



Universidad Nacional Experimental del Táchira

Vicerectorado Académico

Decanato de Docencia

Departamento de Ingeniería en Informática

Trabajo de aplicación profesional

Proyecto especial de grado

Autor(es): Jesús David Escalante Rodríguez

C.I.: 21.220.841

jesusd.escalante@unet.edu.ve

Tutor(es): Rossana Timaure

rttg@unet.edu.ve

San Cristóbal, Julio de 2018.2018



Universidad Nacional Experimental del Táchira
Vicerrectorado Académico
Decanato de Docencia
Departamento de Ingeniería Informática
Trabajo de Aplicación Profesional
Proyecto Especial de Grado

**Aprobación del Tutor para presentación de la Propuesta del Proyecto Especial de
Grado**

Yo, Rossana Timaure titular de la cédula de identidad No. V-12.021.405. en mi carácter de Tutor(a) del Proyecto Especial de Grado titulado: , presentado por el bachiller: Jesús David Escalante Rodríguez, titular de la cédula de identidad No. 21.220.841, por medio de la presente autorizo la presentación de la Propuesta de Proyecto Especial de Grado ante los jurados designados por la comisión de Trabajo de Aplicación Profesional del Departamento de Ingeniería en Informática, en virtud de considerar que reúne los requisitos establecidos en el artículo 16 de las Normas para el Trabajo de Aplicación Profesional de la UNET.

Nombres y apellidos del tutor
C.I. V- 12.021.405
Rossana Timaure

Introducción

La producción de papa criolla (*Solanum Phureja*), colombiana o yema de huevo como también es conocida, siempre ha tenido muchos desafíos por parte de la industria, debido a los grandes intereses provenientes de los diferentes consumidores que la misma posee, siendo así que es uno de los productos agrícolas más consumidos y con mayor importancia en el mundo, después del arroz, maíz y trigo. Colombia es el mayor productor de papa criolla en Latinoamérica y por lo cual es un país que posee grandes exigencias industriales de la misma (Ligarreto G - Suarez M, 2003).

Los tubérculos de papa criolla tienen aproximadamente entre dos y ocho centímetros (2-8cm), y se pueden clasificar en tres calibres según su diámetro promedio de donde radican los intereses comerciales de la misma. Los tubérculos de entre dos y medio y cuatro centímetros son preferidos para encurtidos y pre-cocidos mientras los tubérculos promedio entre cuatro y seis y medio centímetros son preferidos para frituras en hojuelas o con más de cinco centímetros para frituras en tiras (CORPOICA, 2009).

El crecimiento vegetal es definido por Cabezas en 2005 como «El aumento irreversible del tamaño y peso seco de las plantas (altura, área foliar, diámetro, número de células y cantidad de protoplasma) o los cambios que ocurren en una planta o población de plantas a través del tiempo, fenómeno acompañado del aumento en la complejidad estructural metabólica del organismo (diferenciación celular, número de hojas), por procesos de división y alargamiento celular, incorporación de materia y energía del ambiente (fotosíntesis, absorción de agua y de iones) y metabolización subsiguiente, la cual se traduce en multiplicación y diferenciación celular. Este proceso está íntimamente relacionado con algunos factores internos como fotosíntesis, respiración, transpiración, condiciones de estrés, concentración enzimática, balance hormonal y expresión genética» (Píneros C, 2009).

En la papa son dos los procesos fisiológicos asociados directamente al rendimiento de la misma,

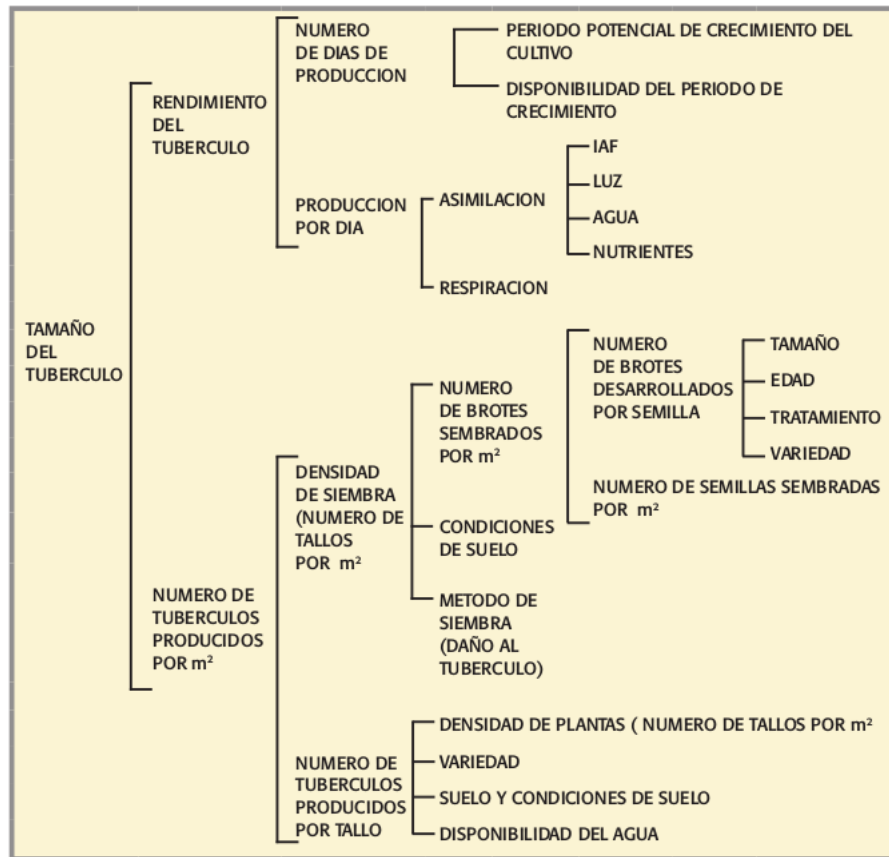
la fotosíntesis y la respiración. Y estos dos son procesos íntimamente asociados, ya que durante la fotosíntesis se producen carbohidratos que son consumidos durante la respiración, pero una gran cantidad de estos carbohidratos contribuyen al proceso de producción incrementando el tamaño de los tubérculos, el nivel y periodo de crecimiento de los tubérculos es una variable que responde directamente a la expresión de rendimiento expresada como producción por día, siendo así que cerca de un noventa por ciento (90 %) del peso acumulado de los tubérculos producidos por una planta es producto de la asimilación de dióxido de carbono, así como otros procesos similares (Beukema, 1979). Sin embargo, estos procesos se ven afectados por condiciones externas tales como la intensidad de luz, temperatura, longitud de los días, condiciones del suelo que son variables no controladas, o por condiciones como la cantidad de riego, fertilizantes aplicados y productos de abono para la mejora de los suelos que son variables controlables pero que requieren de costos adicionales de producción (Píneros C, 2009). Y es por ello que se sigue buscando con el manejo de variables controladas el control sobre la cantidad total producida y tamaño de los tubérculos sin producir costos adicionales, siendo la densidad de siembra una de estas variables, en la cual nos enfocaremos en esta investigación (Darghan E - Bernal E - Rodríguez L, 2017).

Las variables que influyen en el rendimiento de la papa pueden ser observados en la 1. Pág.3.

Sin embargo la densidad de siembra es una variable de tipo espacial, ya que se define como la distancia entre plantas y entre linderas, siendo así una coordenada en X y una coordenada en Y en un plano cartesiano, por lo cual un estudio estadístico de la misma para clasificar o predecir un modelo no puede ser del tipo de regresión lineal o de un análisis de varianza debido a los residuales que deben ser independientes en cada punto, no tomando en cuenta una relación entre vecinos como si ocurre en los cultivos de papa, las plantas compiten por los recursos y por consiguiente por producir más, haciendo así los residuales dependientes y así mismo no optima la aplicación de dichos análisis (Píneros C, 2009).

Se plantea el uso de regresión espacial entonces, pero el manejo de la densidad de siembra como variable es del tipo cualitativo con un valor de distancia entre linderos y plantas fijo para una cantidad de densidades definidas, implicando tener que aplicar regresión espacial para cada uno de los valores de estas densidades. Sin embargo aunque estos análisis me arrojen resultados con un comportamiento y una distribución estadística similar, no se puede asegurar una relación entre cada una de las densidades definidas basados en una normalización n-variante, donde n sería el número de densidades definidas. Por ello se propone en esta investigación el uso de una red neuronal probabilística (PNN) donde no importan los supuestos y definiciones de un modelo lineal.

Figura 1: Variables de influencia sobre el rendimiento de Solanum Phureja.



Una red neuronal artificial (ANN) es un conjunto de nodos de un programa (neuronas) interconectados entre si, simulando el proceso de pensamiento humano, se pudiera considerar como una caja negra entrenada previamente para esperar una entrada y basado en las características o comportamiento de la misma proporcionar una determinada salida, eliminando asi la necesidad de diferentes algoritmos que deban analizar comportamientos cada uno por separado. Una red neuronal probabilística (PNN) no es mas que una ANN que usa funciones estadísticas que escalan la variable no linealmente como una forma de campana o una distribución normal (Stephen G, 2007).

El objetivo de esta investigación es construir un clasificador de densidades de siembra a través de una red neuronal probabilística, además de comparar los resultados provenientes de una regresión espacial versus datos observados, para concluir cuan efectivo puede ser aplicar una regresión espacial para estos casos.

Capítulo 1

Preliminares

1.1. Planteamiento y formulación del problema

Una red neuronal artificial (ANN) es un modelo que crea una relacion entre una configuracion de senales de entrada y una senal de salida usando un modelo derivado de nuestro entendimiento de como el cerebro responde a determinados estímulos de entradas sensoriales. Así como un cerebro usa ese conjunto de células interconectadas llamadas neuronas, así las ANN usan una red de neuronal artificiales o nodos para solucionar problemas aprendidos (Lantz B, 2015).

En general, las ANN son aprendices versátiles que pueden ser aplicarse a casi cualquier tarea de aprendizaje en cualquier contexto, clasificación, predicción, reconocimiento de patrones e incluso pudiéramos hablar de tareas de supervisión de procesos. Las ANN, se aplican mejor a los problemas cuyos datos de entrada y salida son bien definidos o bastante simples, sin embargo el proceso que relaciona la entrada con una determinada salida es en extremo complejo, por eso se conocen como un método de caja negra (Lantz B, 2015).

De una manera muy arcaica, las ANN han sido usadas desde hace cincuenta años para simular como el cerebro humano es capaz de resolver problemas. Primero surgieron las simples funciones lógicas AND y OR, que ayudaron a los científicos a entender como biologicamente el cerebro opera. Sin embargo, como el avance tecnológico y computacional ha sido incrementado y continua creciendo potencialmente durante los últimos años, las ANN ha incrementado su complejidad siendo usadas frecuentemente en múltiples problemas como:

1. Programas de reconocimiento de voz y escritura, como los que usan los correo de voz de los servicios de transcripción y máquinas clasificadoras de correo postal.

2. La automatización de dispositivos inteligentes como los controles ambientales de un edificio de oficinas o automóviles autoguiados y drones autoguiados.
3. Modelos sofisticados de patrones climáticos y otros fenómenos científicos, sociales o económicos.
4. La clasificación y reconocimiento de cualquier cosa, basado en sus atributos y características.

Existen muchos métodos y lenguajes que permiten la estructuración y desarrollo de ANN, destacándose lenguajes como Python y softwares como Statgraphics y SPSS, sin embargo como el objetivo de esta investigación es trabajar con una variable de tipo espacial que amerita la aplicación de regresión espacial antes para la normalización de los datos recogidos y observados, se propone el uso del lenguaje R con el paquete `neuralnet` para el manejo y construcción de la red neuronal.

R es un lenguaje de programación y entornos para gráficos y cálculos estadísticos, es una implementación diferente al lenguaje S (Chambers, 1998), y se desarrolla bajo un proyecto de software bajo Licencia Pública General (General Public License, GNU) de la Free Software Foundation (<https://www.fsf.org/>) en forma de código fuente. Se compila y se ejecuta en una amplia variedad de plataformas entre ellas FreeBSD, Linux, Windows y MacOS. (The R Project for Statistical Computing, 2018).

El paquete `neuralnet` es un paquete del lenguaje R de muy sencilla utilización que permite la aplicación, visualización e implementación de redes neuronales. El paquete permite configuraciones flexibles a través de elección personalizada del error y función de activación. Además, el cálculo de pesos generalizados está implementado (CRAN-R Project, 2016).

Esta propuesta permitiría ajustar y entrenar una red neuronal artificial probabilística que permita la clasificación de densidades de siembra de papa yema de huevo (*Solanum Phureja*) teniendo como entrada las siguientes variables:

1. **Peso total del cultivo.** Variable cuantitativa expresada en gramos (gr), que indica el peso de todos los tubérculos recogidos en la siembra.
2. **Total de tubérculos recogidos.** Variable cuantitativa que indica el número de tubérculos recogidos en la siembra.

3. **Diametro ponderado de todos los tuberculos.** Variable cuantitativa expresada en centimetros (cm), que indica el calculo del diametro ponderado (por los tuberculos ser ovoides) de cada tuberculo recogido en la siembra.
4. **Calibre de los tuberculos recogidos.** Variable cualitativa, con 4 valores posibles que definen a que calibre pertenece un tuberculo. Siendo los calibres divididos en tuberculos menores a dos centimetros (<2cm), entre dos y cuatro centimetros (2-4cm), entre cuatro y seis centimetros (4-6cm) y mayores a seis centimetros (>6cm).

La densidad de siembra es la distancia entre plantas y leras o sencillamente cuantas plantas son sembradas por metro cuadrado de siembra.

El objetivo es comparar los resultados de clasificacion alimentando la red con datos observados y con los mismos datos normalizados espacialmente a traves de regresion espacial, para observar si es necesario el analisis de los datos recogidos antes de la clasificacion y a su vez concluir hasta que punto la densidad de siembra afecta las variables antes mencionadas. Otra de las pruebas que se plantea usar en esta investigacion es realizar los analisis eliminando variables de entrada atisbando cuales variables permiten y generan una mejor clasificacion.

La curva ROC es una herramienta estadistica utilizada en analisis de clasificacion para determinar la capacidad discriminante de una prueba diagnostica dicotomica. Esta capacidad discriminante esta sujeta al valor umbral elegido de entre todos los posibles resultados de la variable de decision, es decir, la variable por cuyo resultado se clasifica (en nuestro caso la densidad de siembra) en un determinado grupo. La curva es el grafico resultante de representar, para cada valor umbral, las medidas de sensibilidad y especificidad de la prueba diagnostica (Benavides A, NF).

Posterior a las diversas clasificaciones planteadas, se propone la construccion en lenguaje R de las curvas ROC correspondientes para determinar cuan optima es la clasificacion resultante, permitiendo asi llegar a las conclusiones deseadas.

1.2. Objetivos

1.2.1. Objetivo General

1.2.2. Objetivos Específicos

-

1.3. Justificación e Importancia

Colombia es el mayor productor, consumidor y exportador de papas diploides en el mundo; tiene una ventaja competitiva notable debido a ser centro de diversidad y poseer gran aceptación por los consumidores debido a las características del tubérculo. Adicionalmente, en este país se ha desarrollado una amplia tradición como cultivo tecnificado, con potencial de industrialización y exportación (Rodríguez - Ñustez - Estrada, 2009). Además de contar con uno de los recursos genéticos que ofrece las mayores oportunidades de exportación como alimento procesado exclusivo y sin competencia. Desafortunadamente, muchas de las estrategias de manejo del cultivo de la papa, han sido adaptadas a papa criolla, creando un sistema productivo poco eficiente (Piñeros, 2009). Los altos precios de los insumos agrícolas y el mal manejo de los cultivos agrónomicamente provoca baja productividad y amenaza la competitividad del sistema de producción, por lo que es importante identificar factores limitantes del rendimiento y desarrollar prácticas innovadoras para el cultivo, tales como una gestión nutricional integrada y equilibrada, que es una de las prácticas más eficientes para garantizar a la planta la oportunidad de expresar su potencial genético que eventualmente se reflejara en una mejor calidad y rendimiento (López - Gómez - Rodríguez, 2014).

Varias investigaciones revelan que el tamaño promedio de los tubérculos, su variabilidad y su conteo, definen una única distribución, presentando mayor variabilidad de rangos de tamaño en la etapa final de llenado, con aumento del rendimiento de los mayores tamaños por cuenta de los tamaños inferiores, lo que sugiere que un tubérculo que pasa de un calibre inferior al siguiente, no es sustituido por otro de calibre anterior, además, parece existir una correlación negativa entre la variabilidad relativa del tamaño del tubérculo y el número de tubérculos por unidad de área (Struik, 1991). Es importante reconocer, que debe tenerse cuidado con la interpretación de esta correlación, pues hasta cierto punto, la naturaleza de los datos asociados al número de tubérculos por planta pudiera ser similar a la de los datos composicionales, es decir, si un componente aumenta, el otro está forzado a disminuir, lo que genera correlaciones sin conexión lógica en datos verdaderamente

composicionales, sin embargo, el numero de tuberculos por calibre no suma una constante, pudiera ocurrir que su maximo posea una distribucion especifica, entonces una mayor cantidad de tuberculos de un calibre podria provocar un numero menor de tuberculos en otro calibre especifico, lo que en esencia puede generar correlaciones en las que debe tenerse mucho cuidado al momento de su interpretacion. La importancia del cultivo obliga a muchos investigadores a realizar diferentes estudios para mejorarlo segun el uso que vaya a darse a los tuberculos, generando nuevas variedades resistentes a plagas y enfermedades y de facil adaptacion a diferentes climas, procurando hacer uso del espacio de siembra de forma optima, por lo que varios estudios evidencian el estudio de la densidad de siembra para evaluar sobre todo el rendimiento. En 2017, Darghan evaluo en lugar del rendimiento, un indicador que se asocia directamente como lo es el calibre de los tuberculos, los cuales en la practica se manipulan en cuatro categorias de diametro promedio (hasta 2 cm, de 2 a 4 cm, de 4 a 6 cm y más de 6 cm). La naturaleza de esta variable imposibilita usualmente la comparacion de la respuesta (conteos de tubérculos) para cada densidad de siembra (definidas como 30cm*100cm, 40cm*100cm y 50cm*100cm) mediante analisis de varianza, pues en el caso de conteos, otras distribuciones como la Poisson y la Binomial negativa se adaptan mucho mejor al tipo de dato generado. Los modelos clasicos de regresion Poisson y binomial y negativa, en sus modalidades usuales o en la opción inflada por ceros (Cameron, 1998) en datos de conteo pertenecen a la familia de modelos lineales generalizados (Zeileis, 2008) y los desarrollos recientes permiten generar varias estadisticas que permiten su comparacion con sus contrapartes sin ceros en exceso, lo que resulta util en la eleccion del mejor modelo para relacionar predictores y respuesta en conteos.

Sin embargo, los estudios realizados hasta ahora son todos haciendo suposiciones del tipo lineal y haciendo modificaciones que permitan así su estudio a traves de anova o regresion lineal tomando en cuenta los patrones de vecindad, mas la propuesta consiste en clasificar los datos a traves de una red neuronal que no amerita de suposiciones estadisticas o matematicas, que solo aprendiendo de lo observado es capaz de realizar una clasificacion de densidades de siembra, anadiendo aun así para conseguir un mejor resultado un analisis de regresion espacial previo, con el objetivo de solo con datos de siembra previos conocer a que densidad de siembra se debe cosechar para obtener un deseado calibre para fines comerciales e industriales.

1.4. Alcance y Limitaciones

La red neuronal artificial probabilistica a implementar tiene como proposito la clasificacion de la densidad de siembra en papa criolla o yema de huevo, siendo util para incrementar las posibi-

lidades de obtener un determinado y deseado calibre de tuberculos conociendo la densidad a la que se debe sembrar, anadiendo tambien para su optimizacion analisis de regresion espacial y un conjunto de pruebas que van desde la variacion de las variables de entrada hasta la utilizacion de los datos observados, para definir que tan buena o no es la clasificacion. Sin embargo las pruebas se relizaran bajo solo un conjunto de datos reales que fueron tomados bajo las medidas de cuatro calibres de tuberculos y tres densidades de siembra.

Los datos disponibles del cultivo de papa se tienen en base a tres densidades de siembra, cantidad de plantas sembradas, cantidad de tuberculos recogidos, peso total del cultivo, diametro ponderado de cada tuberculo y calibre de los tuberculos y forman parte de la investigacion de Maestria del estudiante de la Universidad Nacional de Colombia, Nelson Bernal.

Los resultados de las pruebas sobre los escenarios planteados seran validados estadisticamente con ayuda del analisis de las curvas ROC generados por estos mismos resultados.

Capítulo 2

Fundamentos teóricos

2.1. Antecedentes

La clasificacion por medio de redes neuronales ha sido un hito ya marcado en el campo agromico, muchos estudios se han realizado con el objetivo de analizar ciertos comportamientos de plantas y los beneficios que se puedan sacar de ellas. En 2011, el grupo de investigacion de sistemas de procesamiento y control de senales de la Universidad Nacional Tenaga de Malasia desarrollo un sistema de inteligencia con un enfoque novedoso para la clasificacion de frutas usando tecnicas de procesamiento de imagenes digitales y redes neuronales artificiales, con el objetivo de desarrollar un metodo de clasificacion rapido con una meta del 100 % de eficiencia. El estudio se realizo con cinco frutas, manzanas, platanos, zanahorias, mangos y naranjas, extrayendo de ellas siete caracteristicas en funcion de la forma y el color. La captura de las imagenes se realizo con una camara digital convencional y las manipulaciones a los datos y construccion de la red con el software MATLAB. Los resultados obtenidos durante esta investigacion fueron de gran avance en el campo de reconocimiento de patrones en imagenes.

En 2011, Mayabiro E presento en la Universidad Nacional Experimental del Tachira, un prototipo sobre el entorno MATLAB para el calculo de la tasa de germinacion de plantulas de pimenton previamente segmentadas. El entorno desarrollado permitia establecer una clasificacion de las plantulas, hojas u objetos de la misma por medio de redes neuronales. Se realizo el entrenamiento de multiples redes neuronales multicapas con algoritmos de retropropagacion, donde aunque variaban las capas intermedias de las redes y sus funciones de transferencia fueron entrenadas con los mismos datos de entrada, validandolas con una base de datos de pruebas para seleccionar al final una con salidas similares a las deseadas.

Otro estudio realizado en 2013 por Stephen Gang Wu consistia en el estudio teorico de tecnicas de procesamiento de imagenes y datos para el reconocimiento automatico de hojas para la clasificacion de plantas. Doce caracteristicas de las plantas fueron extraidos y distribuidas en cinco variables principales que constituian el vector de entrada de una red neuronal artificial probalistica, que habia sido entrenada con 1800 hojas para clasificar 32 tipos de plantas con una precision superior al 90 %, el autor aseguraba que su metodologia de implementacion de la PNN era facil y rapida en comparacion de otras investigaciones similares.

En 2017, Bernal N realizo un estudio de campo con el cultivo de papa criolla para evaluar la influencia de la densidad de siembra asociada a distancias entre plantas de 30,40 y 50 cm y distancias entre surcos de 100 cm sobre el conteo de tuberculos de calibres inferiores a 2 cm, entre 2 y 4 cm, entre 4 y 6 cm, y de mas de 6 cm de diametro ponderado y sobre el peso fresco en gramos de los tuberculos. Los tuberculos cosechados se clasificaron y contaron mediante tamizado y se pesaron en su totalidad sin discriminar por calibre. Los modelos estadisticos empleados para modelar el comportamiento de la cosecha, evidenciaron el efecto significativo de la densidad de siembra sobre el conteo de tuberculos y calibre y se observo una razon aproximada de 40:40:20:1 desde el calibre menor al mayor. El efecto de la competicion, en todos los modelos probados resulto significativo, aumentando en la mayoria de los casos a medida que disminuia la distancia entre plantas, tanto en el patron de vecindad intrahileras como en el caso de inter e intrahileras.

2.2. Bases Teóricas

2.2.1. Lenguaje R.

R es un conjunto integrado de *software* de código abierto para el almacenamiento, manipulación, cálculo y visualización de datos para computación y gráfica estadística, puede ser compilado y ejecutado en Windows, Mac OS X y otras plataformas UNIX (como Linux), se distribuye usualmente en formato binario (<https://www.r-project.org/about.html>, 2018). El proyecto de *software* R fue iniciado por Robert Gentleman y Ross Ihaka. El lenguaje fue influenciado por lenguaje S desarrollado originalmente en Bell Laboratories por John Chambers y sus colegas. Desde entonces ha evolucionado para el cálculo estadístico asociado a diversas disciplinas para contextos académicos y comerciales. En R, la unidad fundamental de código compartible es el paquete, el cual agrupa código, datos, documentación y pruebas, y resulta simple de compartir con otros. Para enero del 2015 ya habían más de 6.000 paquetes disponibles en la Red Integral de Archivos de R,

conocido comunmente por su acrónimo CRAN, el cual es el repositorio de paquetes . Esta gran variedad de paquetes es una de las razones por las cuales R es tan exitoso, pues es probable que algún investigador o académico ya haya resuelto un problema en su propio campo usando esta herramienta, por lo que otros usuarios simplemente podrán recurrir a ella para su uso directo o para llamarla en un nuevo código (Wickham,2015).

2.2.2. RStudio.

RStudio es un ambiente de desarrollo integrado (*Integrated Development Environment*, IDE) que ofrece herramientas de desarrollo vía consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la depuración y la gestión del espacio de trabajo. RStudio está disponible para Windows, Mac y Linux o para navegadores conectados a RStudio Server o RStudio Server Pro (Debian / Ubuntu, RedHat / CentOS, y SUSE Linux) (<https://www.rstudio.com/about/>, 2018)

2.2.3. Glosario

R :

Capítulo 3

Fundamentos Metodológicos

A continuación se plantea la estructura a seguir por el presente trabajo, detallando el enfoque, tipo, nivel y diseño de la investigación y la metodología a implementar entre otros.

3.1. Enfoque de la investigación

La presente investigación se desarrollará siguiendo un enfoque cuantitativo, puesto que, como lo indican Pallela y Martins (2012) , “la investigación cuantitativa requiere el uso de instrumentos de medición y comparación, que proporcionan datos cuyo estudio necesita la aplicación de modelos matemáticos y estadísticos, el conocimiento está basado en hechos”. Los datos a usar en el desarrollo del paquete provienen de ensayo experimental desarrollado en los Laboratorios de Biofertilizante e Instrumentación Biológica, adscritos al decanato de Investigación de la Universidad Nacional Experimental del Táchira .

3.2. Tipo o nivel de investigación

Este proyecto plantea un tipo de investigación de campo, según como lo indican Pallela y Martins (2012), la investigación de campo “consiste en la recolección directamente de la realidad donde ocurren los hechos, sin manipular o controlar variables” ya que permite indagar los efectos de la interrelación entre los diferentes tipos de variable en lugar de los hechos.

En este punto se debe determinar la profundidad que abarca esta investigación, teniendo en cuenta que de acuerdo con el nivel de la investigación es definido como “grado de profundidad con

que se aborda un fenómeno u objeto de estudio” (Arias, 2012).

En este sentido, se tiene que dadas las características del proyecto, se asocia con un nivel descriptivo, tal como lo indican Pallela y Martins (2012), “hace énfasis sobre conclusiones dominantes o sobre como una persona, grupo o cosa se conduce o funciona en el presente” esto debido a que se medirán los datos extraídos sin alterarlos para ser mostrados en el sistema.

Cuando se habla de un nivel descriptivo junto con una investigación de tipo de campo, en ella no se formulan hipótesis y las variables se enuncian en los objetivos de la investigación que se desarrollará.

3.3. Diseño de la investigación

Según Arias(2012), el diseño de la investigación es “la estrategia general que adopta el investigador para responder al problema planteado” (p.21) por lo que es vital establecer una correcta secuencia de pasos para elaborar el prototipo de software que dará solución a la problemática principal de la investigación.

Con este enfoque, se tiene que este trabajo seguirá un diseño no experimental, enfocado en el uso de información existente, de acuerdo con lo dicho por Pallela y Martins (2012) al definir el diseño no experimental como:

Es el que se realiza sin manipular en forma deliberada ninguna variable. El investigador no sustituye intencionalmente las variables independientes. Se observan los hechos tal y como se presentan en su contexto real y en un tiempo determinado o no, para luego analizarlos. Por lo tanto, este diseño no se construye una situación específica sino que se observan las que existen. Las variables independientes ya han ocurrido y no pueden ser manipuladas, lo que impide influir sobre ellas para modificarlas. (p.81)

Esto indica que no hay manipulación de variables. Esta investigación presenta una modalidad de proyecto especial que, como lo indican Pallela y Martins (2012), los proyectos especiales “destinados a la creación de productos que puedan solucionar deficiencias evidenciadas, se caracterizan por su valor innovador y aporte significativo” (p.92), ya que se creará un *software* aplicable al área de estudio.

3.4. Metodología

Para el desarrollo del paquete se seguirán las pautas estándar establecidas para la creación de paquetes y extensiones en R.

Creación del esqueleto del paquete.

En esta etapa se diseñaran y crearan los directorios, ficheros y objetos que conformaran el paquete.

Registrar el método para el envío y uso de funciones.

Es esta etapa del desarrollo se estableceran las dependencias sobre los paquetes de la base fuente de código R y sus métodos de conexión, considerando el manejo de versiones y los criterios de mantenimiento, además se estableceran los espacios de nombre o las estrategias para la búsqueda y utilización de las variables; unificando estos criterios a las funciones que seran diseñadas.

Diseño y codificación de las funciones.

Los métodos para el diseño de las funciones primarias en R serán los diagramas de flujo; y para su codificación se seguirán las normas de estilo para codificación en R, sugeridas por Wickham (2015) y por el creador del paquete *formatR* Xie(2017), además se establecerán la dependencia con las funciones de código base y las recomendadas para desarrollo en R.

Pruebas unitarias de las funciones.

Debido a que los paquetes en R están conformados, entre otros elementos por las funciones primarias, a cada una de ellas se les realizaran pruebas unitarias en dos fases, la primera con datos sintéticos que permitan comprobar cada estado del diagrama de flujo, que esquematiza la solución numérica que permite el cálculo de cada uno de los índices fisiológicos para entre otra herramientas pueden utilizarse los paquetes de *RUnit* (Zenka, 2015) y *testthat* (Wickham, 2017), y la segunda etapa donde cada función asociada a un índice fisiológico se le realizan prueban

con los conjuntos de datos de prueba que formaran parte integral del paquete y con los cuales se desarrollarían los ejemplos prácticos que conformarían la documentación que acompañaría al paquete R.

Chequear la carga del paquete.

En esta etapa del desarrollo se utilizaran las funciones de chequear paquete que ofrece el código R; cuya finalidad es verificar cada fichero del árbol de carpetas asociadas a cada elemento de la estructura o esqueleto del paquete, que a su vez creara el archivo de documentación en LaTeX y/o HTML, compilará el código fuente y creará las librerías de enlace dinámico (*dynamic link library* DLL).

Construcción del método de distribución del paquete.

Se seleccionara la forma de distribución del paquete desde el repositorio local, creando los ficheros fuentes (en formato *tarball*) y en binario.

3.5. Aspectos administrativos

La realización de la investigación será planificada según lo establecido en el siguiente diagrama:

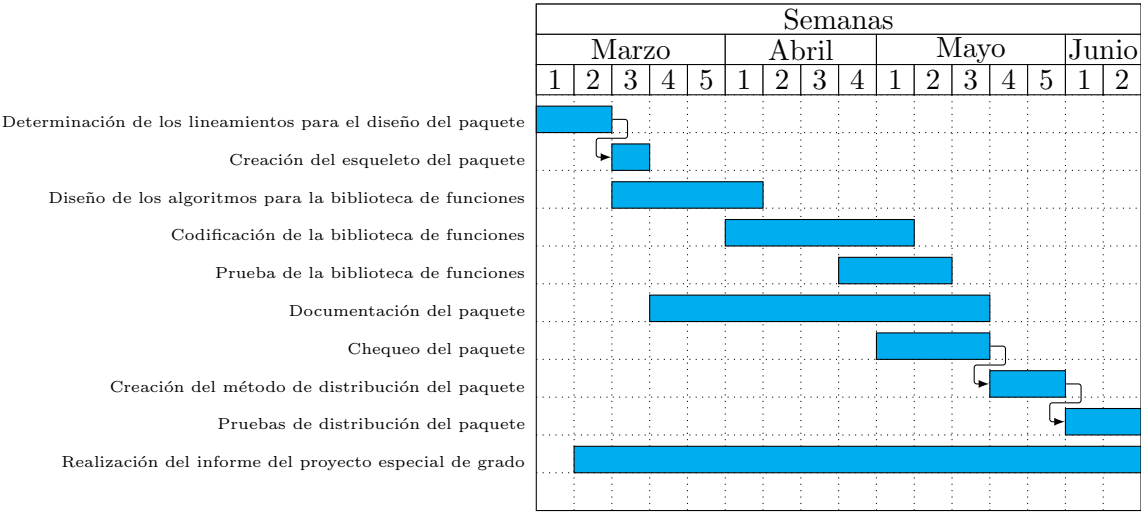


Figura 3.1: Diagrama de Gantt con la planificación del proyecto especial de grado