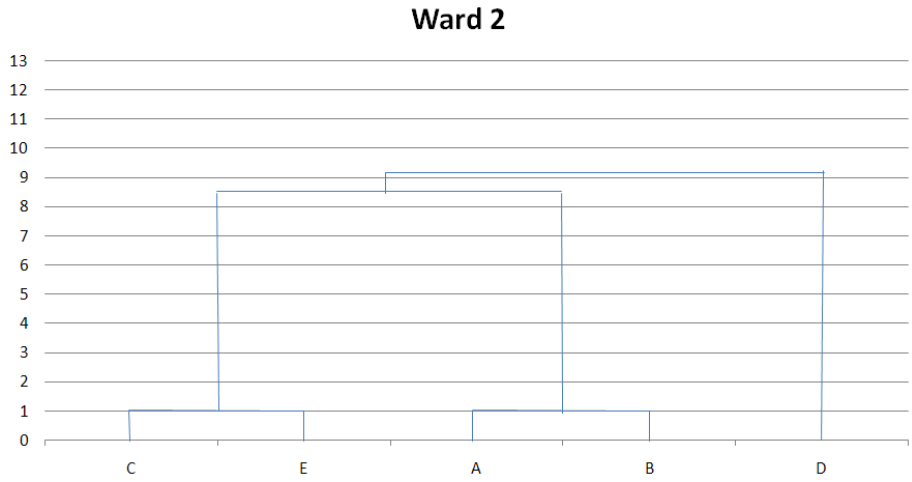

Figura 9. Dendrograma del método Ward2.

El dendrograma de la Figura 9 es el resultado de la agrupación de los clusters de la Figura 5 y las distancias obtenidas son de la unión de los cluster en cada paso de la Tabla 13, Tabla 15, Tabla 17, Tabla 19, a diferencia de los enlaces anteriores es la agrupación ya que este método primero agrupo *C*,*D* con *A*,*B* y al último *D*, debido a que este método obtiene la disimilaridad de cada cluster y va actualizando la distancia euclidiana con los nuevos atributos a diferencia de los enlaces anteriores.

2.1 Estado del arte

En esta sección se presenta la revisión de la literatura del método Ward, el cual es utilizado en diferentes áreas y para diferentes aplicaciones, por ejemplo,

científicas, medicina, biología, etc. A continuación, se describen algunos trabajos desarrollados con éste método.

En *(Govender, & Sivakumar, 2020),* Se presenta una revisión general de dos técnicas de agrupación de uso común, es decir, k-means y algoritmos jerárquicos, que se han aplicado en estudios de contaminación del aire a lo largo de 40 años aproximadamente.

Los algoritmos utilizados fueron enlace simple, promedio, Ward y centroide por parte del método jerárquico y k-means por parte del no jerárquico, al igual que la hibridación de ambos métodos, por ejemplo, Ward y k-means.

En este artículo no hace una comparación de los resultados, solo se habla de las investigaciones que utilizaron estos métodos, lo que si menciona es que el método de k-means es el más utilizado, y el segundo es Ward.

En *(Rampado et al, 2019),* la investigación que presentan consiste en analizar tomografías computarizadas, los datos se generaron utilizando el software de registro de dosis GE Healthcare, Milwaukee, WI, EE. UU, durante 1 año (2017), utilizando los métodos de agrupamiento jerárquico de enlace único, promedio, completo y de Ward y no jerárquico (K-means), la implementación fue con el software R.

Los resultados del análisis fueron favorables para el método de Ward, como se muestra en la Figura 10 de los dendrogramas, Ward obtuvo un encadenamiento mejor y entendible con dichos datos.

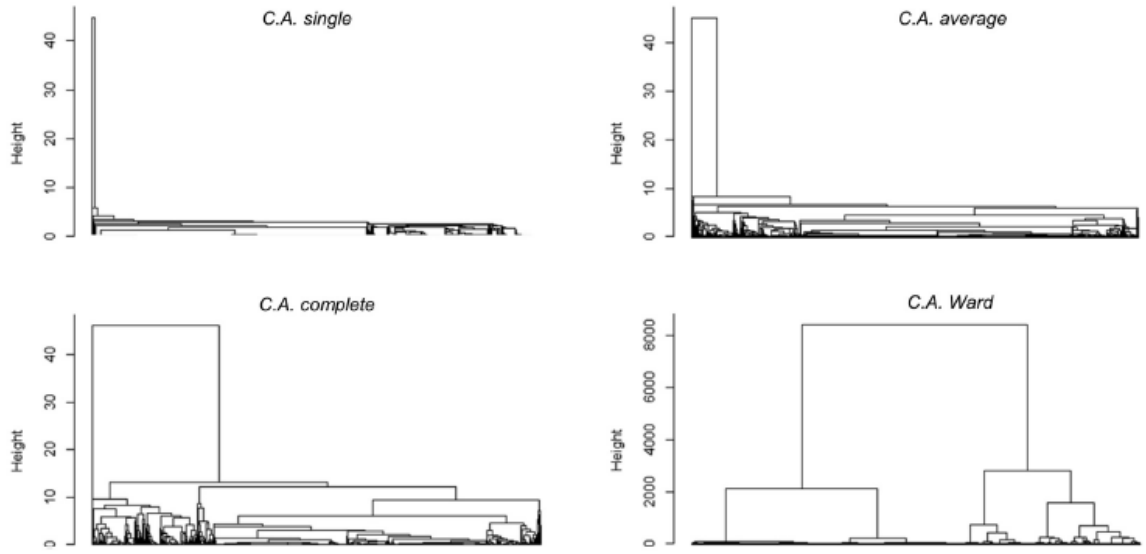

Figura 10. Dendrogramas de los 4 métodos, el que obtuvo mejores resultados de los cuatro es el método de Ward, por su mejor agrupamiento de los objetos (encadenamiento).

En *(Espinel, 2019),* se presenta el estudio del método de Ward con la herramienta del software matemático Matlab, con las funciones: pdist, linkage y dendrogram:

a) Pdist: Se obtiene la matriz de similitud con la distancia euclidiana.
b) Linkage: Se obtiene la matriz de la unión de los cluster en cada proceso.
c) Dendrogram: Se obtiene la representación gráfica del dendrograma.

La prueba consistió en 6 puntos, sus variables fueron los valores de *x*, *y*, también se habló del corte del dendrograma, el cual se dice que el óptimo es donde se da un salto brusco en la unión de dos clusters (eje y del dendrograma), los resultados se muestran en la Figura 11 y se puede observar que Matlab hace el trabajo completo, quiere decir que realiza la agrupación y presenta gráficamente el dendrograma.

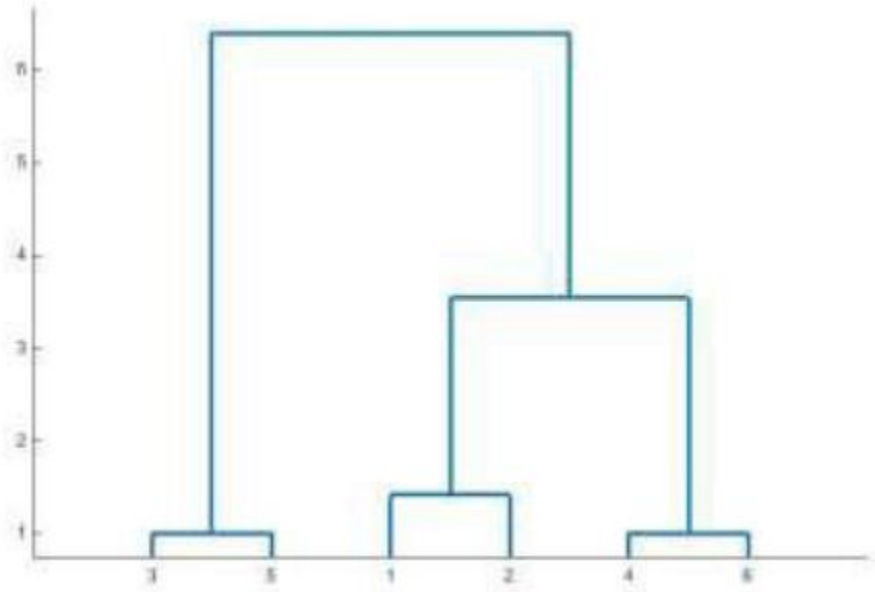

Figura 11. Dendrograma del resultado obtenido de la agrupación de los 6 puntos mediante Matlab.

En *(Dumont et al, 2018),*se muestran los resultados obtenidos del análisis de resistividad de las rocas mediante el método de Ward en la isla de la Reunión, en ella se encuentran dos volcanes, el antiguo volcán Piton des Neiges (noroeste) y el joven Piton de la Fournaise (sureste), en donde las rocas tienen diferente tipo de resistividad, y el estudio consiste en agrupar las rocas que tienen la resistividad más similar, los datos se obtuvieron mediante vuelos a lo largo de 10,400 km de líneas de vuelo durante 3 meses.

Como resultado se obtuvo la agrupación de 11 clusters cada uno con color diferente, como se muestra en la Figura 12.

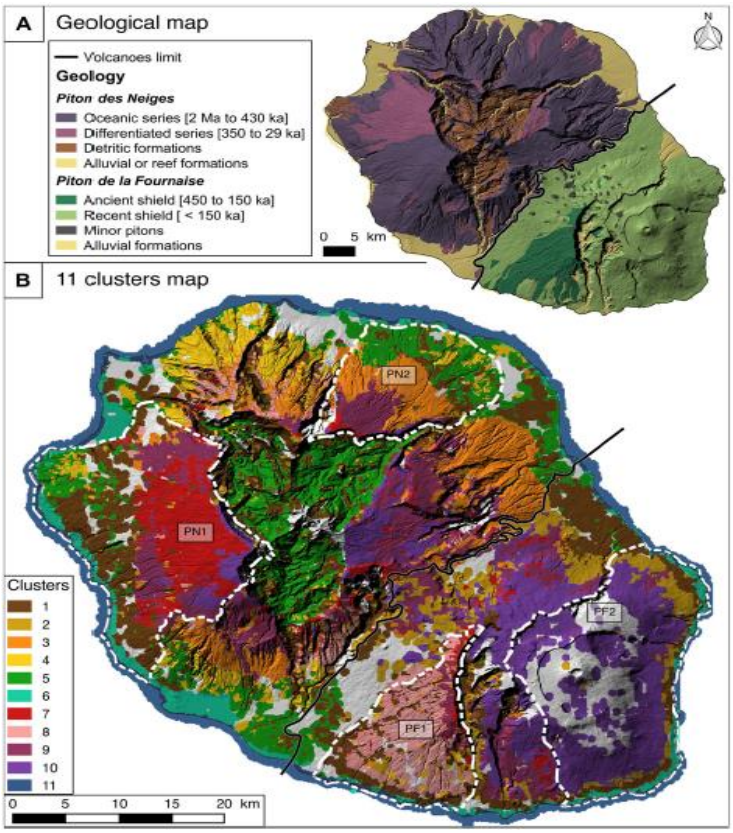

Figura 12. El mapa geológico [A] y [B]. La línea negra separa los dos volcanes. Sus actividades se simplifican en dos períodos con su edad entre corchetes [A], la dispersión de los datos de los diferentes clusters [B].

En *(Kim et al, 2018*), el objetivo de este estudio es clasificar subgrupos entre los intentos de suicidio de Corea de sur, con un total de 888 participantes, los datos se obtuvieron mediante una encuesta, la clasificación se realizó mediante el método de Ward, el criterio de agrupación cúbica (CCC), estadística *F* y estadística de *t* cuadrado se utilizaron conjuntamente para seleccionar un número adecuado de grupos, en total se obtuvieron dos grupos de agrupamiento.

En *(Balugani Kim et al, 2018),* los algoritmos k-means y el método de Ward se utilizan para agrupar elementos en grupos homogéneos para ser administrados con políticas de control de inventario uniformes, y con ello disminuir la carga de trabajo computacional en la fase de simulación que puede ser muy costosa desde el punto de vista informático.

Para la comparación de dichas técnicas de agrupamiento se utilizaron 625 elementos, en tres diferentes estructuras.

- Características fijas.
- Características igualmente espaciadas.
- Características generadas aleatoriamente.

El método de Ward fue mejor en las dos primeras estructuras, todas las características fueron normalizadas.

En *(Crouse et al, 2018),* se expone el análisis de cluster mediante el método de Ward, utilizando la desviación estándar de 9 variables a 135 pacientes para saber la heterogeneidad cognitiva (estado mental), lo cual dio 3 grupos más similares, se trabajó con el software SPSS, versión 22, IBM, Chicago, IL, EE. UU, los valores se estandarizaron.

En *(Eszergár, & Caesar, 2017*), se crearon nuevos grupos de usuarios con el método de Ward, el cual consistió en encuestas reales a usuarios de transporte. Los 5 grupos generados por el método de Ward fueron: estudiante, trabajador, turista, empresario y pensionista. Para estos grupos se consideraron los siguientes aspectos: edad, su motivación para viajar (escuela, trabajo, ocio) y su posible dificultad para viajar (discapacitados).

Se utilizó la distancia euclidiana al cuadrado utilizando los valores normalizados, y el algoritmo se ejecutó en MatLab.

En *(Adamczyk et al, 2017),* se comparó el método de Ward con redes de Kohonen, el análisis consistió en la actividad física de 10 vacas, en 3 meses.

Antes del análisis se realizó una normalización de todas las variables, para el software Statistica (Versión 12.0, 2013).

En la Figura 13 se muestra la agrupación del mes de junio utilizando la desviación estándar.

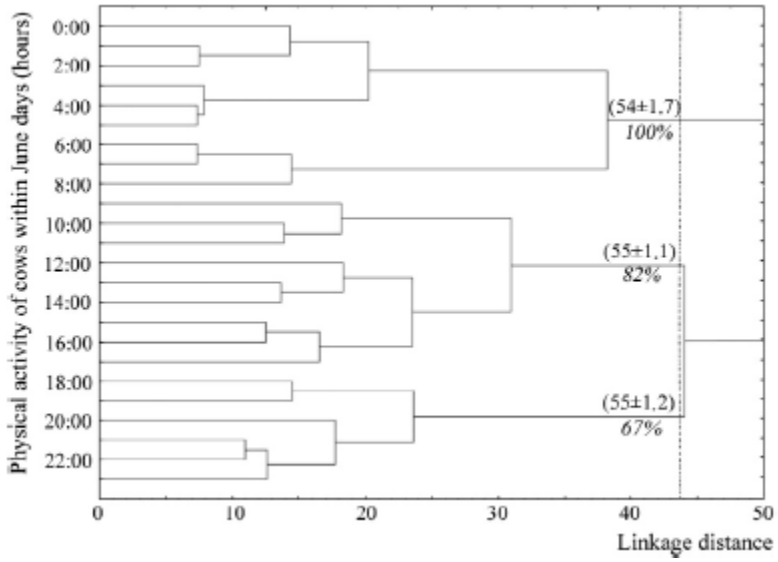

Figura 13. Resultados del mes de junio, donde se puede apreciar la actividad física de las vacas en el mes de junio, representada por un intervalo de hora.

En *(Violán et al, 2016),* Se realizó un estudio con pacientes de un centro de salud en la ciudad de Cataluña en el 2010, las edades fueron de 19 a 44 años.

"El análisis de clusters se realizó con el coeficiente de Jaccard para medir la similitud. Se utilizaron diferentes métodos jerárquicos aglomerativos de análisis de clusters. Todos los métodos menos el método Ward encadenaban sucesivamente las observaciones en un solo cluster. Finalmente, se eligió como primer método, el método Ward" (Violán et al, 2016).

En *(Heredia et al, 2012),* Se realizó un análisis para estaciones de medición de precipitación en el valle del cauca, Colombia. El objetivo es obtener las estaciones atípicas, es decir diferentes al resto, para comprobar la agrupación, posteriormente se utilizó un análisis exploratorio gráfico y cuantitativo con series univariadas para comprobar dicha agrupación.

Se utilizaron 3 diferentes enlaces, medio, de Ward y centroide (promedio), y distancia euclidiana al cuadrado como medida de similitud, todos los valores se estandarizaron.

Se analizaron 150 estaciones de precipitación y con el objetivo de ahorrar tiempo, se tomó la mejor agrupación con 75 estaciones. Para este artículo los dos mejores

enlaces son medio y promedio para el análisis de 75 estaciones o menos, pero si se aplica a un mayor número de estaciones tiene más efectividad el método Ward y el enlace promedio.

En *(Cabello, & Salama, 2012),* Se estudiaron los préstamos realizados entre los años 2000 al 2010, en Argentina. El estudio consiste en determinar si hay diferencias entre provincias de Argentina, en la participación del destino de los préstamos bancarios.

El proceso consistió en analizar los préstamos por sector, referente a cada providencia de Argentina, utilizando el método de Ward, su medida de similitud fue la distancia euclidiana al cuadrado, los datos analizados fueron 25 clusters u objetos (providencia), con 16 variables (sectores) cada uno, obtenidos por el Banco Central de la República Argentina (B.C.R.A.), cabe mencionar que se trabajó con el programa SPSS.

Los resultados obtenidos se muestran en la Figura 14.

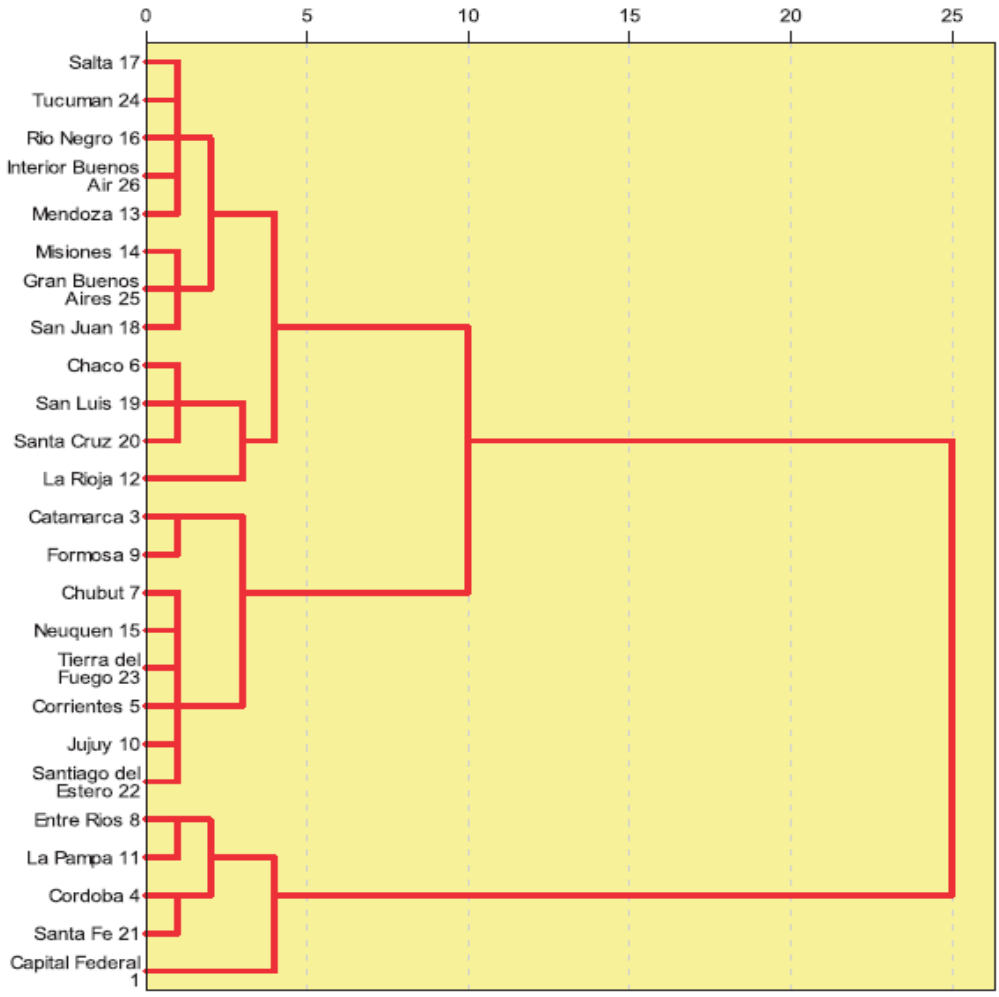

Figura 14. Dendrograma de los préstamos, los cuales se agruparon por providenciales las cuáles son más similares entre sí, según para que ocuparon los préstamos.

En *(Cargnelutti, & Guadagnin, 2011),* se presenta un estudio donde se realizó la agrupación con datos reales del maíz, con dos distancias de disimilitud (distancia euclidiana y de Manhattan) y cuatro métodos jerárquicos (enlace simple, enlace completo, enlace promedio entre el grupo y Ward).

Para comparar las agrupaciones se calculó el coeficiente de correlación cofenética (CCC). El grupo con el puntaje CCC más alto se consideró más consistente.

La consistencia del patrón de agrupación de cultivos de maíz de los métodos aumenta en el siguiente orden: Ward, enlace completo, enlace simple y enlace promedio entre el grupo.

En *(Cargnelutti et al, 2010),* se estudió la planta de frijol, con la combinación de ocho medidas de disimilitud (Euclidiana, Euclidiana estandarizada, Euclidiana promedio, Euclidiana promedio estandarizada, Distancia cuadrada de Euclidiana, Cuadrado de distancia euclidiana estandarizada, Mahalanobis y Mahalanobis estandarizado) y ocho métodos dela agrupación jerárquica (enlace simple, enlace completo, Ward, mediana, enlace medio dentro del grupo, enlace medio entre el grupo, Gower y Centroide).

En *(Solano et al, 2008),* se utilizó el método de Ward para agrupar 19 riesgos y 19 lesiones en accidentes laborales, los datos analizados están estandarizados y con una distribución normal, la distancia utilizada fue Chevichev.

En *(Chavent et al, 2007),* se comparan 3 algoritmos DIVCLUS-T, Ward y k-means, los tres algoritmos utilizan la minimización del criterio de inercia, la comparación fue con seis bases de datos del repositorio de UCI Machine Learning, utilizando la distancia euclidiana. Como resultado se obtuvo que para grandes números de objetos el mejor algoritmo es Ward y k-means.

En *(Rodríguez, & Esquivel, 2005),* trabajaron los datos de 14 yacimientos, cada uno con 46 variables en dos diferentes tipos de tiempo (edad de cobre y del bronce).

Para la medida de similaridad se utilizó la distancia Chi-cuadrado y el método de Ward, el cual agrupó correctamente los yacimientos más similares.

Los resultados se muestran en la Figura 15.

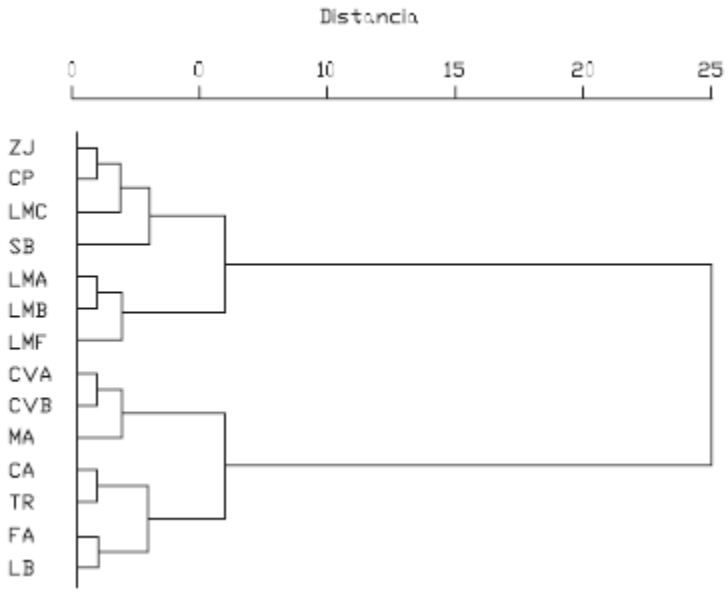

Figura 15. Dendrograma de los 14 yacimientos, que se obtuvo dos agrupaciones (K=2), según su similaridad.

En *(Díaz, & Mormeneo, 2002),* se analizó la temperatura y precipitación de la región pampeana, con 44 estaciones del servicio meteorológico. La información obtenida fue mensual por 30 años de cada estación.

En total fueron 44 objetos y 360 atributos, se utilizó la distancia euclidiana al cuadrado, para cada estación (objeto) se utilizó el promedio, desviación estándar y el coeficiente de correlación lineal de pearson.

El método de Ward superó a otros 7 métodos, jerárquico o difusos, para la zonificación propuesta.

Los resultados se muestran en la Figura 16.

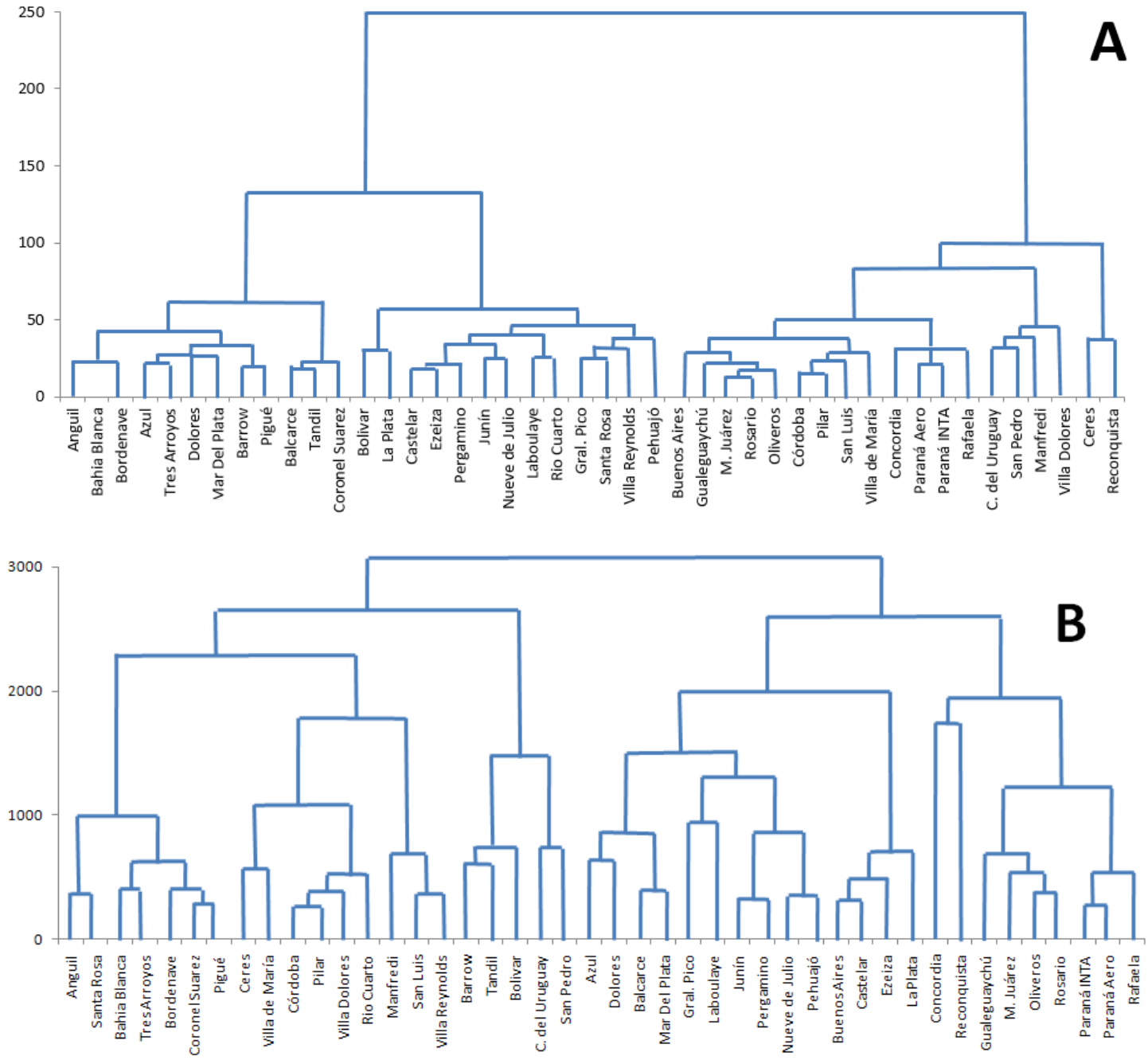

Figura 16. Dendrograma del análisis obtenido de las A) temperaturas, y B) precipitaciones, de la región pampeana.

CAPÍTULO III. DESARROLLO

3.1 Descripcióndel algoritmo utilizado.

En la Figura 17 se muestran los componentes principales del algoritmo desarrollado.

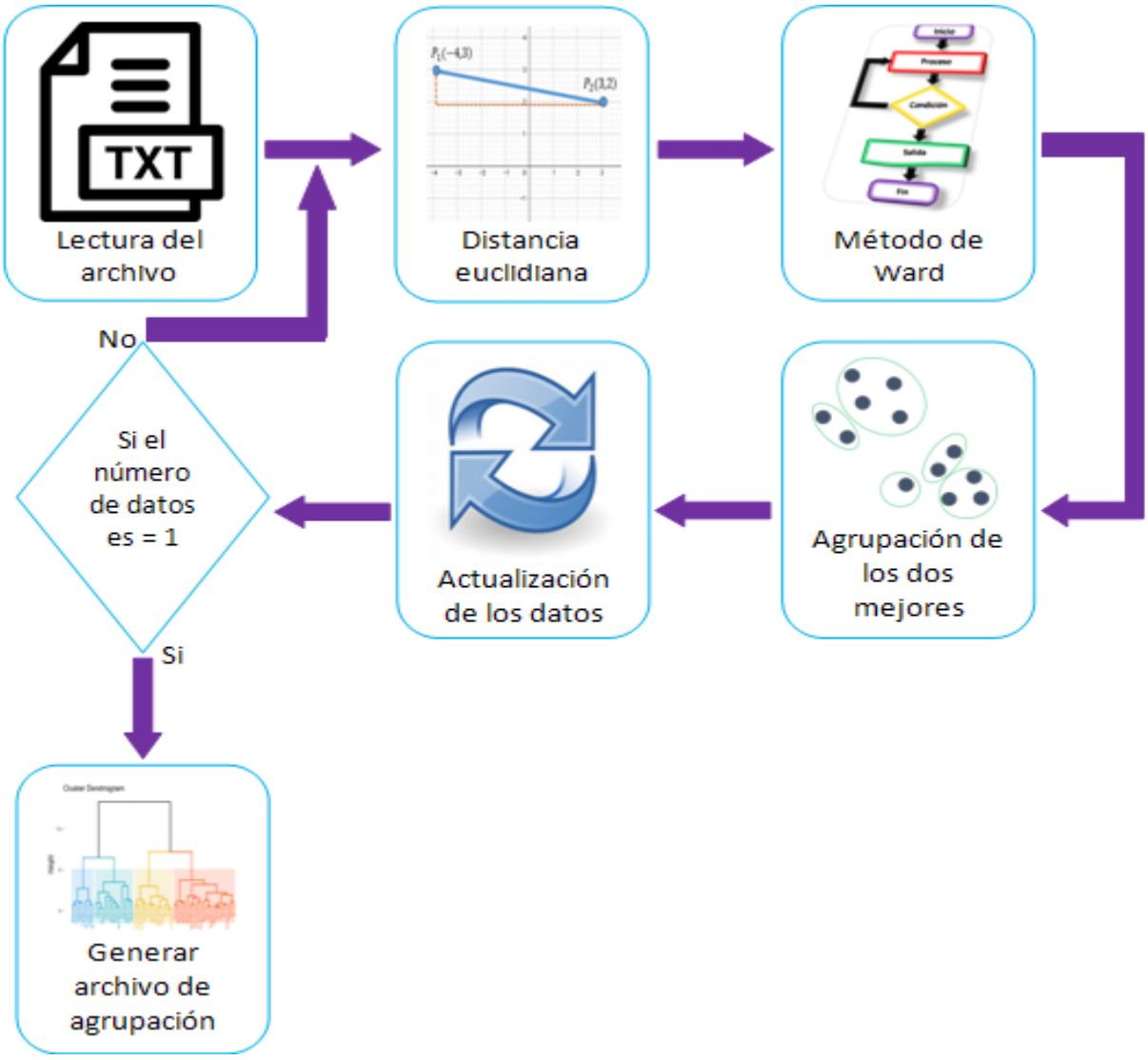

Figura 17. Proceso del algoritmo jerárquico-aglomerativo.

Esta representación del algoritmo es en general, ya que cada recuadro es otro subproceso, el algoritmo se estará ejecutando mientras *n* sea mayor a 1 (*n* es el

número total de clusters a agrupar). En las siguientes subsecciones se detalla cada subproceso.

3.1.1 Lectura del archivo.

Este módulo, lee el archivo .txt para obtener los datos que contiene, por ejemplo, número de clusters, número de atributos, cardinalidad, con ello se llena como el ejemplo de la Tabla 12 que es la matriz de atributos, este módulo sólo se ejecuta una vez.

Figura 18. Módulo de lectura de archivo txt.

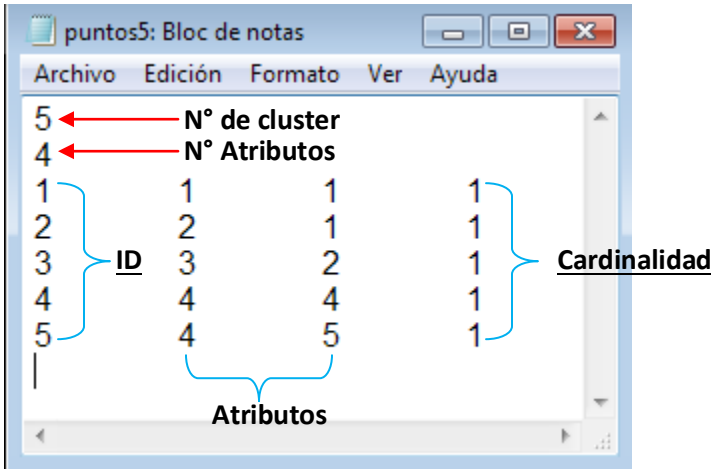

Figura 19. Estructura del archivo de entrada.

En la Figura19 el primer valor de arriba hacia abajó es el número de cluster u objetos (5), el que sigue es el número de atributos (4), posteriormente son los clusters con sus datos, por ejemplo, el primero

número de izquierda a derecha es el id del cluster, los siguientes valores a la derecha son los atributos menos el último ese es la cardinalidad.

3.1.2 Distancia euclidiana.

En este módulo se obtiene la distancia euclidiana con los datos de la matriz de atributos como la Tabla 12 del ejemplo e Ward2, y el proceso se muestra en la Figura 20 y Figura 21.

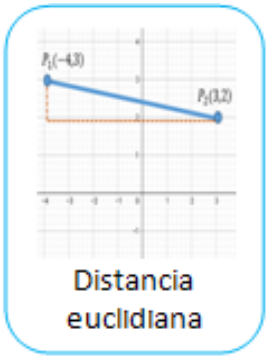

Figura 20. Módulo de distancia euclidiana.

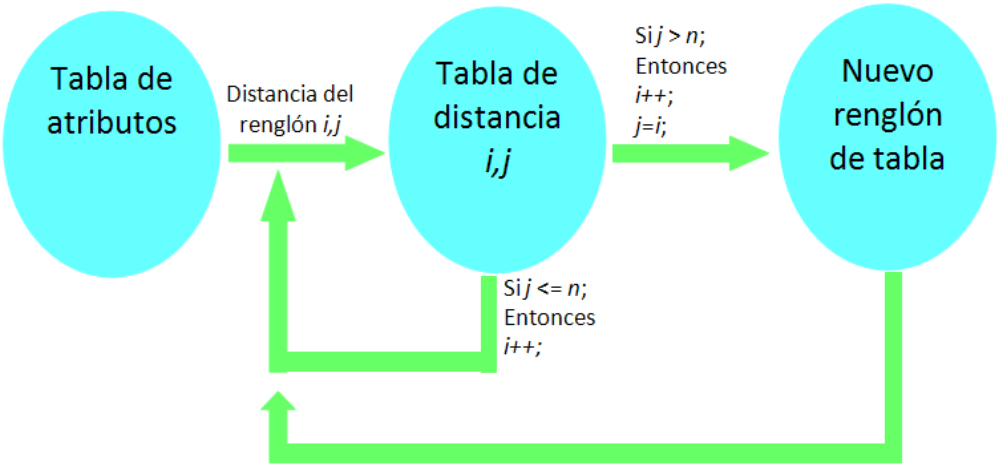

Figura 21. Diseño de alto nivel de la distancia euclidiana.

Para la distancia euclidiana se comienza con la matriz de los atributos un ejemplo es la Tabla 12 del ejemplo de Ward 2, se comienza con el primer cluster por lo tanto siempre $i = 1$, y $j = 1$, ya que se tiene la

distancia del primer cluster se agrega a la tabla de distancia, después se compara si $j < = n$ entonces j ++ que sería $j = 2$, esto se va a repetir hasta que $j > n$, cuando esto se cumple entonces se cambia a otro renglón nuevo tanto para la tabla de atributos y de distancia, i ++ que sería $i = 2$ y $j = 2$, hasta que $i = n$, y con ello se va ir llenando la matriz de distancia euclidiana, un ejemplo de la matriz es la Tabla 13 del ejemplo de Ward2.

Cabe mencionar que esta distancia es no cuadrada y tiene una complejidad de n^2.

3.1.3 Método de Ward.

Para este módulo se requiere de dos tablas, la de atributos como la Tabla 12 del ejemplo de Ward2, y la de la distancia euclidiana Tabla 13 del ejemplo de Ward2, posteriormente se obtiene la disimilaridad de cada uno de los clusters con el resto con la expresión 7. En la Figura 23 se muestra el diseño de alto nivel del método de Ward del módulo de la Figura 22.

Figura 22. Modulo del método de Ward.

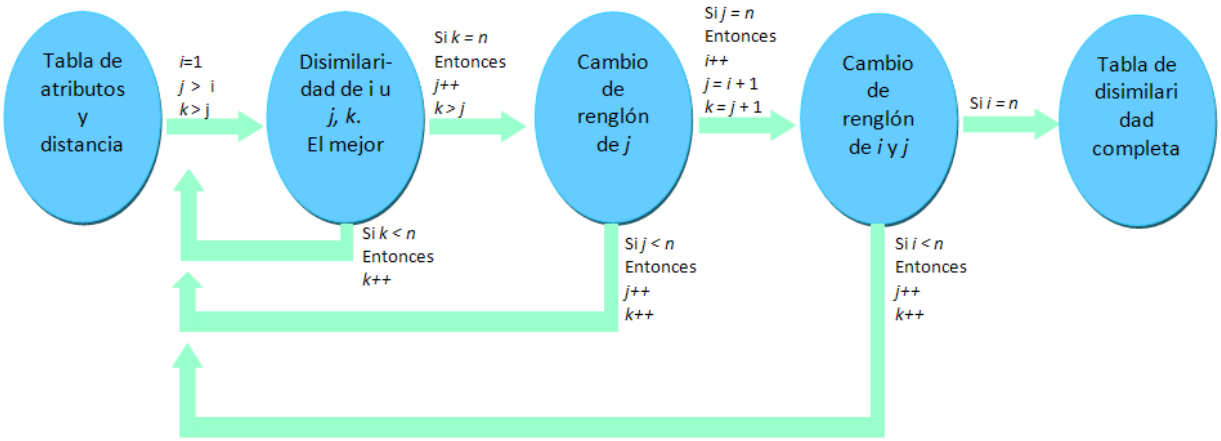

Figura 23. Diseño de alto nivel del método de Ward.

En la Figura 23 se muestra el proceso que consiste en obtener la disimilaridad del cluster 1 ∪ cluster 2 tal que cluster 3, esto es igual a la expresión "*i* ∪ *j*, *k*" sería el paso 1, con la expresión 7, y se verifica *k*, si es igual a *n*, entonces aumenta *j* siempre y cuando *j* sea menor o igual a *n*, si *j* llegara a *n*, entonces *i* aumentaría, de lo contrario solo aumentaría *k*, entonces quedaría cluster 1 ∪ cluster 2 tal que cluster 4, esto quiere decir que se obtendrá la disimilaridad del cluster 1 y cluster 2, con la distancia del cluster 4, este proceso termina hasta que *i* = *n*.

Este módulo tiene la complejidad más alta ya que cada cluster se compara con el resto doble vez, y esto lo hace muy tardado con un número de objetos grande, la complejidad es n^3, por lo tanto, este módulo es el que se pretende paralelizar, para disminuir la complejidad.

3.1.4 Agrupación de los dos mejores.

En este módulo se seleccionan a los dos mejores clusters con mayor disimilaridad obtenidos del módulo anterior (Método de Ward), en la Figura 24.

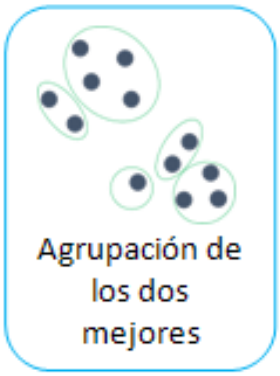

Figura 24. Módulo de la selección de los dos mejores.

3.1.5 Actualización de los datos.

En este módulo se actualizan los atributos de los dos clusters seleccionados como mejor con la expresión 6, para que sólo quede un único cluster, como muestra en la Figura 25 y Figura 26.

Figura 25. Módulo de actualización de atributos.

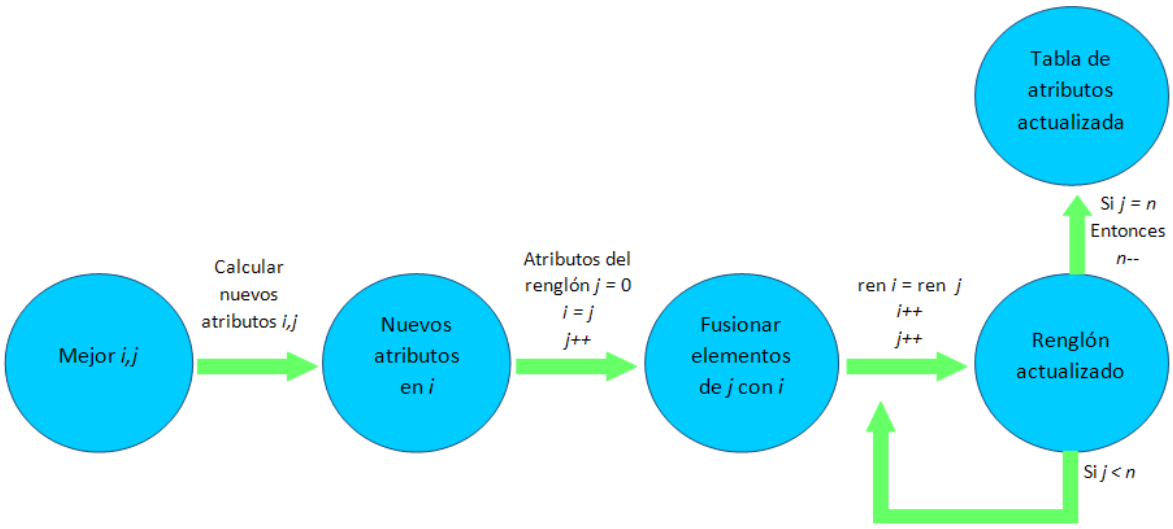

Figura 26. Diseño de alto nivel de la actualización de los atributos.

Se comienza con los dos clusters seleccionados con mayor disimilaridad, posteriormente se actualizan los atributos con la expresión 6 de los dos clusters seleccionados pero los datos se escriben en el primer cluster, después los atributos del segundo cluster seleccionado toman como valor 0, por último todo los cluster siguientes se van recorriendo una casilla hacia atrás, esto quiere decir que si el segundo cluster seleccionado es el 2, entonces el cluster 3 se recorre a la casilla del cluster 2 y al número de cluster se le resta 1.

Para ejemplificar lo anterior supongamos que el valor total de clusters es 5, y se agrupan cluster 1 y 2, en el cluster 1 se actualizan los atributos, luego los valores del cluster 2 toman el valor 0, y el cluster 3 se recorre al 2 y así sucesivamente hasta llegar al cluster 5, posteriormente se elimina un renglón, entonces el valor 5 se le resta 1 y es 4.

Este módulo tiene una complejidad de n^2.

3.1.6 Verificar el número de cluster.

En este módulo se verifica si el número de cluster es mayor a 1, si esto es verdad el proceso se repite desde la distancia euclidiana, de lo contrario termina el proceso, y se genera el archivo txt con la agrupación.

En la Tabla 20 se muestra la complejidad del algoritmo.

Tabla 20. Complejidad del algoritmo de Ward.

PROCESO	COMPLEJIDAD
ALGORITMO	N
DISTANCIA EUCLIDIANA	n^2
DISIMILARIDAD (WARD)	n^3
NUEVOS ATRIBUTOS	n^2

La complejidad en tiempo de ejecución $T(n)$, es el tiempo que se tarda en ejecutar un algoritmo sus instrucciones, para obtener la complejidad total del algoritmo se multiplica toda la complejidad del algoritmo, por ejemplo:

$T(n)=n\ (n^2 + n^3 + n^2)$
$T(n)=n^4$

Teniendo una complejidad alta con mayores objetos, esto se tardaría mucho, por eso, el objetivo de esta investigación es poder disminuir esta complejidad.

3.2 Programación paralela.

En esta sección, se describe la implementación paralela para el algoritmo jerárquico aglomerativo con el método de Ward.

La programación en paralelo tiene como objetivo descomponer el problema en tareas, para que cada tarea se ejecute en un procesador de manera simultánea (Navarro et al, 2013).

El algoritmo se divide en tareas y las tareas se envían a los procesadores para que se ejecutan de forma simultánea, como se muestra en la Figura 27.

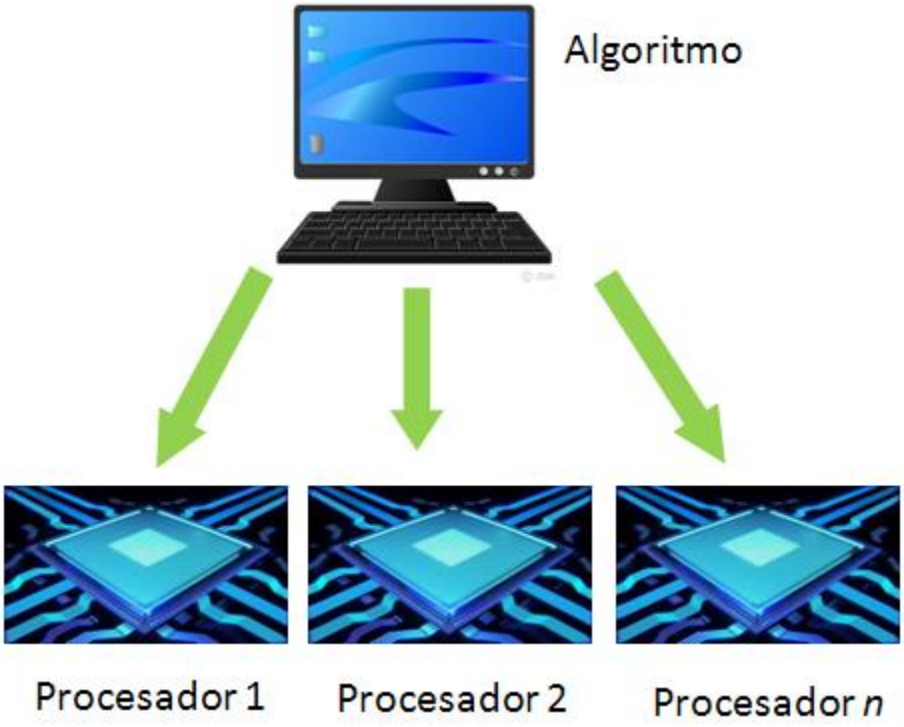

Figura 27. Proceso de programación paralela.

3.3 Modelo de programación paralela.

El modelo utilizado es el de maestro-esclavo(master/slave), el cual consiste en un nodo maestro que genera y asigna trabajo, generalmente llamado tareas, a *N* nodos esclavos que reciben las tareas, las procesan y devuelven un resultado. Por lo tanto, el nodo maestro es responsable de recibir y analizar los resultados calculados por los nodos esclavos (Baldo et al, 2004). Como se muestra en la Figura 28.

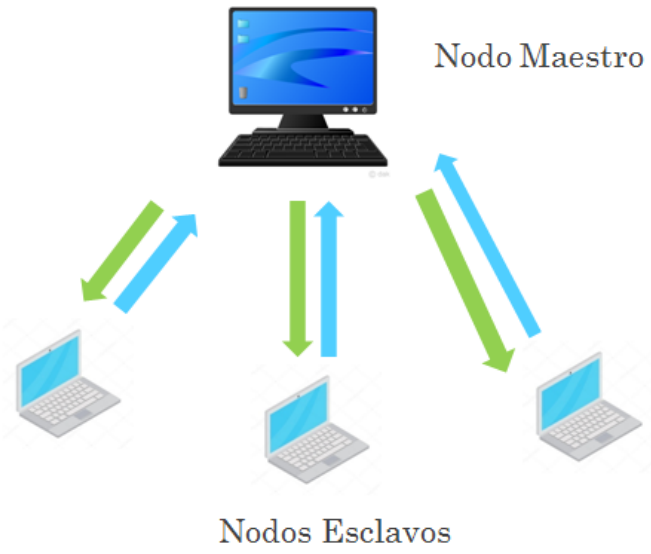

Figura 28. Proceso de nodo Maestro-Esclavo.

Este paradigma es generalmente adecuado para plataformas de transmisión de mensajes o de memoria compartida, ya que la interacción es naturalmente bidireccional (Baldo et al, 2004).

Según cómo el maestro distribuya las tareas y recopile los resultados de los esclavos, se presentan dos tipos diferentes de implementaciones master/slave: programas con interacción síncrona o asíncrona. Se dice que un programa tiene interacción síncrona cuando las tareas deben realizarse en fases, es decir, todas las tareas de cada fase deben finalizar antes de la distribución de las tareas de la siguiente fase. Por el contrario, la interacción entre el maestro y los esclavos se denomina asíncrona cuando el maestro distribuye nuevas tareas cada vez que un esclavo termina su cálculo. Dicha interacción puede reducir el tiempo de espera considerando una distribución no determinista de solicitudes de esclavos (las tareas tienen un costo computacional desconocido). En este contexto, el nodo maestro generalmente implementa un búfer de espera para hacer frente a las llegadas simultáneas de resultados de los esclavos (Baldo et al, 2004).

3.4 Algoritmo paralelo.

El algoritmo paralelo se desarrolló mediante el modelo Maestro-esclavo, utilizando la biblioteca MPI (“Message Passing Interface”), (“Interfaz de Paso de Mensajes”) para el paso de mensajes. Acontinuación, se muestra el reparto de tareas de nuestro modelo Maestro-esclavo implementado:

Nodo maestro:

- Se ejecuta en Ioevolution
- Lectura de datos
- Cálculo de distancia euclidiana
- Distribución de tareas (**MPI_Bcast**)
- Recepción de resultados (**MPI_Recv**)
- Encontrar la mejor agrupación
- Actualizar atributos

Nodo esclavo:

- Recepción de datos (**MPI_Recv**)
- Implementar método Ward a los datos recibidos
- Retornar el resultado del método Ward (**MPI_Send**)

Para entender mejor el proceso, en la Figura 29, se presenta el diseño de alto nivel paralelo, en la Figura 30 el diagrama de flujo y Figura 31 el pseudocódigo.

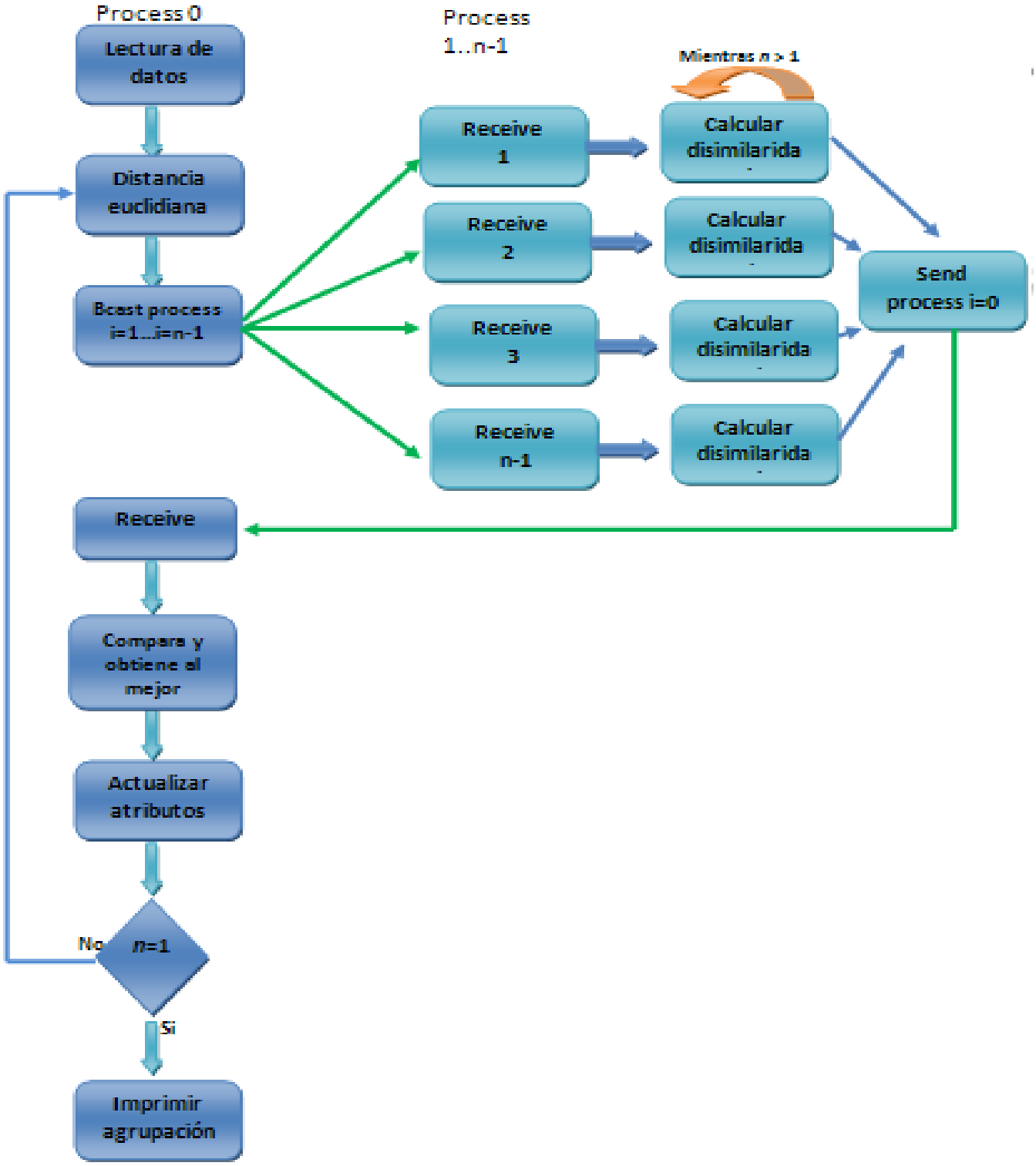

Figura 29. Diseño de alto nivel paralelo.

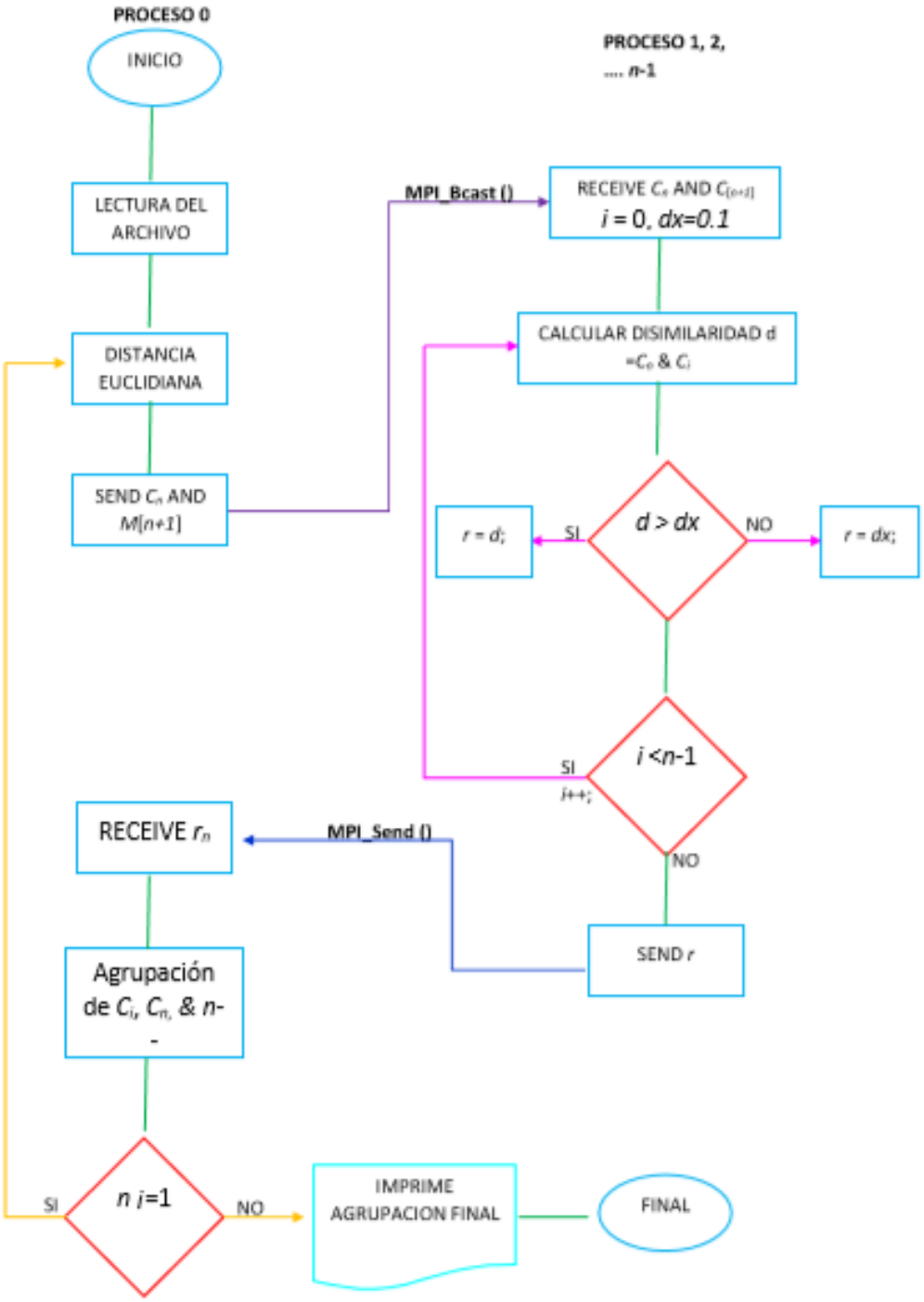
PROCESO 0
INICIO
LECTURA DEL ARCHIVO
DISTANCIA EUCLIDIANA
SEND C_n AND $M[n+1]$
MPI_Bcast ()
PROCESO 1, 2, ... n-1
RECEIVE C_n AND $C_{[n+1]}$
$i = 0$, $dx=0.1$
CALCULAR DISIMILARIDAD d =C_o & C_i
$d > dx$
SI
NO
$r = d$;
$r = dx$;
$i < n-1$
SI
i++;
NO
SEND r
MPI_Send ()
RECEIVE r_n
Agrupación de C_i, C_n, & n-
n i=1
SI
NO
IMPRIME AGRUPACION FINAL
FINAL

Figura 30. Diagrama de flujo del algoritmo paralelo.

```
1   Algoritmo de Ward paralelo.
2   Nodo Maestro
3   mientras n > 1
4           Leer_archivo ();
5           n = número total de cluster;
6           Distancia_euclidiana ();
7           Enviar_nodos ();
8           Recibir_nodos_esclavos ();
9           Obtener_centroide ();
10          n--;
11  Fin mientras
12  Imprimir_agrupacion ();
13
14  Nodos esclavos
15  Recibir_nodo_maestro ();
16  Valor1 = 0, Valor2 = Valor1 + 1, Valor2 = Valor1 + 1;
17  Do
18          Do
19                  Si Valor1 != Valor2 && Valor3  != Valor2 && Valor3  != Valor1 entonces
20                          Disimilaridad ();
21                  Fin Si
22                  Valor3 ++;
23          Fin mientras (Valor3 < n);
24          Valor2 ++;
25          Valor3 ++;
26  Fin mientras (Valor2 < n)
27  Enviar_nodo_maestro ();
28  Fin del Algoritmo
```

Figura 31. Pseudocódigo del algoritmo paralelo.

Como se puede apreciar en la Figura 29 del diseño de alto nivel, Figura 30 diagrama de flujo y Figura 31 del pseudocódigo, el nodo maestro o proceso 0, se encarga de la lectura de los datos del archivo txt, el cual contiene la información de los clusters como su id, atributos, etc, posteriormente con esa información se obtiene la distancia euclidiana, la cual consta de obtener la distancia del punto A al punto B, esto es con todos los cluster a analizar, esta tabla se envía del nodo maestro a los nodos esclavos mediante MPI_Bcast el cual consiste en enviar los datos de un nodo al resto de los nodos conectados al mismo tiempo, cada nodo esclavo recibe los datos con el estándar MPI_Recv que sirve para recibir los datos de otro nodo, con esos datos cada nodo procesa el método de Ward, pero solo de un cluster, es decir, que el procesador 1 o nodo 1 solo va a obtener la disimilaridad

del cluster 1 con el resto de clusters y cluster 2 con el resto, por lo tanto, se requiere el mismo número de objetos que de procesadores, todo simultáneamente y cada uno va guardando el mejor (con mayor disimilaridad). Cuando cada procesador o nodo esclavo termina, envía su resultado (mejor) al nodo maestro con el estándar MPI_Send, que se utiliza para enviar los datos de un nodo a otro, el nodo maestro recibe todos los resultados con MPI_Recv y compara cuál de esos datos tiene la mayor disimilaridad, cuando obtiene la mejor agrupación actualiza los atributos de ambos clusters y elimina un cluster, compara si hay más de dos clusters, si esto es verdadero repite el proceso desde la distancia euclidiana, de lo contrario se termina el proceso y el nodo maestro genera el archivo txt con la agrupación.

El lenguaje de programación utilizado para el desarrollo secuencial y paralelo fue en C.

3.5 Cluster Ioevolution.

La programación paralela se realizó en el cluster Ioevolution, que se encuentra en la Universidad Autónoma del Estado de Morelos en la Facultad de Contaduría Administración e Informática, el cual cuenta con cinco servidores y cada servidor cuenta con un límite de procesadores como se muestra en la Tabla 21.

Tabla 21. El número de procesadores por servidor que se pueden utilizar.

Cluster Ioevolution					
Máquinas	Procesador	Procesadores Disponibles	Núcleos por Procesador	Núcleos disponibles	Hilos Disponibles
Ioevolution	Intel(R) Xeon (R) CPU @ 3.40 Ghz	8	4	32	256
compute-0-0	Intel(R) Xeon(R) CPU E5645 @ 2.40 Ghz	12	6	72	864
compute-0-1	Intel(R) Xeon(R) CPU X3430 @ 2.40 Ghz	4	4	16	64

compute-0-2	Intel(R) Xeon(R) CPU X3430 @ 2.40 Ghz	4	4	16	64
	TOTAL	**28**	**14**	**136**	**1248**

En la Tabla 21 se muestra el número total de procesadores que se dispone para poder ejecutar el algoritmo, pero cada procesador dispone de 4 a 6 núcleos, posteriormente se multiplica por el total de procesadores disponibles por servidor, y tenemos los núcleos totales, pero cabe mencionar que cada núcleo tiene 4 hilos para ejecutar o hacer tareas por lo que tenemos un total de hilos disponibles de 1,248, lo cual nos da un número mayor de procesadores para poder ejecutar mayor número de objetos.

Esto sirve para tener un balance de carga equitativo y así no sobre cargar los servidores.

CAPÍTULO IV. EXPERIMENTACIÓN Y RESULTADOS

4.1 Experimentación.

La experimentación se realizó con las instancias de la literatura y benchmark clustering, que se muestran en la Tabla 22.

Tabla 22. Instancias utilizadas para el algoritmo de Ward, secuencial y paralelo.

Autor	N° de objetos
Gallardo	5
Elaboración propia	10
Elaboración	57

propia (maíz)	
Chang	312

La instancia de Gallardo tiene 5 objetos, como se muestra en la Figura 32.

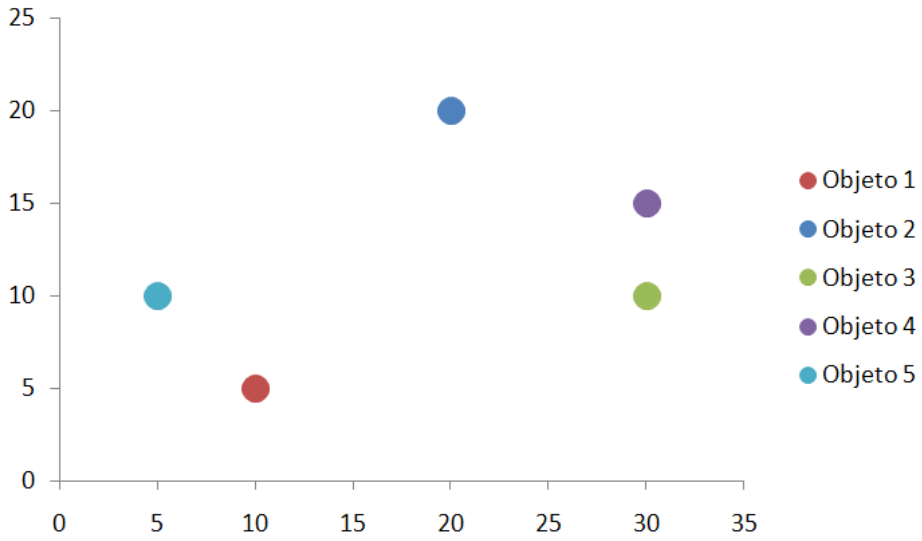

Figura 32. 5 objetos a agrupar de la instancia de Gallardo.

La elaboración de 10 objetos fue propia para trabajar con un mayor número de datos, los cuáles son aleatorios, como se muestra en la Figura 33.

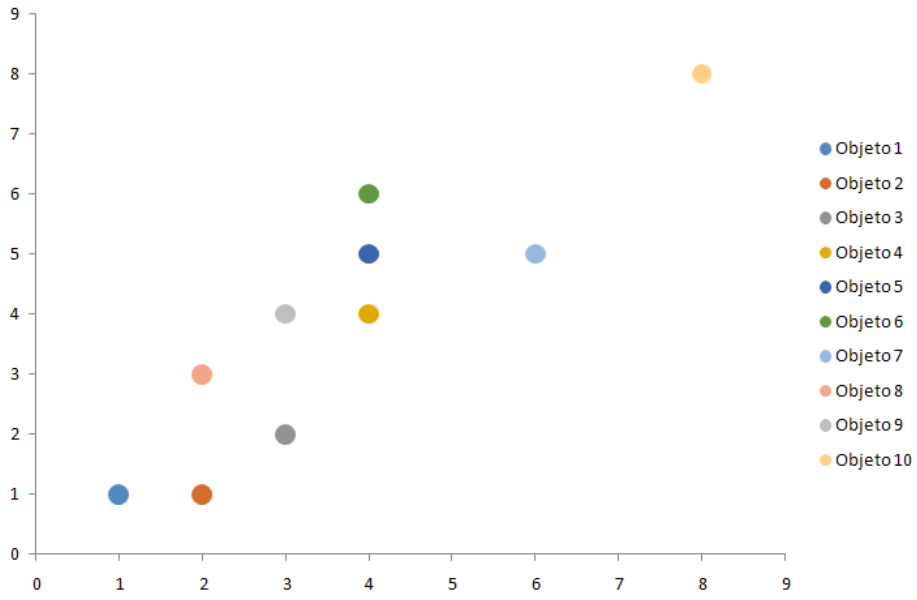

Figura 33. 10 objetos a agrupar de elaboración propia.

Los 57 objetos de maíz fueron tomados de (Cargnelutti, & Guadagnin, 2011), los cuales se dieron aleatoriamente mediante la distribución normal, ya que el artículo solo da la desviación estándar y la media, como se muestra en la Figura 34.

Figura 34. Figura de los 57 objetos a agrupar del maíz cada color es atributos similares.

Por último, de (Chang & Yeung, 2008) se obtuvieron del benchmark clustering, el cual está constituido por tres espirales de diferente color y tiene 312 objetos en total, como se muestra en la Figura 35.

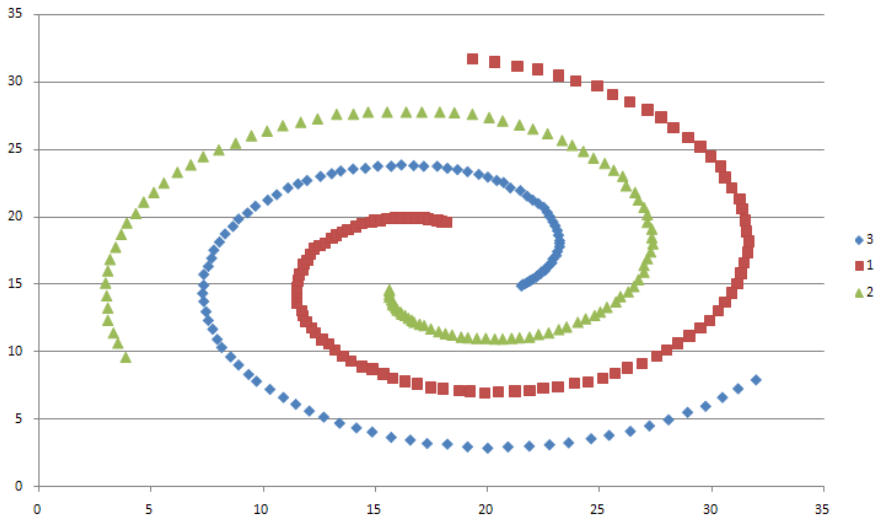

Figura 35. 312 objetos en forma de espiral.

En cada resultado se va a obtener la aceleración (Speedup) S(p) y la eficiencia E(p). S(p) es la comparación entre el tiempo secuencial contra el tiempo de un algoritmo paralelo para el mismo problema (Zavala-Díaz, 2013) y "E(p) de un algoritmo paralelo define la relación entre la aceleración S(p) y el número de procesadores *p*. Este factor indica qué tan bien son utilizados los procesadores" (Zavala-Díaz, 2013).

En este trabajo el número de *p* es igual al número de objetos, ya que cada objeto tiene un procesador.

4.2 Resultados.

Los resultados que se muestran en las Tablas 23-26, son la comparación del tiempo secuencial vs paralelo, el secuencial se corrió en Ioevolution, con procesador Intel(R) Xeon (R) CPU @ 3.40 Ghz.

La eficiencia en cada resultado se refiere a la eficiencia del algoritmo paralelo, donde 100% es mejor y entre menor tenga de porcentaje no es muy bueno para esa instancia.

Tabla 23. Resultados de la instancia de 5 objetos.

Gallardo	
Secuencial	**Paralelo**
0.000354 segundos	0.002607 segundos

En la instancia de Gallardo que se muestra en la Tabla 23, el tiempo de ejecución del algoritmo secuencial es: 0.000354 y el paralelo de 0.002607 segundos, por lo tanto, la aceleración S(p) del algoritmo paralelo es 0.000354 / 0.002607= 0.135 y la eficiencia E(p) es 0.135 / 5= 0.0271 * 100 = 2.71, es decir la implementación paralela es 2.71% eficiente para esta instancia, como se puede observar para esta

instancia no es muy buena la eficiencia debido al número de objetos, el dendrograma se muestra en la Figura 36.

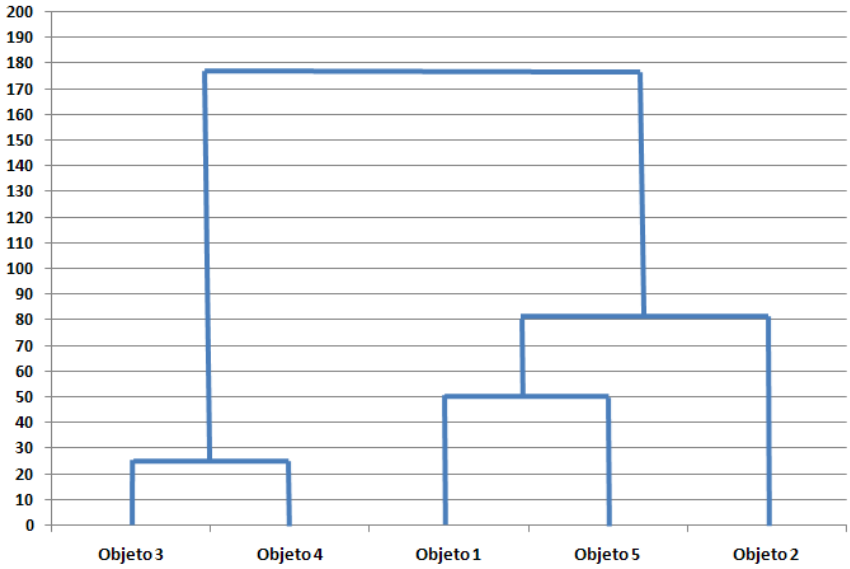

Figura 36. Dendrograma de los 5 objetos de la instancia de Gallardo.

El dendrograma de la Figura 36 presenta una agrupación correcta según la distribución de los objetos, esto con base en lo que se observa en la Figura 32.

Tabla 24. Resultados de la instancia de 10 objetos.

Elaboración propia	
Secuencial	**Paralelo**
0.000458 segundos	0.008991 segundos

En la instancia de 10 objetos elaboración propia que se muestra en la Tabla 24, el tiempo de ejecución del algoritmo secuencial es: 0.000458 y el paralelo de 0.008991 segundos, por lo tanto, la aceleración S(p) del algoritmo paralelo es 0.000458 / 0.008991= 0.0509 y la eficiencia E(p) es 0.0509 / 10= 0.0050 * 100 = 0.50, es decir la implementación paralela es 0.50% eficiente para esta instancia,

como se puede observar para esta instancia no es muy buena la eficiencia debido al número de objetos, el dendrograma se muestra en la Figura 37.

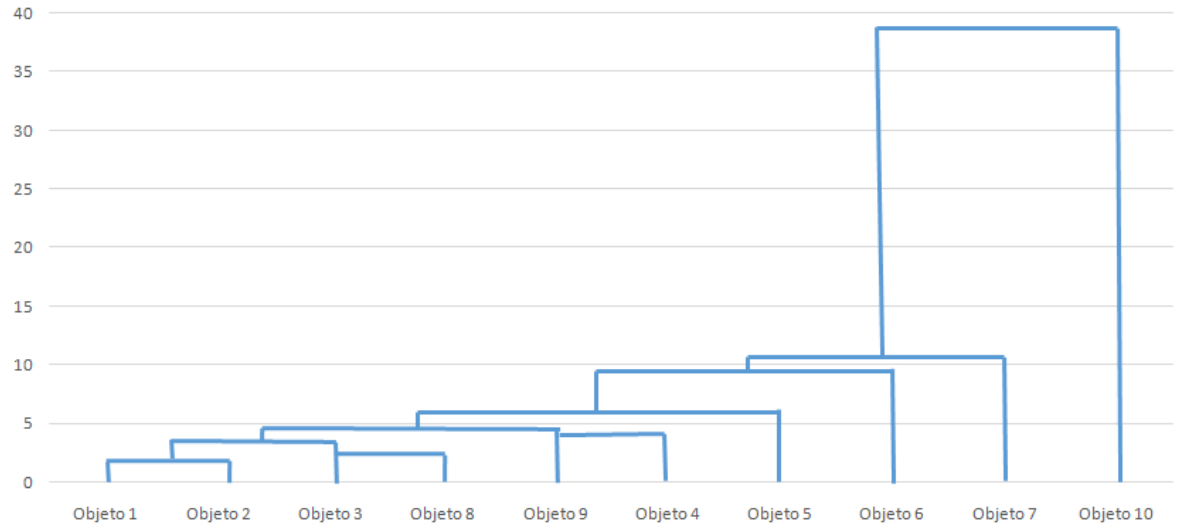

Figura 37. Dendrograma de la agrupación de los 10 objetos.

Tabla 25. Resultados de la instancia de 57 objetos representación del maíz.

Elaboración propia (maíz)	
Secuencial	**Paralelo**
0.560161 segundos	1.095899 segundos

En la instancia de Cargnelutti, & Guadagnin, que se muestra en la Tabla 25, el tiempo de ejecución del algoritmo secuencial es: 0.560161y el paralelo de 1.095899 segundos, por lo tanto, la aceleración S(p) del algoritmo paralelo es 0.560161/ 1.095899= 0.511 y la eficiencia E(p) es 0.511 / 57= 0.0089 * 100 = 0.896, es decir la implementación paralela es 0.896% eficiente para esta instancia, como se puede observar para esta instancia no es muy buena la eficiencia debido al número de objetos, el dendrograma se muestra en la Figura 38.

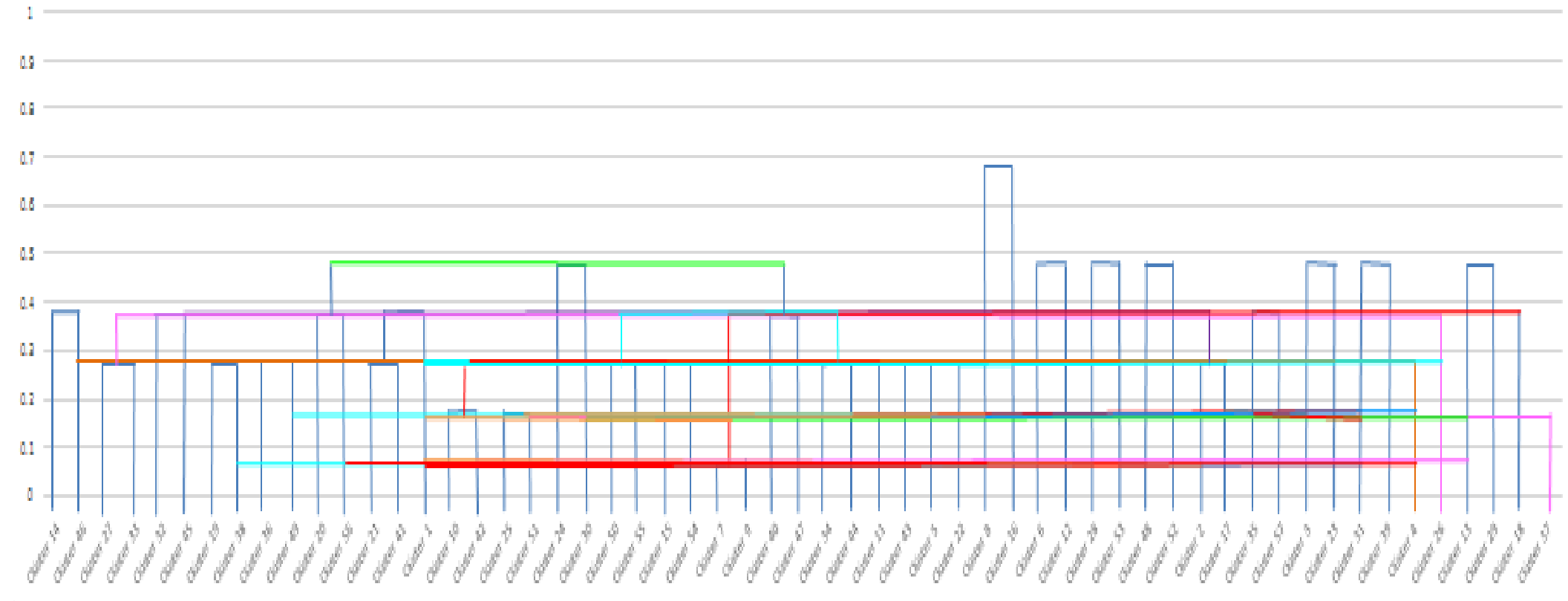

Figura 38. Dendrograma de la agrupación de los 57 objetos de maíz.

Tabla 26. Resultados de la instancia de 312 objetos, en forma de espiral.

Chang	
Secuencial	**Paralelo**
478.261963segundos	13.09 segundos

En la instancia de Chang & Yeung, que se muestra en la Tabla 26, el tiempo de ejecución del algoritmo secuencial es: 478.261963 y el paralelo de 13.09 segundos, lo tanto la aceleración S(p) del algoritmo paralelo es 478.261963 / 13.09 = 36.53 y la eficiencia E(p) es 36.53 / 312 = 0.117 * 100 = 11.71, es decir la implementación paralela es 11.71% eficiente para esta instancia, como se puede observar para esta instancia es mejor la eficiencia debido al número de objetos, el dendrograma se muestra en la Figura 39.

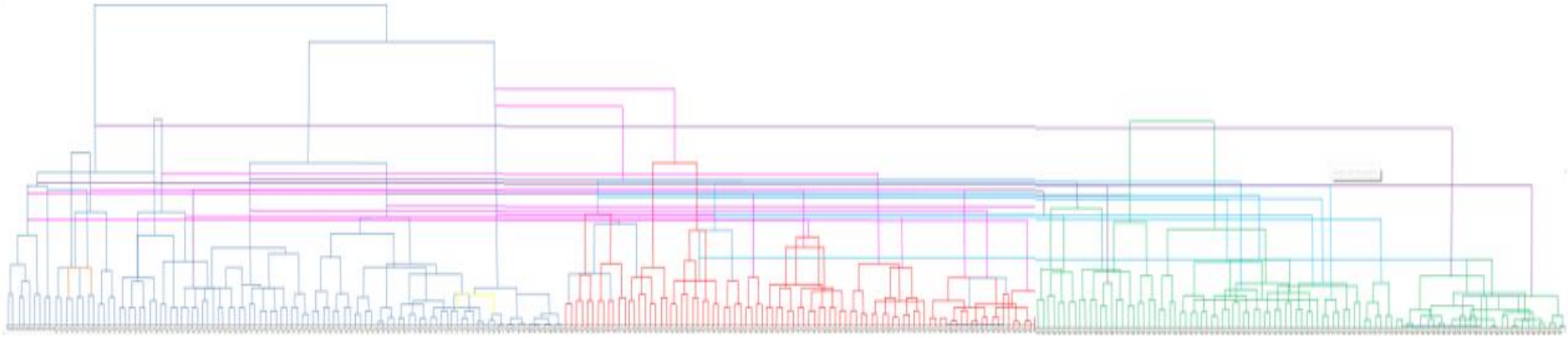

Figura 39. Dendrograma de la agrupación de los 312 objetos de la espiral.

En la Figura 39 se muestran los colores de la agrupación según como los colores de la Figura 35.

Se puede observar que los resultados obtenidos con las instancias pequeñas la $E(p)$ presenta poca diferencia con el secuencial, debido al número de objetos que contienen las instancias que fueron menores de 100 objetos, pero con la instancia que tiene más de 100 objetos, se presenta una gran diferencia entre el secuencia y el paralelo, esto nos indica que entre mayor número de objetos contenga la instancia, es mejor el algoritmo paralelo, esto se resume a que cada procesador ejecutó su tarea rápido, gracias a que se balanceó bien el trabajo en los procesadores disponibles.

Según los resultados, la programación paralela es más rápida, ya que divide las tareas y nuestro algoritmo funcionó correctamente en diferentes instancias (número de objetos y atributos), entonces se logró reducir la complejidad del algoritmo del método de Ward con la programación paralela.

Las instancias con menos objetos (menores que 100 objetos) se ejecutaron en un tiempo pequeño tanto secuenciales como paralelas, debido a que son menores objetos a comparar, en cambio con mayores objetos (mayores que 100) es un poco más lento debido a que se ejecutan más procesos para comparar, pero aun así es rápido el paralelo y con ello se pueden ejecutar problemas con gran número de objetos.

CAPÍTULO V. CONCLUSIONES

Con los datos obtenidos se concluye que:

El método de Ward agrupa mejor los objetos dispersos pero similares entre sí, ya que fue capaz de agrupar objetos en forma de espiral y con los cultivos de maíz. En este último se agregaron más atributos no solo fueron los valores de "*x*" y "*y*".

Con el algoritmo jerárquico aglomerativo con el método de Ward y la programación paralela, es posible resolver las instancias en menor tiempo que un algoritmo secuencial y con ello se puede trabajar con instancias que contiene un número mayor de objetos y variables, por ejemplo, con una instancia de 312 objeto se obtuvo que el algoritmo paralelo es un 11.71% más eficiente que el secuencial.

En la Tabla 27, se muestra una comparación de los resultados obtenidos en cada instancia, según su tiempo de ejecución y su eficiencia del algoritmo paralelo vs el secuencial.

Tabla 27. Resultados concentrados.

Instancia	Tiempo secuencial en segundos	Tiempo paralelo en segundos	% de eficiencia del algoritmo paralelo
Gallardo	0.000354	0.002607	2.71%
Elaboración propia	0.000458	0.008991	0.50%
Elaboración propia (maíz)	0.560161	1.095899	0.896%
Chang	478.261963	13.09	11.71%

Se observó que entre mayor número de objetos contenga la instancia es más eficiente el algoritmo paralelo, y con ello se puede resolver el problema de la

complejidad y ya se pueden probar instancias con mayor número de objetos que no existen en la literatura para dicho problema.

Con lo antes mencionado, la hipótesis fue verdadera y el objetivo general fue alcanzado, ya que con la programación paralela y el método de Ward se pueden agrupar objetos diferentes de un mismo conjunto en un menor tiempo.

Con esta aportación, hay gran posibilidad de trabajar con instancias y grandes conjuntos de objetos como los que se utilizan en Big Data u otras aplicaciones que contengan una gran cantidad de objetos y de diferentes variables, será más fácil y rápido la búsqueda y agrupación de dicha información.

5.1 Trabajos futuros.

- Poder trabajar con datos reales del maíz.
- Trabajar con otros tipos de datos.
- Comparar el tiempo de ejecución y la calidad de la agrupación entre el algoritmo de k-means vs Método de Ward paralelo.

Referencias

Adamczyk, K., Cywicka, D., Herbut, P., & Trześniowska, E. (2017). The application of cluster analysis methods in assessment of daily physical activity of dairy cows milked in the Voluntary Milking System. *Computers and Electronics in Agriculture*, 141, 65-72. http://dx.doi.org/10.1016/j.compag.2017.07.007

Baldo, L., Brenner, L., Fernandes, L. G., Fernandes, P., & Sales, A. (2005). Performance models for master/slave parallel programs. *Electronic Notes In Theoretical Computer Science*, 128(4), 101-121.

Balugani, E., Lolli, F., Gamberini, R., Rimini, B., & Regattieri, A. (2018). Clustering for inventory control systems. *IFAC-PapersOnLine*, 51(11), 1174-1179.

Berkhin, P. (2006). A survey of clustering data mining techniques. *In Grouping multidimensional data* (pp. 25-71). Springer, Berlin, Heidelberg.

Cabello, A. G., & Salama, A. (2012). Un estudio sobre la distribución regional de los préstamos en la Argentina por sector económico, 2000-2012. Una aplicación del análisis de cluster. *Analítika: revista de análisis estadístico*, (3), 43-57.
ISSN 1390-6208
e-ISSN 1390-7867

Cargnelutti Filho, A., & Guadagnin, J. P. (2011). Consistência do padrão de agrupamento de cultivares de milho. *Ciência Rural*, 41(9), 1503-1508.
ISSN 0103-8478

Cargnelutti Filho, A., Ribeiro, N. D., & Burin, C. (2010). Consistência do padrão de agrupamento de cultivares de feijão conforme medidas de dissimilaridade e métodos de agrupamento. *Pesquisa Agropecuária Brasileira*, 45(3), 236-243.
ISSN 0100-204X

Chang, H., & Yeung, D. Y. (2008). Robust path-based spectral clustering. *Pattern Recognition*, 41(1), 191-203.

Chavent, M., Lechevallier, Y., & Briant, O. (2007). DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis*, 52(2), 687-701.

Crouse, J. J., Moustafa, A. A., Bogaty, S. E., Hickie, I. B., & Hermens, D. F. (2018). Parcellating cognitive heterogeneity in early psychosis-spectrum illnesses: a cluster analysis. *Schizophrenia Research*, 202, 91-98.
https://doi.org/10.1016/j.schres.2018.06.060

Díaz R.A & Mormeneo I. (2002). zonificación del clima de la región pampean a mediante análisis de conglomerados por consenso. *Revista Argentina de agrometeorologia*, 2, 125-131.

Dumont, M., Reninger, P. A., Pryet, A., Martelet, G., Aunay, B., & Join, J. L. (2018). Agglomerative hierarchical clustering of airborne electromagnetic data for multi-scale geological studies. *Journal of Applied Geophysics*, 157, 1-9.
https://doi.org/10.1016/j.jappgeo.2018.06.020

Espinel, P. (2019). Procedimiento para efectuar una Clasificación Ascendente Jerárquica de un Conjunto de Puntos utilizando el Método de Ward. *Infociencia*, 9(1), 13-18.

Eszergár-Kiss, D., & Caesar, B. (2017). Definition of user groups applying Ward's method. *Transportation Research Procedia*, 22, 25-34.

Gallardo, J. s.f. Introducción al Análisis Cluster. *Universidad de Granada, Granada, España*. Disponible en http://www.ugr.es/~gallardo/pdf/cluster-g.pdf (2019)

Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Research on air Pollution*, 11(1), 40-56.
https://doi.org/10.1016/j.apr.2019.09.009

Heredia, L. M. C., Escobar, Y. C., & Díaz, Á. J. Á. (2012). Análisis cluster como técnica de análisis exploratorio de registros múltiples en datos meteorológicos. *Ingeniería de Recursos Naturales y del Ambiente*, (11), 11-20.
ISSN 1692-9918

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8), 651-666.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning, *New York: springer*, 112, 3-7.
ISSN 1431-875X
ISBN 978-1-4614-7137-0 ISBN 978-1-4614-7138-7 (eBook)
DOI 10.1007/978-1-4614-7138-7

Kim, H., Kim, B., Kim, S. H., Park, C. H. K., Kim, E. Y., & Ahn, Y. M. (2018). Classification of attempted suicide by cluster analysis: A study of 888 suicide attempters presenting to the emergency department. *Journal of affective disorders*, 235, 184-190.
10.1016/j.jad.2018.04.001

Larranaga p, Inza I & Moujahid A. (2012). Tema 14. Clustering. 2019, *Departamento de Ciencias de la Computaci´on e Inteligencia Artificial Universidad del País Vasco–Euskal Herriko Unibertsitatea* Sitio web: http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t14clustering.pdf

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?. *Journal of classification*, 31(3), 274-295.
DOI: 10.1007/s00357-014-9161-z

Navarro, C. A., Hitschfeld-Kahler, N., & Mateu, L. (2014). A survey on parallel computing and its applications in data-parallel problems using GPU architectures. *Communications in Computational Physics*, 15(2), 285-329.

Rampado, O., Gianusso, L., Nava, C. R., & Ropolo, R. (2019). Analysis of a CT patient dose database with an unsupervised clustering approach. *Physica Medica*, 60, 91-99.https://doi.org/10.1016/j.ejmp.2019.03.015

Rodríguez-Ariza, M. O., & Esquivel, J. A. (2005). Una valoración de la paleovegetación del sureste de la península ibérica durante la prehistoria reciente a partir de aplicaciones estadísticas en antracología. *In VI Congreso Ibérico de Biometría. Avances en Arqueometría* pp. 263-272.

Solano, J. C. C., Sevilla, F. C., Felipe, A. I. G., Membreno, A. M., & Calvo, E. R. (2008). Interpretación de las relaciones intragrupales de riesgos y de lesiones mediante análisis cluster jerárquico. *Revista de matemática: Teoría y Aplicaciones*, 15(2), 175-186.
ISSN: 1409-2433

Violán, C., Foguet-Boreu, Q., Roso-Llorach, A., Rodriguez-Blanco, T., Pons-Vigués, M., Pujol-Ribera, E., & Valderas, J. M. (2016). Patrones de multimorbilidad en adultos jóvenes en Cataluña: un análisis de clusters. *Atención Primaria*, 48(7), 479-492.

Zavala-Díaz, J. C. (2013). Optimización con cómputo paralelo: teoría y aplicaciones. Universidad Autónoma del Estado de Morelos.